Giorgio Bongiovanni · Gerald Postema
Antonino Rotolo · Giovanni Sartor
Chiara Valentini · Douglas Walton *Editors*

Handbook of Legal Reasoning and Argumentation

Springer

Handbook of Legal Reasoning and Argumentation

Giorgio Bongiovanni · Gerald Postema
Antonino Rotolo · Giovanni Sartor
Chiara Valentini · Douglas Walton
Editors

Handbook of Legal Reasoning and Argumentation

Springer

Editors
Giorgio Bongiovanni
Dipartimento di Scienze Giuridiche
and CIRSFID
Università di Bologna
Bologna
Italy

Gerald Postema
Department of Philosophy
University of North Carolina
Chapel Hill, NC
USA

Antonino Rotolo
CIRSFID
Università di Bologna
Bologna
Italy

Giovanni Sartor
Department of Law
European University Institute
Florence
Italy

Chiara Valentini
Department of Law
Universitat Pompeu Fabra
Barcelona
Spain

Douglas Walton
University of Windsor, Centre for Research
in Reasoning, Argumentation
and Rhetoric (CRRAR)
Windsor, ON
Canada

ISBN 978-90-481-9451-3 ISBN 978-90-481-9452-0 (eBook)
https://doi.org/10.1007/978-90-481-9452-0

Library of Congress Control Number: 2018933503

Printed on acid-free paper

This Springer imprint is published by the registered company Springer Nature B.V.
The registered company address is: Van Godewijckstraat 30, 3311 GX Dordrecht, The Netherlands

Contents

Part I Basic Concepts for Legal Reasoning

Part II Kinds of Reasoning and the Law

Part III Special Kinds of Legal Reasoning

Contributors

Colin Aitken School of Mathematics and Maxwell Institute, The University of Edinburgh, Edinburgh, UK

Amalia Amaya Instituto de Investigaciones Filosóficas, Universidad Nacional Autónoma de México, Mexico City, Mexico

Kevin D. Ashley School of Law and Graduate Program in Intelligent Systems, University of Pittsburgh, Pittsburgh, PA, USA

Carla Bagnoli Dipartimento di Studi Linguistici e Culturali, Università di Modena e Reggio Emilia, Modena, Italy; University of Oslo, Oslo, Norway

Giorgio Bongiovanni Dipartimento di Scienze Giuridiche and CIRSFID, Università di Bologna, Bologna, Italy

Bartosz Brożek Department for the Philosophy of Law and Legal Ethics, Jagiellonian University, Kraków, Poland

Cristiano Castelfranchi Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche (CNR), Rome, Italy

Samuele Chilovi Departament de Filosofia, Universitat de Barcelona, Barcelona, Spain

Marcello Di Bello Lehman College - City University of New York, Bronx, USA

Jaap Hage Faculty of Law, Maastricht University, Maastricht, The Netherlands

Kenneth Einar Himma School of Law, University of Washington, Seattle, WA, USA

Lewis A. Kornhauser School of Law, New York University, New York, NY, USA

Emiliano Lorini IRIT-CNRS Toulouse University, Toulouse, France

Fabrizio Macagno IFILNOVA, Instituto de Filosofia da Nova, Universidade Nova de Lisboa, Lisbon, Portugal

Andrei Marmor Cornell Law School, Cornell University, Ithaca, New York, NY, USA

J. J. Moreso Departament de Dret, Universitat Pompeu Fabra, Barcelona, Spain

Veronica Rodriguez-Blanco School of Law, University of Surrey, Guilford, UK

Antonino Rotolo Dipartimento di Scienze giuridiche, Università di Bologna, Bologna, Italy

Giovanni Sartor Dipartimento di Scienze Giuridiche, Università di Bologna, Bologna, Italy; European University Institute, Florence, Italy

Burkhard Schafer Law School, The University of Edinburgh, Edinburgh, UK

Chiara Valentini Department of Law, Universitat Pompeu Fabra, Barcelona, Spain

Bart Verheij Faculty of Science and Engineering, University of Groningen, Groningen, The Netherlands

Douglas Walton University of Windsor, Centre for Research in Reasoning, Argumentation and Rhetoric (CRRAR), Windsor, ON, Canada

Wojciech Załuski Department of Philosophy of Law and Legal Ethics, Jagiellonian University, Krakow, Poland

Introduction

Since ancient times, there has been a presumption by judges, lawyers, and other legal professionals that legal reasoning is based on some kind of rationality that an agent who carries out actions and makes decisions can be presumed to be operating with. For example, legal reasoning does sometimes explicitly appeal to the existence of such a rational agent, called a rational person. But on the other hand, there is much skepticism and controversy about this presumption and more specifically how it applies to legal reasoning. For one thing, there is popular skepticism about whether there is something that can be called legal logic. For another thing, what people traditionally have often seemed to have had in mind is that legal logic fits the model of legal reasoning called mechanical jurisprudence. On this model, deductive logic is used to draw the rational conclusion in a given case at issue, say in a trial, by fitting a strictly universal generalization (All X without allowing any exceptions are Y), to a legal fact X, and drawing a conclusion Y. The problem with this model, although it fits occasionally, it is not applicable to the broad majority of cases in law, where the arguments used to support or attack a conclusion are defeasible (subject to exceptions). So, in the past, we have remained stuck in a dilemma where we are forced to concede that either legal reasoning does not have a logic, or if it does, it is one that is not applicable to the majority of cases being adjudicated on a daily basis.

Recent research on argumentation, especially in the field of artificial intelligence and law, offers a way out of this dilemma, by two means. Argumentation can be defined as a method for identifying, analyzing, and evaluating the pro and con arguments on both sides of a disputed issue where the factual knowledge base needed to resolve a dispute may be incomplete or inconsistent, and fallible arguments are used on both sides to arrive at a provisional conclusion based on a standard of proof appropriate for the case.

Examples of how argumentation tools can be applied to real legal cases are given in this Handbook. One such tool is the use of argumentation schemes, common forms of argument that can be deductive, but for the most part represent forms of reasoning that are defeasible, as they are subject to criticism or rebuttal by the asking of critical questions. The other is to apply legal reasoning in a dialectical framework which uses burdens and standards of proof, along with other devices, to take the context of use of an argument in a specific setting (e.g., in a trial in a particularly legal system) into account. Moving forward with this task means that linguistic interpretation of legal terms needs to be treated as inherently pragmatic in nature. Such an approach must not only take into account the semantic meaning of words and expressions, but also their pragmatic aspects considering how they are used in a communicative context. This latter approach requires not only considering the rationality of both a single agent with individual goals, but also the rational decisions and actions of several agents who reason together to deliberate to carry out their collective goals.

This Handbook shows, from a number of different angles and perspectives, and using a number of different tools, many of which may be new to readers, how this new approach to legal reasoning can be applied to many different important aspects of legal reasoning, throwing light on number of key problems and providing new avenues for solving them. By this means, the reader is allowed to look at legal reasoning in a fresh way, and thereby move forward to overcome the traditional dilemma about whether there is a legal logic or not.

One of the problems confronting the traditional approaches to legal reasoning is the uncertainty among legal scholars at this point in time, about the relationship between argument, at least argument in the sense of the term representing rationality, and reasoning. As one might expect, at this time there are also differences among the theoreticians on how to define the notion of an argument in precise enough terms to make the concept useful for the study of computational models of legal argument. Moreover, as noted above, accepting standard models of reasoning that have been dominant in the past, such as those of classical deductive logic, represents legal reasoning as mechanical jurisprudence based on absolutely universal rules of law not subject to exceptions. Hence, the problem of how to define the notion of an argument in a way that enables the drawing of distinction between reasoning and argument is one that pervades attempts to model legal reasoning by some more flexible notion of argument that can do justice to defeasible legal inferences of the most typical kind. Many of the chapters in the Handbook confront this problem.

Although some theorists now prefer the language of arguments, more traditional theorists would like to have a point of entry into this family of concepts by first of all defining what seems be the less problematic notion of a reason. Giorgio Bongiovanni, in Part I, Chapter "Reasons (and Reasons in Philosophy of Law)," provides a theory to answer this question and to classify different kinds of reasons. In this chapter, he distinguishes between normative, motivating, and explanatory reasons, and provides a theoretical framework for drawing a distinction between reasons for belief and reasons for actions. This chapter helps the reader move forward to the other chapters based on the assumption that some sense can be made of the distinction between the fundamental idea of presenting reasons to reasonably accept or reject a disputed claim, and the idea of presenting pro or con arguments for this same purpose. Overcoming this linguistic and underlying theoretical problem paves the way for the explorations of legal reasoning in the rest of the book.

Reasons in law are of course closely related to reasons in moral philosophy. Carla Bagnoli, in Chapter "Reasons in Moral Philosophy" of part I, clarifies the functions of moral reasons by drawing a distinction between explanatory reasons that make an attitude or action intelligible, and normative reasons that guided an agent's activity by offering considerations in support of or against actions. Even though this distinction by Bagnoli is not taken to be mutually exclusive, it provides a pragmatic basis for understanding how reasons have rational bite in different contexts where reasoning is typically used. It also offers a basis for distinguishing between subjective and objective reasons, which is useful for understanding how one agent can have authority over another, based on the assumption that normative reasons can be founded on arguments from authority. These distinctions are relevant to legal reasoning because they can lead to a better understanding of norms of basic rationality that can help a citizen to deal with conflicts of moral reasons, a kind of problem widely confronted in legal argumentation. This problem raises the question of how such common legal conflicts can be resolved in a legal setting by logical arguments. For example, they raise the question of how evidence-based reasons can help legal adjudicators decide outcomes in cases where there are moral reasons on both sides of a conflict.

Chapter "Legal Reasoning and Argumentation" of part I shows how the arguments on both sides of a case can each be based on evidence, such as witness testimony and so forth, that can support rational arguments and that in themselves can represent rational arguments. This treatment of such arguments by Douglas Walton in the chapter lends support to Wigmore's view that there is some kind of science of proof apart from deductive logic that underlies legal reasoning. The examples treated in the chapter show how such arguments are evidence-based. This basis in evidence gives us a structure for analyzing and evaluating how legal argumentation works in general (and, more relevantly for this Handbook, should work) as applied to particular cases. The analyses of the examples of legal reasoning in this chapter show us how to apply typical defeasible argumentation schemes, such as argument from witness testimony, argument from expert opinion, abductive reasoning, and so forth, to arguments put forward on either side of a contested case. It also shows us how to analyze and evaluate sequences of argumentation by chaining together such individual arguments based on schemes in a context such as that of a trial. Additionally, Chapter "Legal Reasoning and Argumentation" shows how the argumentation in such a sequence has three stages, an opening stage, an argumentation stage, and a closing stage. The middle part, that of the individual arguments making up the chain of pro and con argumentation, represents the reasoning used in a case, whereas the other two parts represent argumentation in the fuller the dialectical (procedural) sense of the word. They provide essential parts of the pragmatic aspect of the argumentation, taking us from the burden of persuasion at the opening stage to the decision made during the closing stage.

The concept of an autonomous rational agent carrying out an intelligent goal-directed action is fundamental to computing, especially in multiagent systems and robotics, and to understanding legal reasoning and argumentation. Yet little has been done to apply work on action theory to legal reasoning and argumentation in trials and other legal settings. In Chapter "Norms in Action: A Logical Perspective," Emiliano Lorini provides a clearly written survey and introduction covering some most important results in this field that can be applied to legal reasoning. Drawing on rich historical sources, and the formally developed logical systems of norm and action that can be found in the writings on action by logicians and philosophers, Lorini explains the state of the art on the most promising development in this area, the so-called *STIT*, or the logic of "seeing to it that." This logical model represents the basic idea of a rational agent bringing it about that a particular proposition is true or false by means of carrying out an action. The underlying idea is that this concept can be modeled as an agency operator in a modal logic system using a Kripke-style semantics. The formal semantics of *STIT*, which is both elegant and intuitively understandable, can be applied to almost any example of legal argumentation of the kind found typically in the courts in any jurisdiction. It offers a logical structure for framing reasoning about choices, actions, and time that is easily applicable to the evidential reasoning in legal cases.

The logical formalization of *STIT* has become an intricately built framework in recent years, and Lorini provides the service of presenting an outline of the main results and applications of the system, showing how they can be applied to the formalization of such key legal notions as responsibility and influence. It is shown, for example, how *STIT* can be applied to the type of responsibility that consists in one agent inducing another agent to violate a certain norm so that the influencer becomes indirectly responsible for the norm violation, and is subject to a sanction. This connection leads to Chapter "Of Norms," which is on the role of norms in legal reasoning.

Norms are, to put it briefly, social and/or legal requirements that separate actions into three categories, those that are required (obligatory), those that are permitted, and those that are forbidden (prohibited). Chapter "Of Norms," by Jaap Hage explains that there are different theories of norms, and different usages of the word "norm" in English, and corresponding terms in other languages. A norm is sometimes described as a prescriptive guide to acting as a command that empowers, proscribes, or allows actions. Obviously, norms are very important in law and ethics. Recently, study of norms has become important in the development of multiagent systems in artificial intelligence, and new formal argumentation systems incorporating norms have thrown light on how reasoning is based on norms in law, ethics, and other applications.

In particular, Hage distinguishes between two kinds of norms, one that tells us what to do, and one that informs us what ideally should be a case. Hage concentrates on norms that have the function of guiding human behavior, clearly a mainstream concern in legal reasoning. He shows us how norms are closely related to reasons for action and clarifies the distinction between norms and facts by distinguishing among various kinds of facts, the way this term is conventionally used. Hage explains how norms are related to what are called possible worlds in the standard semantics for modal logic. He explains how norms are related to duties and obligations. He clarifies the notion of a norm by defining it as a rule that leads to deontic consequences. This chapter is fundamentally important for understanding legal reasoning because it brings out not only how norms are fundamental to legal reasoning, but also how our reasoning with norms can be modeled by the form of modal logic called deontic logic. In Chapter "Deductive and Deontic Reasoning" of part II, Antonino Rotolo and Giovanni Sartor present an introduction to deontic reasoning, the logic of norms.

What are values, and how can we identify them in particular cases? These are the questions addressed Carla Bagnoli, in Chapter "Values," shows how theories of value answer questions such as how we can judge the adequacy of a theory of value, and how values can have a normative capacity to guide action. She shows how theories of values offer answers to these questions, and can throw light on some disagreements in legal cases that depend on arguments about the incommensurability of values.

Values have become especially noticeable as a current topic in artificial intelligence and law now that it has been shown that practical goal-directed reasoning is not always purely instrumental, but is commonly based on values as well, especially in a legal setting. However, philosophical questions remain about the nature of values, and how they play an especially important role in legal reasoning.

Chapter "The Goals of Norms," authored by Cristiano Castelfranchi, is about the relationship between norms and goals. One immediate connection that he identifies at the outset and brings out in this chapter is that norms are artifacts for social coordination through a rational agent's manipulation of its own or others' goals (whether the agent is a machine or a human). This interconnection is shown by Castelfranchi to go both ways. Norms are used in the legal reasoning of goal-directed agents whose actions depend on their free decisions, but norms also have goals and that they are built by agents and used for something. They are societal tools. Norms depend on goal-directed practical reasoning because they have the function of trying to actuate the intended effects corresponding to goals. Goal-directed practical reasoning is often called teleological reasoning, referring to an agent's purpose in carrying out an action. One does not have to go very far into this network of concepts to appreciate how fundamental they are to understanding legal reasoning.

Castelfranchi lists six main structural relations between norms and goals. First, norms are designed to influence autonomous goal-directed actions of a rational agent. Second, norms presuppose the postulation of goals in the mind of an agent carrying out an action. Third, norms are aimed at governing our conduct and often give us new reasons for or against a goal. Fourth, norms are internalized and adopted for a goal that an agent has. Fifth, norms have goals, because they are aimed at bringing about certain outcomes. Sixth, norms are based on collective expectations about the goals of other agents.

An especially helpful part of this chapter on the use of the term "goal" in legal reasoning is that it explains how in modern science there are two different theoretical approaches to the concept of a goal. One is provided by evolutionary approaches, while the other is provided by the control theory of cybernetics. Bringing these two approaches together and explaining both of them is very helpful for seeing where the study of goal-based practical reasoning in artificial intelligence and law is going. Goals, motives, and intentions are fundamental notions in legal reasoning, especially in criminal law. By showing how these concepts are related, and in turn related to norms, this chapter throws considerable light on fundamental concepts of legal reasoning. In legal reasoning, goals are typically based on values, producing the kind of reasoning called value-based reasoning, as opposed to purely instrumental reasoning.

Authority has long been recognized as an important concept for law, but now with the recent literature in artificial intelligence on evidential reasoning based on authority, such as expert opinion evidence in trials, the concept of authority has become more important than ever for understanding how legal reasoning works. In Chapter "Authority," Kenneth Einar Himma draws a distinction between two kinds of authority. Epistemic authority is the source of reasons to believe that a proposition is true or false, acceptable or not, based on the evidence. Practical authority is the source of reasons for action. This chapter is concerned with the notion of practical authority. The chapter identifies properties that make something of practical authority, explains the kinds of reasons that bind subjects of this kind of authority, and examines what conditions standards of practical authority must satisfy to be morally legitimate.

In Chapter "The Authority of Law," Veronica Rodriguez-Blanco poses the philosophical question of how a person as a rational agent can be in control of her own destiny, given that law requires us to carry out innumerable actions which we freely and intentionally perform all the time. To approach this problem, she focuses on the agent and works upward from the practical reasoning of an individual agent to the framework of authority. Instead of trying to explain human actions as being exclusively empirical phenomena, she perceives the need to understand human action in a more fundamental form, seeing it as it operates in a framework of human institutions such as law. This ties in with the need for paying further attention to action theory by applying logical models such as STIT to study the notion of a rational agent carrying out an intelligent goal-directed action, concept one that is fundamental to both computing and legal reasoning.

Drawing on the philosophical literature on intention, she points out that intentional action involves knowledge that is not of an observational kind, even though it might be expedited and supported by observations. On this approach, understanding an action according to reasons or intentions should begin by way of asking a why-question. On this way of viewing an explanation of a person's actions, we grasp it from the person's own description of his action given as an answer to a why-question. This dialectical multiagent approach to the evidential basis for reasoning to intentions is compatible with recent work on legal argumentation.

In Chapter "Deductive and Deontic Reasoning" of part II, Antonino Rotolo and Giovanni Sartor present an introduction to deductive and deontic reasoning, as these formalisms apply to legal reasoning. Since there are already many formal systems of deontic logic, this chapter starts with deductive logic. The formal systems of deontic logic took were developed a framework of deductive logic (classical logic), so to understand much about formal systems of deontic logic, you have to start with deductive logic. But from there, the chapter offers an account of the basics of deontic logic that uses simple examples that are easily transferable to a legal context, so that the reader can easily appreciate and understand how it works in law.

The chapter explains how statements about obligations and permissions work as modal operators in a classical system of deductive modal logic and how such systems apply to conflicts of obligations and permissions of the kinds familiar in law. The chapter outlines the basics of the Kripke semantics as it applies to deontic logic, and from there outlines some axioms and theorems in the system as they apply to common logical inferences of the deontic kind using ordinary legal examples. The logical system is shown to be sound and complete. An interesting feature is that the chapter explains why more advanced normative notions, such as the notion of a right, cannot be exclusively built on the basis of obligations and permissions, because rights can only be analyzed by making reference to the interests of an agent.

Now that Chapter "Deductive and Deontic Reasoning" has covered how deductive logic applies to legal reasoning, Chapter "Inductive, Abductive and Probabilistic Reasoning" proceeds to investigate the role of probabilistic reasoning in law. Probability is always a difficult and contestable subject on the issue of how it applies to legal reasoning, especially when those of us without specialized knowledge of probability theory and the Bayesian axioms try to apply them to real arguments. By using common legal examples to illustrate how it works, and how examples of its application have been subject to interpretation and controversy, Chapter "Inductive, Abductive and Probabilistic Reasoning" is especially valuable for those of us who are uncertain about just how far Bayesian probability can go in analyzing and evaluating the kinds of evidential reasoning commonly found in trials.

In Chapter "Inductive, Abductive and Probabilistic Reasoning," Burkhard Schafer and Colin Aitken survey the history of the relationship between probabilistic reasoning and jurisprudence, showing how the emergence of the subjectivist view of probability has come to be of pivotal importance. On this view, probability represents the subjective degree of belief of a proposition. This view has turned out to be particularly important for legal reasoning because the inferences drawn by jurors are based on background knowledge and common sense assumptions that are difficult to reduce to objective statistical propositions.

As things have turned out, the most widely explored route to try to find a satisfactory application of probability to legal reasoning in a broad majority of cases has been the subjective Bayesian approach based on Bayes' theorem. There have been many differences of opinion on how to apply Bayes' theorem to areas of legal reasoning, such as evidence based on witness testimony, or the arguments from precedent and counter-arguments attacking these arguments. By using a number of relatively clear and simple ordinary examples throughout this chapter, Schafer and

Aitken explain how probabilistic reasoning and abductive reasoning, the latter usually associated with inference to the best explanation, relate to ongoing jurisprudential debates and match forms of argument commonly used in legal reasoning.

In Chapter "Defeasibility in Law," Giovanni Sartor surveys the leading formal and computational theories of defeasibility that have been prominent in artificial intelligence and law, but at the same time shows how the idea of defeasible reasoning can be traced back to Aquinas and Leibniz, and even to Cicero. Sartor explains how the process of defeasible reasoning reflects the natural way in which legal reasoning proceeds, given that law is applied to particular situations, typically in cases where conflicting legal rules may apply, so that adjudication must work with conflicts between the rules as they apply in a case. This chapter approaches defeasible reasoning from an argumentation point of view, where the evidence is evaluated by a process of critical questioning along with weighing pro and con arguments that are relevant within a framework where there is a burden of persuasion is to decide the outcome.

Sartor visually represents the argumentation in a number of interesting examples of legal reasoning by using argument diagrams. The examples, once analyzed, show some interesting features of legal argumentation. One is that when new arguments are introduced into what is called an argument framework representing the set of arguments constructible from a given set of premises, the status of a given argument can change relative to that framework. For example, an argument that was previously justified in the framework can now be overruled. Structuring argumentation in this way, an argumentation framework can be used not only to evaluate a complex sequence of argumentation taking the form of an argument graph or argument diagram. It can also be used to extrapolate the argumentation forward to support or attack an ultimate claim to be proved. By this means, Chapter "Defeasibility in Law" provides an overview of how argumentation systems provide dialectical frameworks representing the pragmatics of legal reasoning as well as its semantics. This capability is fundamental to understanding how legal reasoning works in a case-based setting.

In Chapter "Analogical Arguments" (*Analogical Arguments*), Bartosz Brożek begins by surveying the topic of argument from analogy from Greek philosophy to the recent tradition in philosophy of science and psychology that portrays analogy as a kind of cognition. To illustrate uses of argument from analogy in science, everyday conversational reasoning and legal reasoning, Brożek analyzes a series of examples that shows how analogical reasoning works, taking an argumentation approach in which problems give rise to certain kinds of questions, notably in some cases open-ended questions. Taking a formal approach, he shows how, once the problem situation has been identified, logical argumentation proceeds by retrieving a set of previously decided cases that are similar to the problem situation in certain respects.

From that point, the chapter applies the formal analysis to some well-known legal cases such as the case of Adams v. New Jersey Steamboat Company. Of special interest to the readers of this Handbook is the analysis of how relevant similarity works as a key component in legal analogical arguments. This is shown by using the same extended legal examples and fitting them to the theoretical framework for argument from analogy. It is shown that there are two widely accepted general methods of evaluating legal arguments from analogy, one called the theory-based approach and the other called the factor-based approach. These two approaches are combined into a general structure for analogical arguments that can be applied to cases legal argumentation, L as illustrated by the examples analyzed in the paper. One thing that comes out clearly in the chapter is the need to take the dialectical dimension of analogical arguments into account in order to adequately model their uses in legal reasoning.

In Chapter "Choosing Ends and Choosing Means: Teleological Reasoning in Law," Lewis A. Kornhauser provides a theory to explain the process of teleological reasoning by articulating its nature in relation to rational choice theory. Teleological reasoning is basically goal-directed reasoning by a rational agent who could, for the purposes of this Handbook, be either a machine or a human, and can comprise a group of agents forming a team for the purpose of deliberating on what to do. Teleological reasoning is identified in argumentation studies by the argumentation scheme for practical reasoning, stating that if an agent has a goal, and knows of the means to carry out that goal, then other things being equal, the agent should go ahead and carry out the action that is the means. Teleological reasoning, in this sense, is a defeasible kind of argumentation subject to default if critical questions are asked when new information comes in, such as the question of whether alternative means are available, or the question of whether carrying out the action would have negative consequences for the agent. According to Kornhauser's account, a rational agent must pay attention to what aspects of consequences are involved as well.

One important feature of teleological goal-directed legal reasoning according to Kornhauser is that legal goals do not need to always be moral ones, meaning that they have to be based on values as well as goals. An important observation made is that legislatures often have to operate instrumentally when they decide which legislation to enact and promote through statutes. An example he cites is the enactment of the Clean Air Act by the US Congress on the basis of the reasoning that the passing of the act will have the consequence of reducing pollution. Cases like this show a process of stepwise reasoning from the actions of one agency to those of another. For example, one institution sketches a goal, a second institution elaborates the goal, and a third agent has the task of implementing the goal, say by framing and implementing a law.

It is becoming more and more obvious in artificial intelligence and law how teleological reasoning is both widespread in legal institutions and how it is fundamentally important generally for understanding how legal reasoning works in practice. But there is already such an extensive literature on consequentialist theories of value and ethics and on rational choice theory as a framework of goal-directed decision-making that it is intimidating to most readers without a background in these areas to get any clearer idea of what teleological reasoning is and how it applies to law and legal decision-making. This chapter uses simply explained examples that guide the reader through this bramble bush of interrelated writings and theories with the clarity that enables him or her to see what elements of them can be helpful to clarify legal teleological reasoning.

There is a special type of decision-making called interactive decision-making which takes place in multiagent settings where what each agent does is dependent on not only with the other agents do but also on their expectations of what the other agents in the decision-making group can be expected to do. Wojciech Załuski, in Chapter "Interactive Decision-Making and Morality," explains how interactive decision-making called strategic decision-making in game theory can contribute not only to moral philosophy but also to our understanding of legal reasoning as a goal-directed form of argumentation.

He begins by outlining the assumptions put to work in classical game theory, showing how there can be stronger and weaker assumptions about the knowledge that the players have and the degrees of rationality that they and the other players can be assumed to have. This strategic approach assumes that the players can make mistakes, and that each player can take advantage of the mistakes made by the other players. This approach has advantages for a good fit to analyzing legal reasoning in an adversarial system, bringing out important aspects of the argumentation that traditional theories of legal reasoning often tended to overlook or minimize. Załuski presents some basic examples of classical game theory that throw light on these aspects. It is also shown in this chapter how game theory can work as a tool for criticizing moral assumptions and theories of the kinds often applied in legal argumentation. Particularly important is the distinction between instrumental teleological rationality and value-based teleological rationality, notions that permeate legal reasoning and deliberation.

Chapter "Evidential Reasoning" of Part III is a general survey of the main problems of evidential reasoning that have been studied in artificial intelligence and law, and explains the leading theories that have been proposed as a way of solving these problems or moving ahead to provide a more unified account of legal reasoning by connecting the theories or even merging them. An attractive feature of this chapter is that even someone with a limited background in artificial intelligence or logic, or related technical subjects, can apply it to understanding how the nuts and bolts of legal and logical reasoning are put together. Common examples of legal reasoning are used, mainly from criminal law, and argument diagrams are presented so that the reader can visualize the basic structure of the reasoning in each case easily and clearly. In each case, an explanation enables the reader to understand how tools from artificial intelligence can be applied.

Di Bello and Verheij cover such basic kinds of evidence as witness testimony evidence, and scientific expert testimony evidence, such as DNA evidence of the kind that has now become so common in criminal cases. They go on to how we should understand conflicts between pieces of evidence, how we should evaluate strength of the evidence, how we should interpret the available evidence, how we should decide about the facts given the evidence, and it should be decided that an investigation has been exhaustive enough so that the closing stage of a criminal proceedings is reached.

An extremely useful part of the chapter is the outline of the three normative frameworks that have been put forward as systematic and well-regulated methods for examining, analyzing, and weighing the evidence in a case. These are the argumentation framework, the probability-based framework, especially Bayesian methods, and the scenario framework that sees the determination of the outcome and a legal case, for example in a criminal trial, as an argumentation-based rational decision between competing stories. The various leading problems in applying these normative frameworks to everyday legal argumentation are explained and discussed, and suggestions are made on how to solve them.

Chapter "Interpretive Arguments and the Application of the Law," by J. J. Moreso and Samuele Chilovi, surveys the literature on current theories of how to interpret the law, addresses criticisms of them, and present their own theory. The first theory considered is the communicative content theory of law, which holds that legal interpretation is modeled on utterance interpretation. This view holds that facts about the nature of language, taken together with facts about the nature of law, are all that is needed to drive a legal interpretation as a conclusion and tell whether it is correct. However, Moreso and Chilovi hold that there is a conflict between this theory and the doctrine of the rule of recognition which suggests that the communicative content theory is problematic.

The next theory considered is the communication theory of law, which holds that legal content is determined in the same way that propositions and other elements of linguistic texts are interpreted in ordinary language. This theory uses what is called a principle of epistemic asymmetry according to which the producer has a message he wants to get across is a particular form of words, and the consumer operates on the assumption that the producer meant something. If the consumer interprets the producer correctly, then the consumer is taken to have succeeded in identifying what the producer meant.

The difficulty with the communication theory of law, according to Moreso and Chilovi, is that it requires the consumer to select among the various intentions that the speakers might have and select the one that is relevant for the legal application. The problem with this approach, they contend, is that it remains unclear what the object of interpretation precisely is. The difficulties are that there is the Gricean problem of meaning something without saying it, using implicature. Another problem is that there can be a communication failure where the rational here takes the speaker to have meant something different from what she actually said. This is shown in detail by an examination of the Gricean maxims as they might be applied to problematic cases of legal interpretation.

After a close examination of these theories, resting on a series of examples, Moreso and Chilovi, put forward their own theory as an alternative that, they claim, can reply to all the objections they encountered in treating the existing theories. According to the theory of Moreso and Chilovi, the existing legal theory already has an assemblage of types of arguments, such as argument from analogy and argument *a contrario*, that can be applied to norms to facts to generate an interpretation (such as one of a statute) using only deductive reasoning. From this premise, they conclude that no form of logic other than classical deductive logic is needed to provide a logical structure to underpin their theory of interpretation, except for an extension of deductive logic to deontic logic.

A valuable feature of this chapter is that it reveals significant weaknesses in the leading traditional theories of legal interpretation, and thereby provides an interesting survey of the theories themselves in the difficulties inherent in them that would enable the study of statutory interpretation to move ahead.

In Chapter "Statutory Interpretation as Argumentation," Douglas Walton, Giovanni Sartor and Fabrizio Macagno show how the traditional canons of interpretation can be represented as argumentation schemes that are defeasible forms of argument pro or con the interpretation of a given statutory or legal text. The formalization of the schemes given in the chapter makes possible the modeling of legal interpretation using formal argumentation systems from artificial intelligence. After introducing some of these formal systems and applying them to two cases, the chapter develops a logical model for reasoning with interpretive canons from a text using defeasible rules to draw an interpretive conclusion.

The chapter begins with a list of eleven interpretive arguments, including argument from ordinary meaning, argument from technical meaning, argument from contextual harmonization, argument from precedent, argument from analogy, argument from a legal concept, argument from general principles, argument from history, argument from purpose, argument from substantive reasons, and argument from intention. The names themselves roughly indicate the nature of each type of argument. This list of eleven types of argument which can be used to support or attack a legal interpretation is compared to an overlapping list of fourteen types of arguments previously identified in the literature. The chapter shows how these interpretive arguments can be classified into subtypes and how each of the schemes representing them need to be formulated so that in this form can be used to derive interpretations in several key examples. These interpretative schemes provide ways of dealing with vagueness and ambiguity in law.

In Chapter "Varieties of Vagueness in the Law," Andrei Marmor distinguishes between different kinds of vagueness in law and explains some of the ways in which legal decision-makers reason with the language. Slippery slope arguments of the kinds one finds in law occasionally, sometimes turn out to be very controversial because they depend on the vagueness of a key legal term. Vagueness, or open texture as it has been called in legal contexts, is inevitable both in ordinary language and in legal communication and reasoning. Legal reasoning itself takes place in natural language, and so as Marmor shows, law cannot entirely avoid linguistic vagueness, even though it has ways of dealing with it. Terms such as "reasonable care," "due process," and so forth, can be made more precise for legal purposes by precedents and criteria set by law, but the inherent vagueness in them is unavoidable because all natural language terms are open-textured. There are always going to be borderline cases. Vagueness is something that case-based legal reasoning can deal with, and it has to contend with on an ongoing basis.

Marmor draws a distinction between semantic vagueness, which concerns the relations between the meanings of words and the objects they apply to, and conversational vagueness, which has to do with borderline cases and relevance. Both kinds of phenomena occur in legal reasoning. As Marmor shows, the normal procedure in law when regulating with vague standards is to put the decision for sanctions for violation to the courts so they can decide whether the standard was violated or not in a given case. By this means, the precedent meaning set by the courts makes the standard less vague in a certain respect.

Marmor shows that this procedure of using legal reasoning to make a standard more precise is context sensitive, and hence is a matter of pragmatics in linguistics (the study of meaning that takes contextual factors into account when drawing implications about what a word or phrase may be taken to mean in a specific instance of its usage). A pragmatic approach is necessary for statutory interpretation, as shown by Walton, Sartor, and Macagno in Chapter "Statutory Interpretation as Argumentation."

In Chapter "Balancing, Proportionality and Constitutional Rights," Giorgio Bongiovanni and Chiara Valentini explain how and why proportionality review is a widespread decision-making model that lies at the core of the debates on rights in education, where it has raised questions about the nature and distinctive features of legal reasoning. In this chapter, they examine the relation of different forms of proportionality to the balancing of rights. This chapter explains how conflicts of interests which lie at the foundation of rights illustrate the need for legal reasoning of a kind that has the capability to distinguish between principles and rules.

The chapter reviews several leading theories that propose models of constitutional rights, revealing that they show the need to apply the canon of proportionality, requiring an approach in which laws operate on different hierarchical levels. Such an approach is shown to require an account of value-based legal reasoning whereby value can be based on the interests of an agent. On this view, if the agent has an autonomy interest in an activity, that activity must be protected by a right. The most influential model of proportionality–balancing holds that constitutional rights need to be treated as defeasible principles that may conflict with other rights or interests, where such conflicts need to be adjudicated by a process of optimization. Taking this approach, it is shown how justification of proportionality review is the legal instrument that has been and needs to be adopted by the courts. It is shown how proportionality takes two fundamental forms, optimizing proportionality and state-limiting proportionality. Several other alternative approaches to proportionality–balancing are considered as well.

In Chapter "A Quantitative Approach to Proportionality," Giovanni Sartor addresses the extent to which the operations involved in balancing and proportionality assessments may include quantitative reasoning, and be subject to arithmetic constraints. Relying on some work on cognitive and evolutionary psychology he argues that processing non-symbolic approximate continuous magnitudes is a fundamental cognitive capacity, which seems to be deployed also when we are reasoning with values, as scalable goals are being pursued. A model is proposed for determining the impact of a choice on different values, assessing the utilities so produced and merging these utilities into an overall evaluation, which may be used in comparisons. The usual standards deployed in proportionality assessments, such as suitability, necessity, and proportionality in a strict sense, are specified relatively to this model. Finally, it is discussed how proportionality assessments can lead to the formulation of rules, and how quantitative proportionality assessments may be constrained by the requirement of consistency with precedents.

Coherence has recently been revived as an alternative to foundationalist theories of justification and as an alternative to the Bayesian model of reasoning, both in psychology and philosophy. Coherence has also been appealed to legal theory as a standard of rational justification. The three main kinds of theories of normative coherence are reviewed and evaluated: principle-based theories, case-based theories, and constraint-satisfaction theories. In Chapter "Coherence and Systematization in Law," Amalia Amaya focuses on normative coherence as dependent on legal coherentism. She addresses the problem of coherence bias, formulating this problem in detail, and argue that a modified version of coherence bias can address this problem. The chapter comparatively evaluates the value and limits of coherentist reasoning in law.

In Chapter "Precedent and Legal Analogy" of part III, Kevin D. Ashley shows how argumentation schemes representing argument from analogy and arguments from precedent can be applied to structure arguments employed in court opinions whether or not to apply precedent or a legal analogy in specific cases. This exercise is valuable for helping law students, legal professionals, and theorists of legal reasoning to both support arguments and to attack them in a carefully reasoned manner. Chapter "Precedent and Legal Analogy" builds on recent work in argumentation. Artificial intelligence has provided a repository of argumentation schemes that can provide a prima facie reason tentatively accepting the conclusion of an argument based on the acceptability of its premises. Such schemes can be used to link arguments together, so that a connected network of such arguments can be evaluated by weighing the pro arguments against con arguments using formal argumentation systems from artificial intelligence.

Chapter "Economic Logic and Legal Logic," written by Lewis A. Kornhauser, compares legal reasoning to the kind of reasoning used in economics and uses this comparison to argue that the latter deepens our understanding of the former. He shows how economic models abstract from the details of complex social phenomena. He suggest that for this reason, it is not surprising that economic conclusions are generally arrived at using mathematical models that rely on deductive reasoning and statistical probability. They have also tended to require that each agent in the decision-making situation as a complete ranking of the outcomes. Therefore, although economic reasoning, like legal reasoning, is based on goal-directed means-end reasoning, it typically concentrates on the means and takes the agent's goals as given. For these reasons, economists have tended to shy away from value-based practical reasoning and do not take used to take the rational agent's values into account.

However, behavioral economists now recognize that agents systematically deviate from these strict rationality assumptions, and this departure presents prospects for the application of argumentation to the study of our reasoning takes place when a (somewhat) rational agent carries out an action such as choosing to buy or sell something. Argumentation would take the approach that this procedure works by the agent deliberating on both sides of the issue using pro–con argumentation, where the sequence of argumentation on both sides is based on argumentation schemes. Argumentation schemes can be deductive or probabilistic in the statistical sense in some instances, but as shown in this Handbook, studies in argumentation have recently shown that in cases of this kind, deciding whether or not to buy something for example, agents use defeasible argumentation schemes, such as the scheme for goal-directed practical reasoning and the scheme for argument from negative consequences. For these reasons, judging from what Kornhauser has shown in this final chapter of Part III, there is a brave new world out there ready to explore how argumentation applies to economic reasoning.

Douglas Walton

Part I
Basic Concepts for Legal Reasoning

Reasons (and Reasons in Philosophy of Law)

Giorgio Bongiovanni

1 Premise

Notwithstanding the fact that reasons have been at the centre of the reflection on normativity and action for at least forty years,[1] there is not complete agreement about the concept and the features of reasons. What is a reason, what kinds of reasons there are, what their different qualifications are, and so on are the subject of a wide discussion.[2] As noted by J. Searle (referring to P. Foot)[3] what is a reason for action seems "to be frightfully difficult," even though "we deal with reasons for action every day" (Searle 2001, 97).[4] This difficulty has different sources, ranging from the "heterogeneity in the use of the term 'reason'" to the different meanings we can attribute to it. As Alvarez (2010, 8) notes, "it is common," in dealing with these questions, "to introduce the discussion by drawing a distinction between different *senses* of the term 'reason,' or between different *kinds* of reason."[5] But distinguishing between reasons can be done in different ways, with different points of reference,

[1] Broome (2004, 28) notes that "within the philosophy of normativity, the 1970s was the age of the discovery of reasons."

[2] For Alvarez (2010, 1), the analysis of reasons raises the following questions, among many others: "What are reasons? Are there different kinds of reasons? Are reasons beliefs and desires? If not, how are they related to beliefs and desires? And what role do they play in motivating and explaining actions?"

[3] Searle (2001, 97) relates that Philippa Foot once wrote, "I am sure that I do not understand the idea of a reason for acting, and I wonder whether anyone else does either."

[4] Dancy (2000, 1), underlines that "There are not so many things that we do for no reason at all. Intentional, deliberate, purposeful action is always done for a reason, even if some actions, such as recrossing one's legs, are not—or not always, anyway."

[5] Alvarez (2010, 7) notes that "this territory is […] quite complex. And with it comes the temptation to suppose that the machinery required to find one's way around it must be correspondingly complex."

G. Bongiovanni (✉)
Dipartimento di Scienze Giuridiche and CIRSFID, Università di Bologna, Bologna, Italy
e-mail: giorgio.bongiovanni@unibo.it

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_1

and it is not easy to find a common denominator.[6] A possible and general initial point is that of differentiating reasons according to their roles in action and reasoning. This analysis, which starts from the awareness that the same reason may play different roles in different contexts,[7] seems to be prodromic to the classification of reasons, it does not commit one directly to an ontological point of view (although it is functional to a unified view of reasons), and it enables us to "explore reasons broadly" (and after focusing on the different kinds of reasons). In this way, this analysis appears useful, as it makes it possible to provide a first classification of reasons and specify the different questions that an analysis of reasons poses, and, at the same time, it could make it possible to order the questions on a different level of generality.

The following sections will analyse (a) the definition of the different classes of reasons (normative, motivating, explanatory) in relation to their role; (b) the problem of the ontology of reasons (facts or mental states); (c) the kinds of reasons and in particular the distinction between reasons for belief (epistemic) and reasons for action (practical); (d) the modality and the strength of reasons; and (e) reasons and the law.

2 The Different Classes of Reasons: Normative, Motivating, Explanatory

Reasons play different roles in our life: even at a glance, we can see how they can be used to guide, motivate, justify, understand and evaluate our behaviours and believes. From this perspective, it is possible to note an assortment of roles that reasons can play, namely, to "motivate and guide us in our actions (and omissions)"; to "be grounds for beliefs, desires, emotions, etc. […] to evaluate, and sometimes to justify, our actions, beliefs, desires, and emotions"; and to be "used in explanations" (Alvarez 2010, 7).[8]

A general distinction often introduced in contemporary literature to summarize these different roles is that between *normative* and *motivating*[9] reasons: the former are the kind that, as is well known, favours or "counts in favour of" a specific behaviour or acting in a specific context,[10] while motivating reasons are the kind "for which

[6] Alvarez (2010, 10) stresses, in this sense, that there are "different ways of partitioning."

[7] Alvarez (2016a) argues that "in itself and out of context, a reason is not a reason of any particular kind, say normative or motivating. It is only in a particular context, where the reason plays a specific role and can be cited to answer a particular question that it can be qualified as being of this or that kind."

[8] In the same sense, Raz (1999a, 15–16) notes that "as well as reasons for actions there are reasons for beliefs, for desires and emotions, for attitudes, for norms and institutions, and many others […]. Reasons are referred to in explaining, in evaluating, and in guiding people's behaviour."

[9] Sometimes, the distinction is placed between *normative* and *explanatory* reasons. See Bagnoli, chapter 2 , part I, this volume, on "Reasons in Moral Philosophy," para 2.

[10] Of course, this idea of reasons as "favouring" comes from Raz (1999a) and Scanlon (1998). Broome (2004, 41), who sees as "a commonplace" the idea that "the reasons for an action are considerations which count in favour of that action," denies (2013, 46ff.) the primacy of reasons

someone does something, a reason that, in the agent's eyes, counts in favour of her acting in a certain way" (Alvarez 2016a). This distinction, largely present in different authors (see, e.g., Dancy 2000; Raz 1999a; Parfit 2011), can be further articulated considering that motivating reasons can play a double role: reasons that motivate someone to act can also be one that, in a third-person role, could also be used to explain a behaviour. In this way, it is possible to arrive at a "three-part classification of reasons:" *normative*, *motivating*, *explanatory* (Alvarez 2016a). The need to distinguish between motivating and explanatory reasons, despite the fact that "one might think, fundamentally," that they are "the same," can be seen in the fact that "even if the same reason sometimes answers the two questions about motivation and explanation, this is not always so" (ibid.). Although a reason that motivates an action can always explain it, a reason that can explain the action is not always the reason that motivates it" (ibid.).[11] From this perspective, as Audi (2010, 273ff.) remarks, we can individuate "three overlapping kinds" of reasons: "normative reasons—which include moral reasons as a major subset—motivational reasons, and explanatory reasons": "Normative reasons are reasons (in the sense of objective grounds) there are for doing something"; motivational reasons are "reasons someone has to do something"; explanatory reasons "are reasons why someone acts."[12] Following a suggestion by Alvarez (2016a), referring to Dancy (2000, 2ff.), this distinction between reasons can be seen as the way in which we can answer the different questions that reasons pose in relation to action and reasoning: in this way, normative reasons will be those that answer the question of "whether there was good reason to act in that way [...], *any reason for doing it* at all"; motivating reasons will be those that answer questions "about *his reasons for doing it*" (Dancy 2000, 2); and explanatory reasons will be those that answer a general "reason why" different "sorts of things"—like "the occurrence (or nonoccurrence) of an event; the

in normativity and supports that of the "ought." To pro tanto reasons, he adds the presence of "pro toto reasons" (or "perfect reasons"). He does not "take the idea of a reason as primitive" (ibid., 54), and he does not consider "counting in favour [as] the basic normative notion" (Broome 2004, 41). For Broome, the idea of a pro tanto reason does not account for the explanatory role of reasons. On these aspects, see Crisp (2014). See also note 35 below.

[11] For Alvarez (2016a), "the advantages of drawing this distinction will be spelled out in examining debates concerning motivating reasons and the explanation of action. [...] [A]pparently competing claims about motivating reasons and the explanation of action are often best understood and resolved as claims about motivating or explanatory reasons, respectively. To be precise, a reason that plays a motivating role for a particular action can (arguably) always play an explanatory role for that action, but the converse does not hold." She argues that "This way of categorising reasons [...] enables us to deal with a range of cases that the binary classification cannot accommodate" (ibid.). To explain the difference between the two types of reasons, she uses the example of Othello and Desdemona: while Othello's jealousy can be seen as one of the reasons explaining the killing of Desdemona, the reason motivating Othello is the belief, induced by Iago, that Desdemona betrayed him (ibid.).

[12] Audi (2010, 275) notes that normative reasons can be, for example, moral or prudential. He also notes that "some normative reasons for—roughly, counting in favour of—an action are reasons for any normal human being. Other normative reasons, however, are person-specific: reasons there are for a specific person."

obtaining (or non-obtaining) of a state of affairs; someone's or something's φ-ing (or not φ-ing)"—happened or not (Alvarez 2010, 27).[13]

This distinction between normative (reasons that favour), motivating (reasons for which we act), and explaining (reasons why) should be developed not only to clarify the exact role of reasons but also to explore what it means that reasons favour, motivate, and explain.

2.1 Normative Reasons[14]

2.1.1 Normative Reasons and the Source of Normativity

The idea that a reason is normative refers to a general notion of what normative is: *normative* means "prescriptive" in relation "to some norm or value and, by implication, concerning correctness," that is what makes something "right or appropriate" in relation to what is prescribed by "norm or values."[15] Normative is therefore what we can consider "right or wrong with reference to what is prescribed by the […] norm, or what furthers the […] value" (Alvarez 2010, 9),[16] and in this sense, normativity implies the correctness of actions in relation to norms or values.

For a reason to be normative, it must therefore "involve the idea that reasons can be invoked to support claims about what it would be right (for someone) to do, believe, want, feel, etc." Normative reasons are those that "can be invoked to support claims about what it would be right (for someone) to do, believe, want, feel, etc." (ibid.),[17]

[13] Broome (2004, 34) underlines that "a useful distinguishing mark is that 'the reason' in this non-normative sense is usually followed by 'why.'" For Alvarez (2016a), "the basis for doing so was said to be the existence of three distinct questions about reasons: whether a reason *favours* an action; whether a reason *motivates* an agent; and whether a reason *explains* an agent's action. Accordingly [...] we should recognize three kinds of reasons: normative, motivating and explanatory."

[14] Normative reasons are often qualified "in terms of justification: a reason justifies or makes it right for someone to act in a certain way. this is why normative reasons are also called 'justifying' reasons" (Alvarez 2016a). Dancy (2000, 6–7) criticizes this way of marking normative as justifying reasons: "there is a sense of 'justify' in which I can be said to have justified doing what I did. but this does not show that the balance of reasons was in favour of the action. It wasn't. […] So I prefer to keep the notion of a reason that justifies separate from that of a normative reason. Broome (2004, 54), too, notes that "'justify' is an ambiguous word."

[15] Norms and values are understood here in a broad sense, that is, as inclusive of different values (moral, prudential, etc.) and norms (legal, social, principles, codes, etc.). For Alvarez (2016a), "the existence of these norms or values depends on a variety of things": "logical and natural relations, conventions, rules and regulations, etc."

[16] From the same perspective, Alvarez (2016a) notes that "'normative reason' derives from the idea that there are norms, principles or codes that prescribe actions: they make it right or wrong to do certain things." Dancy (2000, 1) sees normative reasons as "good reasons for doing the action. So they are *normative,* both in their own nature (they *favour* action, and they do it more or less strongly) and in their product, since they make actions right or wrong, sensible or unwise."

[17] Some authors note that the idea of favouring is an ambiguous one. Schroeder (2007, 11), for example, notes that the idea that reasons "count in favour […] is slippery." Hieronymi (2005,

or, in other words, they are “a reason [that] justifies or makes it right for someone to act in a certain way” (Alvarez 2016a). In this way, that a “reason can favour φ-ing” means that “it can make φ-ing right or appropriate:” to favour is to contribute to or support (i.e. to “recommend, warrant, demand”) the rightness or appropriateness of an action (Alvarez 2010, 3, 11). Making right or appropriate a certain action can depend on many factors, such as values, norms, codes, rules, desires, purposes, and natural facts.[18] In addition, norms and values especially may refer to different fields and may be “moral, prudential, legal, hedonic (relating to pleasure) or of some other kind” (Alvarez 2016a). Finally, the relevance of these factors may radically change depending on the different contexts.[19]

The normative aspect of reasons is expressed “by talk of ‘a reason *for* φ-ing’ (or ‘a reason *to* φ’)”: in this sense, we have reasons for acting, believing, deliberating, wanting something, feeling an emotion, and so on,[20] and, as noted, “there are reasons for the variety of things where we are, typically, responsive to reasons.”[21] Having a reason for φ-ing is often expressed through the concept of “ought”: saying “that if there is a reason for someone to φ” could be seen as saying “that person [...] *ought* to φ.” This does not mean having a moral ought, nor does it mean that the behaviour favoured by a reason is “obligatory”: what “ought to” amounts to varies from case to case, depending on the circumstances (Alvarez 2010, 11).[22]

437–438, 456–457) makes a deeper critique, stating that “we should not understand ‘counting in favour of an action or attitude’ as the fundamental relation in which a consideration becomes a reason.” She stresses that identifying a “reason as a consideration that counts in favour of an action or attitude [...] generates a fairly deep and recalcitrant ambiguity; this account fails to distinguish between two quite different sets of considerations that count in favour of certain attitudes, only one of which is the ‘proper’ or ‘appropriate’ kind of reason for them.” This ambiguity, which for Hieronomy, leads to the so-called problem of the “wrong kind of reason,” referring to the fact that seeing a reason as “a consideration that counts in favour of an action or attitude [...] generates a thoroughgoing ambiguity in reasons for certain attitudes–we cannot distinguish precisely between ‘content-related’ and ‘attitude-related’ reasons.” Instead of seeing a reason as counting in favour of something, she proposes that we see it “as a consideration that bears on a question [...] in a piece of reasoning.” This makes it possible to “distinguish between questions, thus distinguishing these classes. The attitude-related reasons count in favour of the attitude by bearing on whether the attitude is in some way good to have; the content-related reasons count in favour of the attitude by bearing on some other question.” On this analysis, see also Hieronymi (2013).

[18] It is possible to relate these factors to values and norms understood as reference points for desires and purposes, among other things. However, this relation need not be necessary. In any case, we will refer to values and norms as categories in relation to which something can be considered as having been made “right or appropriate.”

[19] As Alvarez (2010, 11) notes: “Thus, if A ought to φ, the circumstances of each case will determine whether A’s φ-ing is merely recommended, or whether it is also required, or mandatory.” On the way in which it is possible to consider the role of context, see Sect. 5.1 below.

[20] Or, conversely, “reasons for not φ-ing, that is reasons for not doing or for not believing something, for not feeling something, etc.” (Alvarez 2010, 10).

[21] Alvarez (2010, 10), referring to Raz (1999b).

[22] This is true for the positions that consider reasons as the fundamental dimension of normativity. As we have seen, this is not, for example, the position of J. Broome.

This definition (normative reasons as what favour *φ-ing*) can be interpreted in different ways in connection with what can be seen as the source (ground, basis, capacity)[23] of the normativity of reasons, that is, of their ability to favour action. Schematizing and simplifying, we can individuate three main positions: the first *desire-based*, the second *value-based*, and the third *rationality-based*. On the first position, "all reasons for acting, intending, and desiring are provided by the fact that the agent wants something or would want it under certain conditions" (Chang 2004, 56). In this perspective, "all practical reasons are grounded in the present desires of the agent; justification has its source in the fact that I do or would want it" (ibid.).[24] In this sense, desires can be seen as "a necessary condition for a consideration to provide an agent with a reason" (Sobel and Wall 2009, 3). On the second position, "no practical reasons are provided by the fact that one desires something." In "'value-based' accounts, reasons for acting, intending, and desiring are provided by facts about the value of something, where being valuable is not simply a matter of being desired" (ibid., 57).[25] There are two principal versions of this second approach: the "buckpassing versions" and the one linked to idea of the existence of evaluative facts. On the buckpassing version, it is not strictly the evaluative fact that provides a reason but the facts on which the evaluative fact supervenes, while on the second, reasons come from evaluative facts. In this way, "value-based views ground all practical reasons in evaluative facts or the facts that subvene them" (Chang 2004, 57),[26] and "justification has its source not in the fact that one wants something but in facts about what one wants" (ibid.).[27] The third (rationality-based) approach is Kantian, and at the centre of normativity, it places the idea of an autonomous and rational subject: "Kantians [...] hold that rationality is the source of practical normativity." From this perspective, "rational deliberation legislates [...] what to do and an action's being the rational thing to do is what makes it true you ought to do it." As it is well know, "Kantians hold that there are non-instrumental rational deliberative procedures guaranteed to issue true normative conclusions," that is "principles

[23] Different authors mention this aspect in various ways. For example, Robertson (2009, 18) speaks of "source," Alvarez (2016a) of "basis" and "capacity," Sobel and Wall (2009, 3) of "grounds."

[24] Chang (2004, 56) explains this approach in this way: "My reason for going to the store, for example, is provided by the fact that I want to buy some ice cream, and my reason for wanting to buy some ice cream is provided by the fact that I want to eat some." This is a form of "subjectivism," that is a view that assumes the capacity of reasons to favour desires, plans, motivations, projects, etc.

[25] Note that, in this case, "my reason to go to the store is provided by the value of what is in question—namely, eating some ice cream—and the value of eating some ice cream is given by the fact that doing so would be valuable in some way—for example, that it would be pleasurable. It is not the fact that I want ice cream that makes having some pleasurable; having ice cream might be pleasurable even if I don't desire it" (Chang 2004, 56).

[26] Chang (2004, 57) describes the difference in this way: "So, for example, my reason to have the ice cream might strictly be given not by the evaluative fact that it would be pleasurable but rather by the natural facts upon which its being pleasurable supervenes, such as that it would be pleasant or that I would enjoy it."

[27] One of the aspects in which these approaches diverge is that of the internal versus external vision of reasons for acting. On this question, see Sect. 3 below.

or laws that any rational agent could recognize and that thereby apply to any rational being" (Robertson 2009, 11, 19). As noted, "on the Kantian view, ideal rationality significantly constrains what is desired or willed in the authoritative way" (Sobel and Wall 2009, 4).

As we will see, these different approaches (and in particular the desire- and value-based ones)[28] can have different implications with regard to the different roles of reasons: in discussing the normative aspect, one of these aspects is the relation to the concept of rationality (understood as the way to arrive at normative conclusions). Very briefly, it is possible to note that the desire-based approach underlines the relevance of an instrumental (means-ends) rationality and the requisite of the "connections between an agent's various attitudes"[29]; Kantian approaches emphasize a procedural form of rationality[30]; the value-based approach stresses a "substantive conception of rationality" grounded in the idea that "there are substantive normative truths" (Robertson 2009, 20).

2.1.2 The Structure of Normative Reasons

A normative reason has a "relational" structure: "it establishes a relation between a fact, an agent, and an action kind" (Alvarez 2016a). As has been noted, "the concept of a reason is itself relational. Basic reason statements of the form 'A has a reason to φ' or 'there is a reason for A to φ' indicate a relation holding between some agent A and some act φ (e.g. an action, belief, feeling)—a reason is a reason *for* someone and *to* or *for* something" (Robertson 2009, 9).[31]

[28] The Kantian approach can be assimilated to a "subjective" one characterized by the control that the rationality of an "ideally rational agent" imposes on desires and attitudes. See Sobel and Wall (2009, 3ff.).

[29] A model of practical rationality that refers to mental states has been developed in Bratman's theory of action (1987, 2014): this is the belief–desire–intention (BDI) model. This model has as its main reference the concept of plans for action and has been used in the field of artificial intelligence research.

[30] Robertson (2009, 19) states that "Kantian principles of rationality are formal in that they lack specific, substantive, normative content, a practically rational deliberator being one who satisfies relevant formal procedures of reasoning."

[31] Robertson (2009, 12), sees it as "uncontroversial […] that the conceptual structure or logical form of a reason is that of a relation." Raz (1999a, 19) notes that "we usually think of reasons for action as being reasons for a person to perform an action." Blackburn (2010, 6) underlines that "reasons are reasons for something: the primary datum is relational. The field of the relation is less clear, or rather, more diffuse." For Searle (2001, 99), "reason statements are relational in three ways. First, the reason specified is a reason for something else. Nothing is a reason just by itself. Second, reasons for action are doubly relational in that they are reasons for an agent-self to perform an action; and third, if they are to function in deliberation, the reasons must be known to the agent-self. To summarize, to function in deliberation a reason must be for a type of action, it must be for the agent, and it must be known to the agent." Alvarez (2016a) stresses that, for some authors like Skorupski (2010) and Scanlon (2014), "the relation involves not just a person, a reason and an action, but more aspects: a time, circumstances, etc."

What determines the way "in which a reason makes φ-ing right, and hence in which something may be right or justified," depends not only on the different factors (like types of norms and values) but also on "what φ-ing is" (Alvarez 2010, 13). We have significant differences between the case of believing, that of acting, and that of feeling emotions. In the first case (believing), "the rightness or appropriateness of φ-ing [...] concerns the concept of truth"; in the second (acting and wanting), "it concerns the concepts of what is valuable and of the good"; in the third (emotions), making right has to do with the appropriateness/reasonableness of feelings or emotions, that is with their being "fitting or proportionate to the facts" (ibid.). On this basis, it is possible to distinguish, within normative reasons, among epistemic reasons related to the concept of truth; those that are practical, relating to action and deliberation; and those related to "emotions."[32] The distinction between epistemic and practical reasons (for action) is particularly relevant because, as we shall see in Sect. 4, the former do not imply a choice between different values and do not seem "person-relative," while the latter do imply a choice between different values and *are* "person-relative."[33] As we shall see, this distinction is related to the fact that epistemic reasons have a single (or prevalent) reference criterion (the truth), while practical reasons refer to a "variety of values."

Normative reasons are in general defeasible, that is, that they can be "defeated by a reason for not φ-ing" (Alvarez 2010, 12). This means that normative reasons are "pro tanto" (or prima facie) reasons. This indicates that a reason "can" favour an action (belief, act, evaluation, etc.) to a certain extent, but that this possibility is limited by the presence of reasons which instead favour an omission. So "I have a *pro-tanto* reason" to do something "and a different *pro-tanto* reason" not to do it (Alvarez 2016a).[34] As Broome (2004, 41) has suggested, a pro tanto reason refers to "a weighing explanation" and to "the distinction between the for-φ role and the against-φ role in a weighing explanation."[35] There is, however, no agreement that all reasons are pro tanto: the same author argues that, next to pro tanto reasons, there

[32] We will not discuss these reasons. Raz (2009, 47ff.) sees "reasons for or against having an emotion" both as standard (adaptive) reasons ("affect-justifying reasons") and as non-standard reasons. On this distinction, see Sect. 4 below.

[33] The notion of person relatedness can be connected with the distinction between agent-relative and agent-neutral reasons: it is a distinction that was introduced by Parfit (1984) on the basis of the distinction between "objective" and "subjective" reasons elaborated by Nagel (1970). For Parfit (1984, 142), "Nagel calls a reason *objective* if it is not tied down to any point of view. Suppose we claim that there is a reason to relieve some person's suffering. This reason is objective if it is a reason for everyone—for anyone who could relieve this person's suffering. I call such reasons agent-neutral. Nagel's *subjective* reasons are reasons only for the agent. I call these agent-relative [...]. When I call some reason agent-relative, I am not claiming that this reason *cannot* be a reason for other agents. All that I am claiming is that it may not be." On the different ways of understanding this distinction, see Ridge (2011).

[34] Dancy (2000, 5) argues that "there can be good reasons not to do an action even when there are better reasons to do it. That an action was right does not show that there were no (good) reasons not to do it. Equally, if an action was wrong, this does not mean that there was no reason to do it; it merely means that there was insufficient reason."

[35] Broome (2004, 42ff.) is against what he calls "protantism," that is the view that "there is a case for thinking that every ought fact has a weighing explanation." For Broome, "even if every ought fact

are "perfect reasons to φ," in which "to φ is defined as a fact that explains why you ought to φ," that is when there is a direct link between a non-normative (explanatory) fact and the proposition that someone ought to φ. Broome differentiates these two kinds of reasons as follows:

> A perfect reason for you to φ is a fact that explains why you ought to φ. A pro-tanto reason for you to φ is a fact that plays a characteristic role in a potential or actual weighing explanation of why you ought to φ, or of why you ought not to φ, or of why it is not the case that you ought to φ and not the case that you ought not to φ. (Broome 2004, 55)[36]

The same author, however, seems to emphasize that whether a reason is conceived as "perfect" or as "pro tanto" depends on a specific philosophical position (evidentialism vs. pragmatism): this means that it is possible to consider reasons in the two ways. In addition, one can add that a perfect reason can be a reason that does not have a "current" reason for not φ-ing, but that this reason can still be hypothesized. If we hold the hypothesis that normative reasons are, in general, pro tanto reasons, an important consequence arises, that is as we shall see in Sect. 5, that in a process of weighing, reasons will have different strengths, roles, and interrelations; this means that the "final" reason for doing φ (acting, believing, deliberating, etc.) should be considered as an overall, or "all things considered," reason to φ, that is a reason that "overrides" or "defeats" other competing reason(s).

2.2 *Motivating Reasons*

A "motivating reason"[37] is "a reason for which someone does something," that is "a reason that, in the agent's eyes, counts in favour of her acting in a certain way" (Alvarez 2016a). In greater detail, it can be seen as a reason that an agent "took to make his *φ*-ing right and hence to speak in favour of his *φ*-ing, and which played a role in his deciding to *φ*" (Alvarez 2010, 35).[38] In other words, it is "a reason that the agent takes to favour her action, and in the light of which she acts" (Alvarez 2016a). A motivating reason, that is a reason that can be recognized "when an agent acts

does have a weighing explanation, many ought facts also have more significant explanations that are not weighing ones." As noted, Broome denies that reasons are the primary category of normativity. In his view, this space is composed of different types of reasons (perfect, or pro toto, and pro tanto) and normative requirements. The "weighing conception" of reasons is also questioned by Horty (2007, 1–2) who proposes to conceive reasons as "defaults:" he sees the weighing conception as "incomplete as an account of the way in which reasons support conclusions."

[36] For Broome (2004, 42–43), "a putative example of an ought fact that has no weighing explanation" is that "You ought not to believe both that it is Sunday and that it is Wednesday."

[37] Alvarez (2010, 54) stresses that "the term 'motivating reason' does not have much currency outside of philosophy and can rightly be regarded as a term of art."

[38] Audi (2010, 275) distinguishes between "motivational" reason from a "motivating" one: the former is a "potential" reason "even if I never act on it," while the latter is one that "explains why I do" something. Motivational reasons must be "*possessed* reasons: reasons someone *has* to do something."

[...] in light of that reason," works as a premise "in the agent's (implicit or explicit) reasoning about φ-ing," that is as "a premise in the practical reasoning [...] that leads to the action" (Alvarez 2010, 35). It should also be noted that speaking of a singular reason "of an agent's motivating reason" or of "the agent's reason" is, of course, a "simplification," both because "an agent may be motivated to act by more than one reason," and because "a fact will seem a reason for me to act only in combination with other facts" (Alvarez 2016a).[39]

This definition leaves some issues open in relation to what a motivating reason is and, consequently, what the elements that constitute it are. Two are the main problems: on the one hand, the question of the role that the subject's desires, goals, and willing have in motivating action and, on the other hand, the role of the beliefs. In the first case, the problem relates to the ability that the reasons have to motivate agents to act (or not act): this is a matter of understanding the transition from *having reasons* to φ-ing to *doing* φ. We have to explain "how thinking that there is a reason for me to do something can motivate me to act, and to act *for* that reason" (Alvarez 2016a) and so to answer the question "about the conditions that determine when a reason for acting applies to a particular agent" (ibid.). In the second case, the problem is that of the role that knowledge and beliefs have in motivating an action: whether they are requisites of motivation and can be considered as motivating reasons. This seems particularly relevant in the case of "false belief," that is when it seems that an agent acts on the basis of a untrue belief.[40]

In relation to these questions it is possible to identify two positions: the first, related to the desire-based approach (and more generally to a Humean view of reasons), is that of "psychologism," while the second is that of "non-psychologism" or "factualism" (Alvarez 2016b). According to the first position, the answer to the two questions implies the central role of desires and beliefs (even disjointly), while according to the second, desires and beliefs can only be seen as aspects that generally motivate, but not as specific reasons for acting. According to the first position, desires and beliefs have a central role, "whether the reasons that apply to you depend on your desires and motivations," and at the same time, "for a reason to motivate you it must be a reason you have," requiring that you "possess" (Audi 2010, 275) the reason, and therefore, that you "must know or believe the consideration that constitutes the reason" (a mental state) (Alvarez 2016a); the second position, by contrast, denies the role of desires and of beliefs.

According to the first position, "it seems right that when an agent acts for a reason, he acts motivated by an end that he desires (an end towards which he has a 'pro-attitude') and guided by a belief about how to achieve that end" (Alvarez 2016a). According to the second position, the role of desires and beliefs is instead not primary, and motivating reasons are facts. This position distinguishes between

[39] Alvarez (2010, 127) notes that reasons for φ-ing can be either "independent" of or "related" to one another. The criterion for the distinction "is the relation between reasons for acting and the goodness or value of the action for which I take them to be reasons that explains why sometimes my reasons for doing something are independent of each other and why sometimes they are not."

[40] As we have already noted, Alvarez (2016a) highlights this problem using the example of Shakespeare's *Othello*.

"being motivated" (inclined) to do something and having a "motivating reason": desires, motivations, willing, purposes, and so on can be seen as things that motivate, but not as real motivating reasons (Alvarez 2010, 53ff.).

These two positions refer to different views about the ontology of the reasons. As we will see in Sect. 3, this is the question of the "conceptual category or categories" to which reasons belong: the category of mental states or that of facts (Alvarez 2010, 32). What must be noted is that these two positions include or exclude different aspects according to ontological choice. Thus, on the first vision, they will be "someone's goal or intention in acting, which is something that the agent desires" (Alvarez 2016a), plus an actor's belief, while on the second "they do not fall under the category 'motivating reasons'" and they can at most be states "that encompass motivation" (ibid., referring to Mele 2003) and thus general motivation, "not motivating reasons" (ibid.).[41]

2.3 Explanatory Reasons

We can define *explanatory reasons* as the "reasons why someone acts" (Audi 2010, 275). This means that the "explanans is 'the reason why' someone acted" (Alvarez 2010, 161).[42] A reason explanation is the "reason in the light of which [a subject] φ-ed" (ibid., 36).[43]

There are various kinds of action explanation[44]: following a suggestion of Alvarez, we can distinguish, with reference to the explanantia, three principal groups of explanatory reasons: "explanations of action that take the form 'A φ-ed because q', where 'q' is (i) the reason for which the agent acted […]; (ii) a reason why A φ-ed which is a fact concerning A's beliefs and desires, knowledge, emotions, feelings, motives, character traits, habits, etc.;" (iii) a reason that explain "by citing a purpose or goal that the agent pursued in his action, and they are formally characterized by the use of 'in order to', 'with the purpose of', 'for the sake of', or equivalent expressions" (ibid., 191).

The first ones can be defined as "reason explanations *proper*," the second as "*psychological explanations*," the third as *purposive* (*teleological*, *intentional*) *explanations*. In the first case, we cite in the explanans a reason of the agent that may also

[41] In Othello's case, according to the first position, the reason for killing Desdemona lies in his desire "to restore his reputation," crippled by Desdemona betrayal, while according to the second, the reason lies in "the putative facts that she is unfaithful to him and that killing her is a fitting way to restore his reputation." For a detailed analysis, see Alvarez (2016a).

[42] Alvarez (2010, 28) notes that "in addition to answers to 'why?'-questions, there are other kinds of explanation, for instance, explanations of how to φ (how to iron a shirt), of how x φ-s (how a steam engine works), etc. But to explain how to φ, or how x φ-s, is not to give a reason."

[43] Alvarez (2010, 166) stresses that the explanation is about "why someone habitually acts, is now acting, has acted, will act, etc., in a certain way," making these actions "intelligible."

[44] The answers that we can give depend also by "the pragmatics of explanation," that is "the context in which the question is asked" (ibid., 26).

be the “reason for which the agent acted” (ibid., 5): so we have a proper explanation because we have an overlap of explanatory and motivating reason.[45] In the second case, the explanans refers to “a psychological fact about the agent, such as the fact that he believed and wanted certain things, or the fact that he had certain motives, or character traits, emotions, habits, and so on” (ibid.). This second sort of explanation can be called a “‘Humean explanation’, which typically has the form ‘he φ-d because he believed that p and wanted x’” (ibid.). In the third case, “actions performed for a reason can be explained with purposive or intentional explanations,” that is “explanations of actions [that] identify the agent’s purpose, which is also normally his intention in acting.” These “are a kind of teleological explanation” of the form of “‘he φ-ed *in order to* ψ’” and “characterized by the use of ‘in order to’ or equivalent expressions (‘so as to’, ‘with a view to’, etc.)” (ibid., 171).

As is the case with motivating reasons, these different kinds of explanations find their foundation in a different ontological view of reasons. This fact is reflected not only in the difference between non-psychologistic explanation (which invokes a fact) and psychologistic explanation, but also in the different relevance that these views attribute to purposive explanation and to the distinction between reasons for action and causes.

Unlike psychologistic approaches, non-psychologistic ones see “purposive” explanations as nonindependent ones that must be “supplemented by reason explanations.” In fact, “these explanations […] do not *state* the fact that the agent had such a goal. And, when they are explanations of things done for reasons, they do not state the reason the agent had for doing what he did, though they often suggest what that reason was” (ibid., 170). On this approach, explanations that refer to the agent’s intentions make it necessary to further identify the reasons for the action, as the purpose and intention are merely the general directions of the action: “the goal or purpose is the end towards which the action is directed,” and the reason is “a fact that guided the agent in his pursuit of the goal mentioned in the explanation” (ibid., 194).[46]

Another distinction relates to the relationship between reasons to act and causes. Of course, it is possible to distinguish between causal explanation and reason explanation on the basis of the fact that the former is a “natural” relationship between facts (unintentional) and the second concerns rational actions. However, what distinguishes psychologistic explanations from non-psychologistic ones is that the former would admit of a causal explanation of the action. As we shall see, this point makes it possible to claim that psychologist explanations are more appropriate: only they can be a premise in a causal explanation. On a non-psychologistic approach, a

[45] For Audi (2010, 276), “they are reasons for which we do something and thereby ground a motivational explanation of our doing it.”

[46] You can have it as your purpose to be on time for a meeting but have different reasons why you are taking a taxi (someone stole your bicycle, you woke up late, the bus is late, and so on). Of course, there are actions that can be explained on the basis of their purpose: they are “merely reactive” (instinctive, mechanical, reactive, habitual, etc.) and “skilled actions” (riding a bike, skiing, dancing, driving, etc.) that “can be explained without the need to attribute to the agent any explicit or implicit calculation or reasoning” (Alvarez 2010, 193, 195).

reason-based explanation does not make it necessary to identify reasons as causes: in explaining an action, it is therefore important to identify the reasons "regardless of whether the explanations in which they feature are causal or not" (ibid., 30).

3 The Ontology of Reasons

Ontological reflection addresses the question of what is a reason, or what conceptual category it is possible to assign it to. As we have already seen in relation to the different roles of reasons, there are different characterizations of how to understand these roles and their actual meaning. So the fact that reasons can play different roles can itself be taken as a clue to their ontological diversity.

There are two main positions: on the one hand, psychologistic ones, which emphasize the role of beliefs and desires, and which consider reasons as mental states, and on the other hand, non-psychologistic (or factualist) ones, which see the reasons as facts.[47] Psychologistic positions assign a double ontological nature to reasons—that of facts and of mental states—while non-psychologistic ones identify only a single ontological determination, that of facts. Prevalent in contemporary literature is the idea that these two positions are in accord in relation to normative reasons, viewed in general as facts, but diverge in relation to motivating and explanatory reasons.[48] Although the divergences mainly concern these two aspects, they also affect the idea of normativity and, in particular, of their normative capacity. We will therefore discuss the different visions in general terms.

What is relevant is not just the subject of discussion (reasons as facts or as facts and mental states) but also, and especially, the implications these two positions have for the analysis of the reasons. In a vast debate, two crucial problems can be identified: the first is to determine what normative and motivating reasons are; i.e., whether or not in order to have a reason or for it to motivate, it must be part of an agent's mental state. The second concerns the role of beliefs: in this case, it is not just a matter of assessing not only the role of mental states, but also the rationality of action. If practical reasoning implies the presence of premises (from which to draw motives, decisions, conclusions, etc.), it seems necessary that these at least be known (and believed) by the agent. These two positions refer to widely debated issues in ethical and metaethical reflection. In particular, the issue of motivation refers, inter alia, to the contrast between internalism and externalism, while the problem of beliefs refers

[47] Alvarez (2016b) identifies different kinds of non-psychologism (factualism). Kantian perspectives are difficult to locate but, as just noted, they seem to be close to subjectivism (psychologism). For Scanlon (2014, 11), Kantians start from "the idea that claims about the reasons an agent has must be grounded in something that is already true of that agent," and this is "something similar might be said by proponents of desire-based."

[48] Alvarez (2016b) asserts that "Factualism is widely accepted for normative reasons, reasons that favour doing something. But things become more controversial when the question concerns the reasons for which we act and the reasons that explain our actions. This has led many to conclude that Factualism is right for normative reasons but not for other kinds of reasons."

to the distinction between objective and subjective (on the model of the distinction between objectivism and perspectivism). Further, distinctions are then present in the various settings: particularly, relevant is the one present in non-psychologistic positions that, in relation to the facts, separates "realism" and "irrealism."

The psychologist position can be seen as the prevailing, and in some ways "orthodox," one, while the non-psychologistic one develops from criticism of the former (Alvarez 2010, 2). This first position was made canonical by Davidson (1963) in an essay titled "Actions, Reasons, Causes," in which he describes the explanation of an action through a reason as rationalization. Davidson identifies what he calls a primary reason. In a well-known passage, he writes:

> Whenever someone does something for a reason, therefore, he can be characterized as (a) having some sort of pro-attitude[49] towards actions of a certain kind and (b) believing (or knowing, perceiving, noticing, remembering) that his action is of that kind." [Consequently] giving the reason why an agent did something is often a matter of naming the pro-attitude (a) or the related belief (b) or both; let me call this pair the primary reason why the agent performed the action. (Ibid., 685–686)

To this consideration of what a primary reason is, Davidson adds that only under these conditions can a reason be part of practical reasoning (which for Davidson is causal reasoning).[50]

As is apparent, in this reconstruction reasons refer to mental states such as wishes and beliefs. As noted, this seems to have important advantages with regard to both the motivation of action and the rationality of action. In the first case, the fact of being dependent on a subject's mental states makes it comprehensible because a normative reason applies to a subject. As noted, "desire-based accounts of reasons may seem to have the edge here" because the idea that a reason determines my behaviour is far more persuasive if it "depends on my antecedent motivations (desires, plans) and therefore that a subject is motivated what he believes" (Alvarez 2016a).

As noted, this issue, from a metaethical perspective, is marked by the distinction between internalism and externalism: the advantage of the first depends on the fact

[49] For Davidson (1963, 685–686), "under (a) are to be included desires, wantings, urges, promptings, and a great variety of moral views, aesthetic principles, economic prejudices, social conventions, and public and private goals and values in so far as these can be interpreted as attitudes of an agent directed towards actions of a certain kind. The word 'attitude' does yeoman service here, for it must cover not only permanent character traits that show themselves in a lifetime of behaviour, like love of children or a taste for loud company, but also the most passing fancy that prompts a unique action."

[50] Buckareff (2014) stresses that Davidson "argues that the relationship between reasons and actions displays the same pattern we discern in causal explanations" and that "if the onset of a primary reason is not a cause of action, we have difficulty accounting for the difference between when an agent has a reason for acting in mind that does not actually explain why she acts as she does and cases where the agent has a reason for acting that does explain why she acts as she does. If we dispense with a causal role for reasons, we may be able to appeal to some reasons of an agent in the light of which the action looks reasonable, but, absent a causal role, it is not clear that the putative justification for action explains the action and, hence, *really* rationalizes it." O'Connor (2010, 130) notes that in the perspective of the "casual theorist," it is possible "determining the true reason(s) for the action."

that the reference to pro-attitudes makes the motivation immediate and requires no additional factors (double pass). In fact, as a general scheme, it is argued from this perspective that "every reason for action must bear relation *R* to motivational fact *M*" (Finlay and Schroeder 2012), while in externalism this relationship is considered unnecessary.[51]

The second aspect that favours the psychologistic approach has to do with the role of beliefs. There are two main features: the first, as mentioned, is linked to the need to identify the reasons for an agent's behaviour requires the agent's "believing or knowing": "a fact that is merely 'out there' cannot explain why you do anything," and therefore the reasons for acting are "mental states (believings, knowings, etc.)" (Alvarez 2016a). The second feature has to do with the ability of the psychologistic approach to explain the action that takes place on the basis of "false beliefs" ("error case"). The fact that actions can be made on the basis of such beliefs seems explicable only from the central role of mental states. Without this acquisition it would also seem impossible to explain the rationality of action (in the minimal sense of drawing conclusions from premises). In fact, some actions take as their starting point a false (albeit sometimes reasonable) belief that justifies and explains them. Reasons in this sense are to be understood in the light of the agent's subjective perspective: psychologistic theories therefore support a "perspectivist" position, as the fact of having a (normative) reason "is not independent of [the agent's] perspective, which includes her beliefs" (ibid.).

Non-psychologistic positions, as mentioned, claim that the reasons (especially normative ones) are facts. They question the main assumptions of the psychologistic theses and point out their conceptual ambiguity. As to the relevance of desires, these positions, on the one hand, emphasize that action often does not occur on the basis of desires and that this can be detected in relation to moral norms and to a wide range of cases, which include, e.g. commands, orders, while on the other hand they distinguish between "being motivated" and reasons to act: from this perspective, desires, beliefs, purposes, pro-attitudes, etc., can be seen as "inclinations," but such inclinations are put into effect on the basis of reasons understood as facts (see Alvarez (2010), 53ff.).[52]

More complex seems the reply to problems posed by false beliefs and cases of error. In this context, the answer that appears most convincing is that reasons are

[51] Finlay and Schroeder (2012) refer this definition to schematic internalism and stress that "different ways of spelling out relation *R* and motivational fact *M* correspond to […] a different thesis—a *version* of reasons internalism." Darwall (1992) distinguishes between "existence" and "judgment" internalism: the latter is based on the idea that "assent to a moral judgment […] concerning what one should do is necessarily connected to motivation (actual or dispositional)"; according to the first, "someone morally […] has to do something only if, necessarily, she (the agent) has (actually or dispositionally) motives to do so."

[52] The scheme of this analysis can be summarized as follows: the motivations (which cannot be reduced exclusively to dispositions, as in Ryle's analysis) are mainly understood as desires ("to give the motive for a deed is to indicate a desire for the satisfaction of which the deed was done"), and a desire is in turn seen as an "inclination to act." See Alvarez (2010, 60, 61, 70), referring to Ryle (1949) and White (1968) (quotation in parentheses) and criticizing Smith (1987) and his attempt to characterize desires through the use of the metaphor of "direction of fit."

"apparent reasons," that is "something [that] appeared to the agent to be a reason but it was not really one" (ibid., 140). This answer is substantiated by highlighting what appear to be the main ambiguities of psychologistic theories, namely the possibility of interpreting mental states either as acts or as the content of acts (the "act/object ambiguity"): "the term 'belief' [and also 'desire'] can be used to refer to one's believing something or to what one believes" (the content of a belief) (ibid., 125). If we use the term *belief* in the second sense, "to say that reasons are beliefs is not to say that they are mental states, because, even if believing that p is a mental state, what is believed (that p) is not itself a mental state" (ibid., 45). The two visions of what a belief is "are quite different": "reasons might be beliefs and desires and yet not be 'believings' and 'desirings' and therefore may not be mental states" (ibid.). Considering beliefs (and desires) as what is believed means referring "to a proposition," and this signifies that if "we think of [...] reasons as what is believed, as facts, as propositions, etc., it is clear that what we are thus thinking about are not psychological entities" and we are not committed "to the view that motivating reasons are psychological entities."[53] In sum, "what is believed is not a psychological entity" (ibid., 154). In relation to error cases, then, this approach "does not lead to any implausible conclusions [...] because in such cases the agent acts only for an apparent reason—and the apparent reason he acts for is not that he believes that p, but rather what he believes, namely that p" (ibid., 146).

What kinds of facts do the non-psychologistic theses refer to? They can be said to generally use a broad concept of fact. The idea that "all reasons are facts or [...] truths, which are expressed propositionally, and which can be premises in reasoning, both theoretical and practical" is declined in a "minimalist" conception (as opposed to an "austere" one) that makes it possible to include the greatest number of facts (including, e.g., negative facts and values).[54] Such a conception can be stated either by saying that a "fact is a true proposition" (Alvarez 2016a) or, in a more undemanding version, by affirming that by "'fact' is meant simply that which can be designated by the use of the operator 'the fact that...'" (Raz 1999a, 17–18).[55] Two aspects are finally to be considered: first, the theories that see reasons as facts can be divided into "realistic" and "irrealistic" ones, and second, that non-psychologistic theories are objectivist and not perspectivist. In the first case, realistic theories support a view of the facts in "some ontologically substantial or robust sense," seeing facts "as the truth-makers for true normative propositions" (Robertson 2009, 13),[56] while "irrealist" theories "deny that such truths are made true by, or obtain in virtue of there

[53] In this sense, Dancy (1993, 32) notes that "what motivates is the matter of fact believed" and that "what motivates is the fact that one believes, which is still a fact."

[54] Alvarez (2016a) points out that there is "disagreement about what facts of any kind are: are they concrete or abstract entities? Is a fact the same as the corresponding true proposition, or is the fact the 'truth-maker' of the proposition? Are there any facts other than empirical facts, e.g. logical, mathematical, moral or aesthetic facts?"

[55] Raz (1999a, b, 18) adds that "a fact is that of which we talk when making a statement by the use of sentences of the form 'it is a fact that...' In this sense facts are not contrasted with values, but include them."

[56] These theories can be distinguished into naturalistic and non-naturalistic ones.

being, normative properties construed as robust or substantial ontological items" (ibid.).[57] Non-psychologistic positions are "'objectivist' in that they presuppose that whether an agent has an (objective) normative reason to act depends solely on the facts and not on the agent's beliefs" (Alvarez 2016a).

4 Epistemic and Practical Reasons

As we have seen, in the sphere of normative reasons it is possible to distinguish reasons in relation to the type of behaviour they may favour. The distinction between epistemic and practical reasons (for action) is based on the type of φ-ing they refer to: the first are those on which basis we "believe something" and/or "make a […] cognitive step within a train of reasoning to a conclusion" (Skorupski 2010, 35),[58] while the latter are those in which our concern is "to perform an action." Epistemic reasons are those that have the concept of truth as a reference point, while the latter may refer to different values, norms, etc. According to the first, rightness or appropriateness is "related to truth," while according to the second, it concerns "the good, broadly conceived, that is, as related to a variety of values" (Alvarez 2010, 3–4).

It follows that "epistemic reasons are governed by one concern: determination whether the belief for which they are reasons is or is not true," while "reasons for a single action may, and typically are, governed by many concerns" (Raz 2009, 41).[59] This can be seen as the most important difference between these two types of reasons: "practical reasons serve many concerns and epistemic ones can serve only one" (ibid., 43).

This difference can be articulated in different ways: as indicated, a first analysis points out that "reasons for believing are not person-relative," while "reasons for acting and wanting are" (Alvarez 2010, 19). From this perspective, this means that, in spite of several possible objections,[60] they do not in principle depend on the agent's perspective. It follows that "if the fact that p is a reason for someone to believe that q, then it is a reason for *anyone* to believe this, no matter what his or her circumstances and goals are" (ibid.). The reasons for the action instead have the "distinctive feature" that "they are the subject of choice:" in addition of referring to different values, this

[57] Robertson (2009, 13) adds that on these positions, "a normative proposition might be true in virtue of satisfying some truth- or knowledge condition—such as those presented by *formal* criteria like universaliability, convergence commitments, the Categorical Imperative, and so on—which, when satisfied, incur no commitment to there being robustly normative properties within the fabric of the world."

[58] Skorupski (2010, 35) brings up as examples such cases in which we "introduce a supposition, make an inference, or exclude a supposition, etc."

[59] For Raz (2009, 37), epistemic reasons are "truth-related"; that is, they "are reasons for believing in a proposition through being facts which are part of a case for (belief in) its truth."

[60] For the objections, see Alvarez (2010), 20–22.

stands in relation both to the "different means of achieving one's goal" and to the "choice between possible but (at the time) incompatible goals" (ibid., 17).[61]

Joseph Raz (2009) has analysed the two types of reason in relation to values, indicating how, in general, reasons for action refer to values (but not only values),[62] while epistemic reasons "are not similarly connected to values, not even to a single value": it is not possible to say that "there is always value in having a true belief" and not even that "it is always a disvalue to have a false belief" (ibid., 43). There follows what can be seen as a specific difference: "the value-independent character of epistemic reasons" (ibid., 44). Unlike reasons for action, these reasons "are not related to values" (Sobel and Wall 2009, 3): they "do not derive from the value of having that belief in the way that reasons for an action derive from the value of that action," and therefore, "reasons for belief are not provided by values in the way that reasons for action are." This is to say that reasons for action can be seen as "presumptively enough," that is enough to justify an action (if there are several reasons, they could be traced to a single reason) or against an action, while "epistemic [reasons] are not necessarily so" (Raz 2009, 43, 44).[63]

In Raz's analysis, epistemic reasons are "adaptive," while those for action are "practical." In the first case, they are reasons that "mark the appropriateness of an attitude in the agent independently of the value of having that attitude, its appropriateness to the way things are," while the latter are "value-related" reasons (ibid., 46). This distinction is matched by that between standard and non-standard reasons: reasons of the first kind (corresponding to epistemic reasons) "are those which we can follow directly, that is have the attitude, or perform the action, for that reason," while reasons of the second kind (corresponding the practical reasons) are "reasons for an action or an attitude [...] such that one can conform to them, but not follow them directly" (ibid., 40). This analysis is aimed at highlighting that normativity is not necessarily bound to values: "the value of a thing provides some reasons, but not all" (Raz 2011, 95).[64] The analysis shows (or should show) that reasons should

[61] Raz (2009, 41) underlines that the possibility of choice regards not only single actions that "can serve or disserve a number of intrinsic values," but also action that refers to a single value that "may serve independent concerns." This happens, for example, when "a single act can advance the welfare of several individuals, when the interest of each of them is a reason, an independent reason, to perform it": in this case, we may face the impossibility of their contemporary practicability or of their possible conflict.

[62] Raz (2009, 43) underlines that "the diversity of concerns manifested in practical reasons is not entirely due to the diversity of values. Diverse values do generate diverse concerns, but so do other factors: for example, being a medically qualified caretaker of sheltered accommodation for disabled people I have a reason to help anyone there who needs insulin injections. There are several such people. So I have a reason to help each of them. Each of these reasons represents an independent concern, and they can conflict with each other, even though they all derive from the same value."

[63] Among the other differences that can be added, there is the one noted by Skorupski (2010, 156, xvii, 40, 41), for whom "epistemic reasons are relative to an epistemic field," that is to a "set of facts knowable to the actors" that create "epistemic dependencies." For Skorupski, "epistemic reasons are relative to their field: whether a subset of facts is an epistemic field constitutes a reason—and how good a reason for belief it constitutes—depends on the other facts of the field."

[64] In an even stronger stance, Raz (2011, 95) has argued that "the difference between practical and epistemic reasons is central to the attempt to understand the normativity of reasons. It defeats any

be seen in the different dimensions (standard, non-standard, adaptive, practical) that the distinction between epistemic reasons and reasons for action makes it possible to emphasize.

5 The Modality (and Strength) of Reasons

One of the central aspects of the analysis of reasons for action is that of evaluating the different normative roles they may have in determining the action of the various subjects. As we have seen, this is linked to the fact that reasons for acting are, in principle, pro tanto reasons: this means that we have reasons for a particular action and reasons against the same action.[65] The possibilities of contrast are wide and at a minimum they refer to the following: the different values an action can express, the choice between two types of actions that have different goals, and the choice of the means with which to achieve a given goal. These possibilities give rise to a phenomenology of reasons that highlights the different modalities (and strengths) they can have.

In an analysis comparing different reasons, it is necessary to distinguish between a conflict between first-order reasons and between first- and second-order reasons: in the first case, this is the contrast between the reasons supporting "different and incompatible courses of action," a contrast that "ordinarily we resolve […] by assessing the relative weight or strength of all the relevant reasons and then deciding in favour of that action which has the greatest overall support" (this is therefore the process of "determining what ought to be done on the balance of reasons") (Perry 1989, 973); in the second case, the reference is a reflective process that determines whether "to act on or refrain from acting on a reason" (ibid.) in relation to reasons of different levels that can orient the weighing process in different ways. In this second case, we have reasons for reasons, such as those which can be determined by the presence of mandatory norms (issued by an authority, be it practical or epistemic).

If we accept the distinction between first- and second-order reasons, we will have two types of conflict between reasons: those between first-order reasons and those between second-order (exclusionary) reasons and first-order reasons.[66]

attempt to explain normativity as having to do with the influence of value on us. Epistemic reasons have nothing to do with value."

[65] Alvarez (2010, 15) takes the example of bungee-jumping: "If bungee-jumping is a good thing to do (at least for some people) this is presumably because it is fun, thrilling, exhilarating, and so on—that is, it is good for what might be called a hedonic reason. On the other hand, given the risks involved, there seem to be prudential reasons against bungee-jumping."

[66] Raz (1999a, 35) notes that "description of conflicts of reasons and their resolution […] is one of the most intricate e complex areas of practical discourse."

5.1 Conflict and Weighing Between First-Order Reasons

There is a conflict between first-order reasons when "p strictly conflicts with q relative to X and φ if, and only if, R (φ)p,x and R(φ)q,x, i.e. that p is a reason for x to φ and that q is a reason to refrain from φ-ing" (Raz 1999a, 25): basically, when you have reasons in favour of a particular behaviour and reasons against that behaviour.

In order to analyse this type of contrast (between first-order reasons), it is necessary to (a) define the role of reasons in relation to the context; (b) evaluate the types of conflicting reasons and their characteristics; and (c) consider the various options in relation to the choice between possible actions.

(a) Comparison between first-order reasons requires a preliminary choice about the way in which the reasons can be considered in practical reasoning and in relation to the context. In this sphere, following Dancy's suggestions (2004a, 9, 132, 94), it is possible to choose between two possible approaches: holism, which claims "that a feature which has a certain effect when alone can have the opposite effect in a combination," that is "that a feature that normally counts in favour of a (sort of) action may on occasion not count in favour at all," and atomism, which is "the claim that if a feature is a reason in one case, it must be a reason (and on the same side) wherever it occurs." This means that on a holistic approach, the "context can affect the ability of a feature to make a difference in a new case" (ibid., 7).[67]

(b) The reasons that may collide can be different: a first distinction concerns the comparison between epistemic and practical reasons. As noted, epistemic reasons are "standard," while those practical are "non-standard:" that means that "a conflict between a practical reason to believe p and an epistemic reason to believe not p is not a genuine conflict," since "the two kinds of reason do not compete." In these cases, "the epistemic reason will win out" (Sobel and Wall 2009, 3). However, the possible contrast between different epistemic reasons is a contrast between first-order reasons.[68]

First-order reasons can be categorized differently: a general classification that outlines their different role is the one proposed by Dancy (2004a), who in the context of the analysis of "contributory reasons" distinguishes between favouring reasons (*favourers*), enabling reasons (*enablers*), and intensifying reasons (*intensifiers*). A "contributory reason for action" is a pro tanto (or prima facie) reason: it is "a feature whose presence makes something of a case for acting, but in such a way that the overall case for doing that action can be improved or strengthened by the addition of a second feature playing a similar role" and that "is not necessarily destroyed by the presence of a reason on the other side" (ibid., 15). In this context, favouring

[67] In relation to the role of moral principles, Dancy (2004a) associates these two positions with two general approaches to reasoning, that of "generalism" (reasons depend on general principles) and "particularism" (reasons do not depend on general principles). For Dancy, "normally, particularists are holists and generalists are atomists" (ibid., 9). He also distinguishes holism from nonmonotonic reasoning.

[68] For Alvarez (2010, 14), "when the reason to believe something and the reason not to believe it are of equal strength, then believing either may be right."

reasons (favourers) are those that provide a reason to act in a certain way; that is, they can make a specific action "right or appropriate." To these must be added enabling reasons, which are those that refer to conditions that make it possible to act in a certain way, namely those that, so to speak, determine the necessary conditions for realizing the act.[69] To these two reasons, we must adjoin reasons that intensify a reason (intensifier): these are the ones that "strengthen" a favouring reason. The overall picture of reasons envisions "three sorts of role that a relevant consideration can play: a relevant consideration can be a *favourer/disfavourer*, it can be an *enabler/disabler* for another favourer/disfavourer, and it can *intensify/attenuate* the favouring/disfavouring done by something else" (ibid., 42).

A classification developed more directly in relation to the possible conflict between reasons is the one proposed by Raz (1999a). The starting point is the identification of a *complete* reason: this is a reason that is "indispensable in any logical explication of reason," which implies that

> the fact that p is a complete reason to φ for a person x, if, and only if, either (a) necessarily, for any person y who understands both the statement that p and the statement that x φ's, if y believes that p he believes that there is a reason for x to φ, regardless of what other beliefs y has, or (b) R(φ)p,x entails R(φ)p,y which is a complete reason. (Ibid., 24)

This definition is not immediately apparent and can be explained in the light of the problem it wants to explain, that is "the difference between completing the statement of a reason and [...] stating a second reason" (ibid., 23). The statement that "wherever φ-ing would increase human happiness one has a reason to φ" is a complete reason if you add the premise that "human happiness is a value": a reason is of this kind if "the fact stated by any set of premises which entail that there is a reason to perform a certain action is a complete reason for performing it" (ibid., 24–25).[70] An *atomic* complete reason can be defined "as a complete reason which would cease being complete if any one of its constituent parts were omitted" (ibid., 25).

A complete reason includes *operative* reasons and *auxiliary* reasons: in the first case, we have reasons that involve an inference, "such that belief in their conclusions entails having a practical critical attitude while no such attitude is required for belief in their premises" (ibid., 33).[71] They are "any reason if, and only if, belief in its

[69] Dancy (2004a, b, 30) explains the different roles with this example: "1. I promised to do it. 2. My promise was not given under duress. 3. I am able to do it. 4. There is no greater reason not to do it. 5. So: I do it." Number 1 is a favourer, while 2, 3, and 4 are enablers (general or specific).

[70] Raz (1999a, 24–25) makes this example: "Suppose John says: wherever φ-ing would increase human happiness one has a reason to φ. Let us assume that Jack denies this. How are we to understand Jack's position? Is he guilty of a mistake in logic? Not necessarily. John does not state a complete reason, though it is easy to see which reason he is invoking. It is that human happiness is a value and that under certain conditions φ-ing increases human happiness. This is his complete reason for φ-ing when those conditions obtain." Of course, Jack can continue to deny that human happiness is a value, but he would make a logical mistake if "the reason of his denial is that values do not always constitute reasons, or that sometimes there will be stronger reasons for not φ-ing despite the fact that it contributes to happiness."

[71] Raz (1999a, 34) stresses that "most operative reasons are either values or desires or interests:" the latter can be called "subjective values," while "values are dubbed objective values" (that is

existence entails having the practical critical attitude": this is the case, for example, if you claim that "respect for persons is a value then there reason for everyone to respect persons," or the fact that my having "promised to φ" means that I "have a reason to φ" (ibid.). Auxiliary (identifying) reasons whose function is "to help to identify the act which there is reason to perform" (e.g. to identify whether a loan can be the most appropriate way to help someone in need) and can be countered "with strength-affecting reasons" (that is, what is more beneficial) (ibid., 34–35). They concretize, so to speak, operative reasons. For Raz, complete and operational reasons must necessarily be together: "every complete reason includes an operative reason and that every operative reason is a complete reason for some action or other" (ibid., 33).

Reasons should be compared: the strength of a reason lies in its "power to override." An assessment of the strength of reasons must take account of any *cancelling* conditions (as in the case in which "a friend has released me from a promise"): they eliminate a reason and therefore do not concern their strength (ibid., 27).

In relation to the strength of reasons, it is possible to identify, in addition to prima facie reason, *conclusive* reasons and *absolute* reasons. In the first case, "p is a conclusive reason for x to φ if, and only if, p is a reason for x to φ (which has not been cancelled) and there is no q such that q overrides p," while, in the second, "p is an absolute reason for x to φ if, and only if, there cannot be a fact which would override it; that is to say, for all q it is never the case that when q, q overrides p" (ibid., 27).[72] This leads to the conclusion that "it is always the case that one ought, all things considered, to do whatever one ought to do on the balance of reasons" (ibid., 36).[73]

A particular role in the balancing process can be assigned to moral reasons. This role, however, should not be seen in the manner of the necessary prevalence of moral reasons, that is "that moral reasons for acting always defeat other reasons," for "it would be sufficient that there was no moral reason *against* φ-ing (Alvarez 2010, 16)."

that "everyone has an operative reason to promote"). On the distinction between subjective versus objective and relative versus neutral, see note 33 above.

[72] Raz (1999a, 28) notes that "not every conclusive reason is absolute. A reason may be conclusive because it overrides all the existing reasons which conflict with it and yet not be absolute because it would be not override a certain possible reason, had it been the case."

[73] In relation to the process of weighing, Parfit (2011, 32–34) distinguishes between *decisive* and *sufficient* reasons: we have a decisive reason "if our reasons to act in some way are stronger than our reasons to act in any of the other possible ways, these reasons are *decisive*, and acting in this way is what we have *most reason* to do. If such reasons are much stronger than any set of conflicting reasons, we can call them *strongly* decisive." We have sufficient reasons if "there is […] nothing that we have decisive reasons to do, or *most* reason to do, because we have *sufficient* reasons, or *enough* reason, to act in any of two or more ways." In these cases "our reasons to do something are sufficient when these reasons are not weaker than, or outweighed by, our reasons to act in any of the other possible ways. We might have sufficient reasons, for example, to eat either a peach or a plum or a pear, to choose either law or medicine as a career, or to give part of our income either to Oxfam or to some other similar aid agency, such as Médecins Sans Frontières."

(c) Often, it seems that a contrast between reasons cannot be solved only on the basis of weighing of reasons, as in the case of multiple options. This can happen in the frequent situations in which "people have a variety of options such that it would accord with reason for them to choose any one of them and it would not be against reason to avoid any of them" (Raz 1999b, 99). This may mean that "in some circumstances reasons are optional" (ibid., 94), that is reasons for which "the fact that there are reasons for a certain response make it an eligible, attractive response, but not one which it is wrong not to adopt." They are situations "where neither believing a proposition nor withholding belief will be irrational" (ibid.). The presence of optional reasons can be explained on the basis of "a special optional type" (ibid., 94) or on the basis of the presence of "incommensurable" reasons. In the first case, the reference is to the distinction between *enticing* and *requiring* reasons: the first would be those that "make an option attractive" (ibid., 100) but do not involve irrationality if they are not followed. This seems to suggest that the choice of an option, because enticing, does not take place on the basis of weighing of reasons. In the second case, the presence of optional reasons would depend on the fact that the reasons are incommensurable and, therefore, on reasons that cannot be assessed on the basis of their weight (lacking a common measure): "Reasons which are incommensurate do not defeat each other" (ibid., 101). These aspects can be analysed in two main ways: on the one hand, as Raz suggests, on the basis of an examination of the different factors of the options involved, by trying to assess the incompatible options on the basis of these factors (and therefore in some way weighing them), while, on the other hand, as Dancy (2004b) seems to suggest, by assessing the role of the conditions of implementation and of context.

5.2 First-Order Reasons and Second-Order (Exclusionary) Reasons

As noted, a further level of comparison between reasons is that between first-order and second-order reasons for acting. This means that "determining" actions "on the balance of first-order reasons is not the only mode of practical reasoning" to which we can refer (Perry 1989, 913). This is because of the presence of secondary reasons, that is, as Raz (1999a, 39) states, "any reason to act for a reason or to refrain from acting for a reason." We can become aware of their presence through "the detailed examination of conflicts of reason which forces the recognition that different reasons belong to different levels, which fact affects their impact on conflict situations" (ibid., 35). The example is that of a subject (Ann) "looking for a good way to invest her money." From a friend she receives, late in the evening, a proposal for what might be a good investment, but "she has to decide the same evening for the offer" (within midnight). She is undecided: she knows it might be a good investment, but she needs to evaluate it by comparing with another offer she has received before. "All she requires is a couple of hours" to evaluate the two proposals, but she does not because she is tired, she had a hard day, and she does not feel able to rationally evaluate the proposal. In short, she rejects her friend's offer, not because she believes that it is not good (compared to the other), but only "because she cannot trust her own judgement at this moment" (ibid., 36). In this way, she "claims to be acting for a reason which is not taken into account" in weighing reasons: what is "special" in this case "is, not that she regards her mental states as a reason for action, but that she regards it as a reason for disregarding other reasons for action" (ibid., 37–38).[74]

This example shows that we can have "a reason for not acting on the balance of reasons" and in particular that we can have "a reason to refrain from acting for a reason" (ibid., 39). These types of reasons are the most important second-order reasons: they are *exclusionary* reasons, that is reasons for which we do not take into account the weighing of first-order reasons (for or against). This shows that there may be a "conflict between a first-order reason and a second-order exclusionary reason" (ibid., 40). What is specific to this type of conflict is that "by a general principle of practical reason [...] exclusionary reasons always prevail, when in conflict with first-order reasons" (ibid.).

Exclusionary reasons "may vary in scope; they may exclude all or only some of the reasons which apply to certain practical problems," and they "may also conflict with and be overridden by another second-order reason" (ibid.). The presence of exclusionary reasons shows that "there are two ways in which reasons can be defeated:" by "conflicting" or by "exclusionary" reasons. This possibility raises the problem of how to "distinguish between the two ways in which a reason can be defeated," that is to "have a test" by which to identify the two types of reasons. This problem could create some difficulties,[75] but it is possible to identify cases in which the individuation of exclusionary reasons is totally clear: this is true for "decisions and mandatory norms [that] can only be explained with reference to exclusionary reasons" (ibid., 41).[76]

As noted, in a conflict between exclusionary and first-order reasons, the first "always prevails," though it can "be cancelled by cancelling reasons."[77] We can also mention the fact that "the scope of exclusionary reasons can be affected by [...] scope-affecting reasons": since "the scope of an exclusionary reason is the class of reasons it excludes," scope-affecting reasons are those that can strengthen (or narrow the scope of) an exclusionary reason (like the high ranking of an authority) (ibid., 46–47). There can also be conflicts "between second-order reasons" that, like

[74] Raz (1999a, 37–39) offers two more examples, relating to the commands of an authority and to a promise.

[75] Raz (1999a, 45) notes that the conflict between first-order and second-order reasons is characterized by "mixed reactions."

[76] Raz (1999a, 47–48) underlines that there are "two main types of exclusionary reasons": "incapacity-based exclusionary" ones and, in general, "authority-based reasons."

[77] The rule of practical reasoning is, in these cases, that "one ought not to act on the balance of reasons if the reasons tipping the balance are excluded by an undefeated exclusionary reason" (ibid., 40).

first-order reasons, can be considered in the light of the strength of the contrasting reasons.[78]

6 Reasons in (Philosophy of) Law

Reflection on the relationship between legal norms and reasons is largely owed to Joseph Raz's work.[79] This reflection, which has many aspects to it,[80] has largely focused on the question of what kind of reason legal rules are.[81] In the philosophy of law, this question has been tied not only to the identification of the type of reason expressed in legal norms, but also to that of legal normativity (and the relation between law and morals).[82]

In Raz's work (1999a), the identification of what types of reasons are legal norms is primarily developed in relation to mandatory norms[83] and on the basis of "content-independent" considerations.[84] The basic thesis is that these norms and some rules

[78] Raz (1999a, 47) emphasizes that he considers exclusionary reasons as the most important second-order reasons and that he does not discuss "second-order reasons to act for a reason."

[79] Redondo (1999, 98) underlines that, for Raz, "in order to account for the concept of legal norm, one must first have a concept of reason for action."

[80] Redondo (1999, 97) stresses that "the concept of reason for action is thought to be relevant for the study of a broad range of questions. In general, it is considered useful for improving the approach to and the explanation of many controversial issues in the field […]. This is the case, for instance, with respect to the problems of normative authority, the existence or validity of a rule, the way how these affect the reasoning of their addressees, etc."

[81] Another important problem, but mainly related to the theory of law, is that of the relation between reasons and law and in particular that of the role that specific legal reasons (such as those related to rights) may have in the processes of applying the law: in this field, the main issue is that of the priorities, in relation to the content of rights (in the form of a material or axiological hierarchy of such content). This, for example, is the case with Ronald Dworkin (1977) and with his theory of rights as "trumps." As is well known, Dworkin argued that, in the processes of applying the law, individual rights always prevail over the community's general goals: in Dworkin's terms, principles (which express individual rights) always prevail over policies. For the analysis of these aspects, see Bongiovanni and Valentini chapter 5, part III, this volume, on "Balancing, Proportionality and Rights."

[82] This problem, which can be seen as the "classic" one of the philosophy of law, will not be analysed in this contribution.

[83] Raz (1999a, 49) states that he prefers "'mandatory' to the more common 'prescriptive.'" The latter "is often used to characterize a type of meaning or a type of speech act […]. 'Prescriptive' also connotes the presence of someone": these are aspects that do not pertain to rules and principles.

[84] Raz (1999a, 51, 50) states that "one is […] forced to look to content-independent features of rules to distinguish rules from reasons which are not rules." With reference to von Wright (1963, Chap. 5), Raz identifies "four elements in every mandatory norms: the deontic operator; the norm subjects, namely the persons required to behave in a certain way; the norm act, namely the action which is required of them; and the conditions of application, namely the circumstances in which they are required to perform the norm action."

are exclusionary reasons[85]: "the notion of exclusionary reasons is essential to the explanation of mandatory norms, especially in order to understand the ways in which their role in practical reasoning differs from that of ordinary reasons for actions" (ibid., 73–74). More specifically, as we will see, "a mandatory norm," for Raz, "is either an exclusionary reason or, more commonly, both a first-order reason to perform the norm act and an exclusionary reason not to act for certain conflicting reasons" (ibid., 58). This double aspect qualifies norms as *protected* reasons, that is "a special kind of reason which combines a first-order reason with an exclusionary reason" (Redondo 1999, 115).[86]

The demonstration of the direct relation between exclusionary reasons and norms takes place in successive steps: firstly, in reference to rules of thumb and those issued by an authority, Raz shows what justifies the fact that they are exclusionary; secondly, the role of the rules is associated with decisions, and this makes it possible to highlight the role played by acceptance and beliefs; thirdly, in the light of the problem of the distinction between rules and norms, Raz analyses the problem of existence/validity of mandatory norms and their characteristics.

What makes exclusionary reasons the rules of thumb and those issued by authority is their role in practical reasoning: they have a precise function that justifies their use. In the first case, what justifies their use is that they are "labour- and time-saving devices [and] error-minimizing devices" (Raz 1999a, 74). They are therefore justified by the task they carry out as tools in relation to "what ought to be done," to save time and work and reduce risks. They are, then, "reasons for having rules," and as such, they "determine the nature of the rules themselves" (ibid., 59): these are specified in the conditions of application of the exclusionary rule. It is necessary to distinguish between the maxim of experience and rules, as "following a rule entails its acceptance as an exclusionary reason for not acting on conflicting reasons even though they may tip the balance of reasons" (ibid., 61). Regarding norms issued by authority, their role is determined by the "nature of authority," which can be "epistemic" or established to ensure social cooperation. In the first case, the authority is "based on knowledge and experience" that "ought to be followed [. . .] when the advice is based on information or experience which the adviser" owns and which we do not or cannot have. This advice is "justified by the wisdom of the authority" (ibid., 63, 74). The rules based on the requirements of social cooperation are justified by the need to ensure such cooperation and are in this sense necessary: in order to cooperate, there must be exclusionary reasons that enable social actors to act on the basis of the instructions

[85] This consideration starts from the criticism of Hart's (1961) practice theory of norms. For Raz (1999a, 53), "the practice theory suffers from three fatal defects. It does not explain rules which are not practices; it fails to distinguish between social rules and widely accepted reasons; and it deprives rules of their normative character."

[86] For Enoch (2014), this type of reason is a "combination of reasons": "A *protected* reason, as I understand it, is such a combination of reasons: If you have a protected reason to φ then you have both a reason to φ and an exclusionary reason excluding at least some of the reasons against φ-ing." In the same essay, he adds to exclusionary reasons *quasi-exclusionary* reasons, which broaden the range of action of the first (for instance, "they include [. . .] reasons not to deliberate in some ways on some reasons").

of the authority. As a result, conceptually, "norms justified by the need to secure coordination must be regarded as exclusionary reasons" (ibid., 74).

To emphasize a further important aspect of exclusionary reasons, Raz uses the analogy with decisions.[87] They are reasons: "a decision is always, for the agent, a reason for performing the act he has decided to perform and for disregarding further reasons and arguments. It is always both a first-order and an exclusionary reason" (ibid., 66). A decision made by a given situation (such as not carrying acquaintances if my car has mechanical problems) can be generalized and become a rule: in this case, it becomes a general exclusionary reason and therefore no longer linked to the evaluation of the various reasons.[88] This role, which "does not depend on whether [someone] came to follow it one way or the other," is based on the fact that "we [...] believe that we are justified in following the rule," that is that the rules are believed to be "valid" reasons for the action (both as first-order reasons and as exclusionary ones) (ibid., 72, 75). Mandatory norms show the same features:

> a person follows a mandatory norm only if he believes that the norm is a valid reason for him to do the norm act when the conditions for application obtain and that it is a valid reason for disregarding conflicting reasons, and if he acts on those beliefs. Having a rule is like having decided in advance what to do. (Ibid., 72–73)

The analogy with decisions leads us to underscore the role of belief in the validity of the rules and so the need to "explain what it means for a mandatory norm to be valid" (ibid., 73).[89]

The analysis of this aspect introduces as reference to the mandatory norms "those who believe in their validity" (ibid., 74), i.e. the participants in the system, and sees the fact of following the rule as a "clue" to their validity; however, the dimension of validity is not linked exclusively to that of belief or acceptance, but remains associated with the justification of exclusionary reasons, that is to their ability to simplify decisions (to be time- and work-saving, etc.) and to the fact that they are produced by authorities.[90] To be valid means both to be able, in given circumstances, to carry out the first task and to come from a "legitimate" authority.[91] In the latter case,

[87] For Raz (1999a, 71), "this analogy provides a key to an understanding of the nature of mandatory norms."

[88] Raz (1999a, 73) states that "when the occasion for action arises one does not have to reconsider the matter for one's mind is already made up. The rule is taken not merely as a reason for performing the norm act but also as resolving practical conflicts by excluding conflicting reasons. This is the benefit of having rules."

[89] For Raz (1999a, 73), the reference problem is that "not every rule is a valid reason."

[90] Raz (1999a, 73) seems to merge two levels: the validity of a norm involves "more than [...] following a rule" (such that "a norm is valid if, and only if, it ought to be followed") and, at the same time, for a norm to be valid requires that "a person follows a rule only if he believes it to be both a valid first-order and an exclusionary reason," that is a "combination of reasons in the validity of which he believes."

[91] The idea of legitimate authority was developed by Raz in his "service conception" (see Raz 1979; 1986). For a review of the various analyses and criticisms of Raz's authority theory, see Ehrenberg (2011). The problem of authority will not be discussed here: for the analysis of this problem, see Himma and Rodriguez Blanco chapter 8 and 9, part I, this volume, on "Authority" and on "The Authority of Law."

the connection between the origin of the rule and its exclusive character is conceptual because, by definition, its coming from authority determines its exclusivity: “to regard somebody as an authority is to regard some of his utterances as authoritative even if wrong on the balance of reasons” (ibid., 65).[92]

In the attempt to “deflate” (Bix 2011, 412), the debate on the normativity of law, David Enoch proposes a different analysis of the relation between legal rules and reasons. In particular, Enoch analyses what it means that “the law [...] gives reasons for action” and reframes the problem of “reason-giving force of the law” (Enoch 2011, 2). This aspect is explored “in the context of a more general theory of reason-giving” (ibid., 3). Enoch identifies three types of reasons for action: *purely epistemic*, *triggering*, and *robust* reason-giving. In the first case, an epistemic reason concretizes when you “indicate to me, or show me, a reason that was there all along, independently of your giving it to me” (ibid., 4), that is when you “call our attention to a reason for action that already applies to us” (Bix 2011, 412).[93] A triggering reason is given when “certain changes in non-normative facts can trigger reasons that already apply to us” (ibid., 413). The example is about the possible reduction in your milk consumption if “your neighbourhood grocer raised the price of milk” (Enoch 2011, 4). Such an increase could be seen as the reason prompting you “to reduce your milk consumption.” That is not so for Enoch, who claims that there is no new reason, simply the grocer’s manipulation of “non-normative circumstances in such a way” as “to trigger a dormant reason that was there all along, independently of the grocer’s actions. Arguably, you have a general reason (roughly) to save money” (ibid.).[94] The third type of reason, robust-giving, “a distinct phenomenon” (ibid.) that comes about when “someone’s statements or actions do not simply remind us of existing reasons, or trigger the effect of existing reasons, but create reasons that were not there before:” this is true in particular of “requests and commands” and of “promises or plans” (Bix 2011, 413).

[92] The correlation between authority and normativity has been questioned by the authors who support the thesis of the connection between law and morality. Nino (1985), for example, has argued that to derive duties from legal norms, working from a positivistic view, is to lapse into the naturalistic fallacy. As noted by Redondo (1999, 112), “on this basis, Nino concludes that it is reasonable to hold, against the thesis of authors like Joseph Raz, that legal provisions do not express operative reasons for justifying decisions, except when they are identified as moral judgments.” In the contemporary philosophy of law of positivist style, Raz’s theory of reasons has been almost exclusively discussed in relation to his conception of authority and not in relation to his conception of reasons. Even those, as Essert (2013) who criticizes the idea that law is directly reason-giving, accept the distinction between first-order and second-order (exclusionary) reasons and see the latter as “reasons to deliberate about other reasons in particular ways.” In this perspective, “to be normative,” law (“legal obligations”) needs “to make a difference in the structure or content of our deliberations:” this is possible on the basis of “considerations which make a practical difference in our deliberations without themselves being reasons for action: second-order reasons” (ibid., 27, 30).

[93] Bix (2011, 412–413) explains epistemic reasons through this example: “before I do something rash, you might remind me of my obligation to be a good role model to my child or to my students. This reason was always present, and your reminding me did not in any way change the reasons for action that apply to me, but you effectively helped me to (re-)discover those already-existing reasons.”

[94] Enoch (2011, 5) underlines that “examples of this triggering case are all around us.”

The decisive step in Enoch's argument is to consider legal reasons not as robust-giving (i.e. as providing new reasons for action) but as triggering reasons.[95] As noted, "he sees no basis for assuming that law always (or "necessarily") gives reasons for action (other than "*legal* reasons for action") (Bix 2011, 214).[96] In essence, Enoch greatly limits the role of law as a reason by reducing it to "non-normative 'triggers' to reasons for action that were always already there" (ibid.). However, he associates this reflection with that of Raz (2006, 1012–1013; 1020) and argues that "in the context of a discussion of authority—plausibly a particular instance of robust reason-giving—Raz […] clearly thinks that the reason-giving involved is an instance of (what I call) triggering reason-giving" (Enoch 2011, 10). In this way, the possibility of giving reasons is shifted to authority and, in some ways, to the reasons why one ought to obey authority.

7 Concluding Remarks

The analysis carried out here has sought to provide a picture of the different types of reasons and, with reference to normative ones, of the source of their capacity to provide reasons for action. Of normative reasons (which can be seen as different ontological entities—facts or mental states), it has been privileged the "weighing conception" (and the vision of the reasons as a pro tanto ones) as it seems the most appropriate in relation to the practical reasoning (and to the legal one). In the philosophy of law, this conception was accompanied by the identification of secondary reasons that should allow the explanation of the normative nature of law and of its authority. In this context, however, the distinction between first-order and secondary reasons was generally accepted and the analysis and research have been mainly focused on the problem of authority of law.

References

Alvarez, M. 2010. *Kinds of reasons: An essay on the philosophy of action*. Oxford: Oxford University Press.

Alvarez, M. 2016a. Reasons for action: Justification, motivation, explanation. In *The Stanford encyclopedia of philosophy,* ed. E. Zalta. https://plato.stanford.edu/archives/win2016/entries/reasons-just-vs-expl.

Alvarez, M. 2016b. *Reasons for action, acting for reasons, and rationality*. https://kclpure.kcl.ac.uk/portal/files/46560247/art_3A10.1007_2Fs11229_015_1005_9.pdf.

[95] Enoch (2011, 8) seems to reduce almost all robust reason-giving to triggering reason-giving: "we must conclude that any case of robust reason-giving is really a case of the triggering of a conditional reason." For a critical examination of Enoch's analysis, see Rodriguez-Blanco (2013).

[96] Enoch (2011, 15) underlines that "we must distinguish, in particular, between the claim that the law gives *legal* reasons for action and the claim that the law gives (genuine, unqualified) reasons for action."

Audi, R. 2010. Reasons for action. In *The Routledge companion to ethics*, ed. J. Skorupski. Abington, Oxon: Routledge.

Bix, B.H. 2011. The nature of law and reasons for action. *Problema. Anuario de Filosofía y Teoría del Derecho*, 5: 399–415. http://www.redalyc.org/pdf/4219/421940003018.pdf.

Blackburn, S. 2010. *The Majesty of reason*. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/S0031819109990428.

Bratman, M. 1987. *Intention, plans, and practical reason*. Cambridge: Cambridge University Press.

Bratman, M. 2014. *Shared agency. A planning theory of acting together*. Oxford: Oxford, Oxford University Press.

Broome, J. 2004. Reasons. In *Reason and value: Themes from the moral philosophy of Joseph Raz*, ed. J. Wallace, P. Pettit, S. Scheffler, and M. Smith, 56–90. Oxford: Oxford University Press.

Broome, J. 2013. *Rationality Through Reasoning*. Oxford: Wiley-Blackwell.

Buckareff, A. 2014. Review of Reasons and causes: Causalism and anti-causalism in the philosophy of action, ed. G. D'Oro and C. Sandis. Basingstoke: Palgrave Macmillan, 2013. *Notre Dame Philosophical Reviews*. http://ndpr.nd.edu/news/reasons-and-causes/.

Chang, R. 2004. Can desires provide reasons for action? In *Reason and value: Themes from the moral philosophy of Joseph Raz*, ed. J. Wallace, P. Pettit, S. Scheffler, and M. Smith, 56–90. Oxford: Oxford University Press.

Crisp, R. 2014. Keeping things simple. In *Weighing and reasoning: Themes from the philosophy of John Broome*, ed. I. Hirose, and A. Reisner, 140–154. Oxford: Oxford University Press.

Dancy, J. 1993. *Moral reality*. Oxford: Blackwell.

Dancy, J. 2000. *Practical reality*. Oxford: Clarendon Press.

Dancy, J. 2004a. *Ethics without principles*. Oxford: Oxford University Press.

Dancy, J. 2004b. Enticing reasons. In *Reason and value: Themes from the moral philosophy of Joseph Raz*, ed. J. Wallace, P. Pettit, S. Scheffler, and M. Smith, 1–18. Oxford: Oxford University Press.

Darwall, S. 1992. Internalism and agency. *Philosophical Perspective (Ethics)* 6: 155–174.

Davidson, D. 1963. Actions, reasons, and causes. *The Journal of Philosophy* 60 (23): 685–700.

Dworkin, R. 1977. *Taking rights seriously*. Cambridge, MA: Harvard University Press.

Ehrenberg, K. 2011. Critical reception of Raz's theory of authority. *Philosophy Compass* 6: 777–785.

Enoch, D. 2011. Reason-giving and the law. *Oxford Studies in the Philosophy of Law* 1: 1–38. https://ssrn.com/abstract=2607030.

Enoch, D. 2014. Authority and reason-giving. *Philosophy and Phenomenological Research* 89: 296–332. https://ssrn.com/abstract=2606995.

Essert C. 2013. *Legal obligation and reasons*. https://law.queensu.ca/sites/webpublish.queensu.ca.lawwww/files/files/Faculty&Research/FacultyProfileDocuments/EssertLegalObligationandReasons.pdf.

Finlay, S., M. Schroeder, 2012. Reasons for action: Internal vs. External. *The Stanford encyclopedia of philosophy,* ed. E. Zalta. https://plato.stanford.edu/entries/reasons-internal-external/.

Hart, H.L.A. 1961. *The concept of law*. Oxford: Clarendon.

Hieronymi, P. 2005. The wrong kind of reasons. *The Journal of Philosophy* 9: 437–457.

Hieronymi, P. 2013. The use of reasons in thought (and the use of earmarks in arguments). *Ethics* 1: 114–127.

Horty, J.F. 2007. Reasons as defaults. *Philosophers' Imprint* 3: 1–26. http://www.umiacs.umd.edu/~horty/articles/007003.pdf.

Mele, A. 2003. *Motivation and agency*. Oxford: Oxford University Press.

Nagel, T. 1970. *The Possibility of altruism*. Princeton: Princeton University Press.

Nino, C.S. 1985. *La validez del derecho*. Buenos Aires: Astrea.

O'Connor, T. 2010. Reasons and causes. In *A companion to the philosophy of action*, ed. T. O'Connor, and C. Sandis, 129–138. Oxford: Wiley-Blackwell.

Parfit, D. 1984. *Reasons and persons*. Oxford: Clarendon Press.

Parfit, D. 2011. *On what matters*, vol. 1. Oxford: Oxford University Press.

Perry, S.R. 1989. Second-order reasons, uncertainty and legal theory. *Faculty Scholarship. Paper* 1354. http://scholarship.law.upenn.edu/faculty_scholarship/1354.

Raz, J. 1979. *The authority of law: Essays on law and morality*. Oxford: Oxford University Press.

Raz, J. 1986. *The morality of freedom*. Oxford: Clarendon.

Raz, J. 1999a. *Practical reasoning and norms*. Oxford: Oxford University Press (1st ed., 1975).

Raz, J. 1999b. *Engaging reason: On the theory of value and action*. Oxford: Oxford University Press.

Raz, J. 2006. The problem of authority: Revisiting the service conception. *Minnesota Law Review* 9: 1003–1044. https://ssrn.com/abstract=999849.

Raz, J. 2009. Reasons: Practical and adaptive. In *Reasons for action*, ed. D. Sobel, and S. Wall, 37–57. Cambridge: Cambridge University Press.

Raz, J. 2011. Reason, rationality, and normativity. In Id., *From normativity to responsibility*. Oxford: Oxford University Press.

Redondo, M.C. 1999. *Reasons for action and the law*. Dordrecht: Springer.

Ridge, M. 2011. Reasons for action: Agent-neutral vs. Agent-relative. *The Stanford encyclopedia of philosophy,* ed. E. Zalta. https://plato.stanford.edu/archives/win2011/entries/reasons-agent.

Ryle, G. 1949. *The concept of mind*. London: Hutchinson.

Robertson, S. 2009. Introduction: Normativity, reasons, rationality. In *Spheres of reason: New essays in the philosophy of normativity*, ed. S. Robertson, 1–28. Oxford: Oxford University Press.

Rodriguez-Blanco, V. 2013. Reasons in action v Triggering-reasons: A reply to Enoch on reason-giving and legal normativity. *Problema: Anuario de Filosofía y Teoría del Derecho* 7: 3–25. http://www.redalyc.org/articulo.oa?id=421940005001.

Scanlon, T.M. 1998. *What we owe to each other*. Cambridge, MA: Belknap Press of Harvard University Press.

Scanlon, T.M. 2014. *Being realistic about reasons*. Oxford: Oxford University Press.

Schroeder, M. 2007. *Slaves of the passions*. Oxford: Oxford University Press.

Searle, J. 2001. *Rationality in action*. Cambridge, MA: The MIT Press.

Smith, M. 1987. The humean theory of motivation. *Mind* 96: 36–61.

Skorupski, J. 2010. *The domain of reasons*. Oxford: Oxford University Press.

Sobel, D., and S. Wall. 2009. Introduction. In *Reasons for action*, ed. D. Sobel, and S. Wall, 1–12. Cambridge: Cambridge University Press.

von Wright, G.H. 1963. *Norm and action: A logical enquiry*. London: Routledge and Kegan Paul.

White, A. (ed.). 1968. *The philosophy of action*. Oxford: Oxford University Press.

Reasons in Moral Philosophy

Carla Bagnoli

While the concept of reason is pervasive in our ordinary practices, there is a large and divisive disagreement about the role of reasons in moral philosophy. Such disagreement depends on three related issues, which concern the definition of “moral reasons,” their sources, and functions.

1 What Is a Moral Reason?

As a working hypothesis, let us establish that a reason is a consideration that counts in favor of something (Scanlon 1998). Typically, there are some facts that count toward believing or acting in a given way. The fact that Laura is an adult Italian citizen counts in favor of believing that she holds voting rights in Italy, and this counts in favor of treating her as entitled to vote in the upcoming elections. What is a *moral* reason? There are two ways to approach this question. On the *material definition*, a reason is moral insofar as it represents some moral facts or moral properties. That is, reasons are moral because their contents are distinctive and peculiar; i.e., they are specifically moral contents. For instance, considerations about harming people or enhancing persons, respecting or undermining their autonomy are typically considered to be moral reasons. How to account for the specific content, import, and strength of such claims is a question for normative ethics. For instance, Kantian theories prohibit harm on the basis of the principle of respect for the dignity of humanity. Utilitarian theories prohibit harm insofar as it does not maximize the utility of all sentient beings (Mill

C. Bagnoli (✉)
Dipartimento di Studi Linguistici e Culturali, Università di Modena e Reggio Emilia, Modena, Italy
e-mail: carla.bagnoli@unimore.it

C. Bagnoli
University of Oslo, Oslo, Norway

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_2

1998). According to virtue ethics, malice is a vice that undermines human flourishing. Such theoretical accounts provide different justifications for their claims, but they often converge on the kind of considerations regarded as morally relevant.

On the *formal definition*, instead, moral reasons are not identified by specific moral contents, but by their internal structure (Frankena 1958). For instance, according to Kantian theories, a consideration counts as a moral reason if it can be the agreed by ideal agents engaged in the activity of co-legislation. This definition does not separate moral reasons from other sorts of reasons, such as epistemic reasons for believing a certain proposition or aesthetic reasons for valuing a certain object. In fact, the underlying claim is that there is no sharp line dividing moral and non-moral reasons. Universality is the formal property of all kinds of reasons, which means that to reason at all we have to be guided by some principles. This is not to deny that there are some moral contents that qualify as moral reasons. The point is that such contents count as moral reasons because of some structural feature such as their function and constitutive aim, rather than in virtue of the fact that they represent a portion of reality. The appeal to the formal structure of moral reasons evokes the metaphysical contrast between form and matter, but it may be interpreted in a way that does not commit to any metaphysics. On this reading, the appeal to the structural arrangement of reason indicates that some sort of consideration counts as a moral reason in virtue of its rational justification. Since rational justification is justification by principles, to count as a reason, any consideration must be principled (i.e., universalizable).

The formal and material approaches to the definition of reasons may converge on the scope of moral reasons, but they importantly differ as to how to account for what makes a consideration a moral reason. For instance, virtue ethics and Kantian ethics agree that there is a moral reason to tell the truth, but the former holds that this is because truth-telling represents a virtue, and the latter holds that this is because lying cannot serve as a viable principle of a universal co-legislation. More fine-grained distinctions emerge when we take into account the more specific functions of reasons.

2 Explanatory and Normative Reasons

Typically, we invoke moral reasons to explain or justify our actions, attitudes, or beliefs. Explaining and justifying are two basic functions of moral reasons, and correspondingly, we may draw the following distinction. *Explanatory* reasons are reasons that make one’s attitude or action intelligible to ourselves and to others. *Normative* reasons are considerations that and guide the agent in deciding what to do. For present purposes, we refer to actions in a broad sense, which is inclusive of all cognitive and affective rational activities in which we engage. Beliefs, feelings, emotions, and attitudes belong to this category, even though their constitutive aims and criteria of success differ from the ones pertaining to the performance of an action. Feelings and beliefs are neither decided nor deliberated in the same ways some actions are, but they can be reflectively endorsed on the basis of reasons.

The distinction between normative and explanatory reasons is not meant to be mutually exclusive. In fact, in order to explain action, we often cite normative reasons for action. Explanatory reasons thus often serve as rationalizations of action (Anscombe 1957). In this case, normative reasons are retrospective and explain past actions rather than guiding prospective actions. According to some, the two sorts of reasons are closely, if not conceptually, connected.

When we attempt to explain an action, we attribute normative reasons to the agent. Typically, observers deploy explanatory reasons and refer to considerations that presumably guided the agent in deciding what to do. For instance, Claire thinks that Fabien refuses to raise a family because he is afraid of losing personal autonomy. The observer's explanation of the action succeeds if it makes the action intelligible on the basis of the agent's reasons for performing it. Claire's consideration makes sense of Fabien's actions.

Of course, Fabien may disagree that his decision to raise a family is based on fear. The observers' explanations do not necessarily coincide with the agent's first-person explanation of the same act. It is a platitude that observers and agents often disagree about the explanatory reasons of the agent's action, but this disagreement is philosophically interesting. On the one hand, agents seem to have a special kind of authority about their own actions insofar as they are their authors. On the other hand, agents do not always know more than their observers about the mechanics of their own action. This is a *problem of opacity* of explanatory reasons. Despite Fabien's protest to the contrary, Claire may be right that fear is the real cause of his decision because Fabien is in a state of denial or is self-deceptive. Claire's attribution is correct, and Fabien is not. To capture these disagreements and investigate their philosophical implications, it is useful to distinguish between *attributive* and *operative* reasons. In this case, Fabien's operative reason is fear.

3 The Issue of Agential Authority

The phenomenon of opacity tells us that observers may be better positioned to attribute correctly causes to agents. In the previous example, Claire appreciates and identifies Fabien's operative reasons better than Fabien himself. However, it is questionable whether observers have the authority to explain actions by correctly identifying their motivational causes. While Claire may correctly classify or understand Fabien's decision to raise a family as caused by fear, it is still Fabien's own exclusive business whether to raise a family or not. This is because normative reasons are importantly related to *reasons for action*. They are important reasons for doing (or omitting) something or for undertaking (or disavowing) some attitudes or other. Fabien's reasons not to raise a family are indeed in a distinctive sense his own reasons. Agents have a special claim on their actions: This is called *first-person or agential authority* (Anscombe 1957; Chap. 4). When we raise questions about what to do, we are engaging action as agents, rather than as bystanders. We are asking for considerations that we endorse as reasons for action. Such endorsement is open exclusively to the prospective agents of the action, and in this specific sense, agents have a special claim on action. That is, agents exercise a special authority on their actions insofar as actions are theirs. The problem is to explain why and how so.

There are competing views of agential authority. According to Kantian constructivism, agential authority implies autonomy, the capacity for self-government, and normative guidance by principles. On an alternative view, autonomy is afforded by reflective endorsement, which is neither principled nor deliberate. Rather, it is a granted by discrete acts of wholehearted identification, which determine the will as well as the bounds of the person (Frankfurt 1988, 1999).

There are also different views about where to situate the philosophical problem of agential authority. According to G.E.M. Anscombe, the problem is both epistemic and practical. The first-person perspective is the perspective of practical knowledge, as opposed to the speculative knowledge of action as an outward performance (Anscombe 1957). By contrast, Christine M. Korsgaards insists that the role of normative reasons is better understood as deliberative, rather than epistemic (Korsgaard 2008, 310–317). It is at the time of deliberation that the question arises whether reasons are efficacious. Normative reasons exert their authority directly, through action itself (Korsgaard 2008, 317). Instead of asking how to put normative truths in practice, it is better to focus on the mechanics of authoring and authorizing action, which pertains to the agent's own perspective.

The issue of agential authority is rooted in the subjective perspective of the agent. Therefore, it raises important issues about the objectivity of reasons for action, their impersonality and impartiality.

4 Subjective and Objective Reasons

Subjective normative reasons are those considerations that the agent takes as relevant because of her partial understanding of the situation, driving interests, and limited information. For instance, Marc justifies his policy of raw-eating on the basis of two kinds of considerations: He does not like cooking and believes that this policy is more ecological in that it has no impact on the planet. These are Marc's subjective reasons for eating raw food. In case Marc's belief that raw food is more ecological is correct, this consideration is also an objective reason to eat raw. However, if this information turns out to be incorrect and actually Marc is wrong in believing that the policy of eating raw food has no impact on the planet, his subjective reasons may still hold. That is because, subjectively, he may still have subjective reasons to eat raw food given what he knows. He may simply be misinformed.

The distinction between subjective and objective reasons comes in degrees, and the two classes of reasons are not mutually exclusive; indeed, they may overlap in most cases. More complex are those cases where subjective and objective reasons prescribe incompatible courses of action. If Marc has a strong preference for raw food, but this kind of food is not healthy, he is criticizable, although he may not be blameworthy for his false belief. A further question is, whether there is a reason to question his position and correct him and if so on which authority.

It might seem that it is a requirement of rationality that subjective reasons should be abandoned or revised when they are shown to be wrong. However, there is a deeper and more general issue at stake concerning the relation between objective and subjective reasons, especially when values are involved. Subjective reasons are in some important sense "reasons." They are not merely illusory or defective. Rather, they are considerations that spring from the agent's own interests and understanding of the situation. For instance, even though Fabien misunderstands the role of fear in his own situation, his subjective reasons may coincide with the operative reason not to raise a family. His account of the situation is perspectival and yet pragmatically fit.

Because rational justification is driven by the quest for objectivity, it is often ignored that partiality plays a large role in the rational assessment of the situation (Nagel 1986). As it emerges in the case of epistemic subjective reasons, however, speaking from a specific perspective is not necessarily an epistemic defect, but an evaluative component of some kinds of reasons (ibid. Elgin 2017). In fact, some core moral values are for their own nature partial and perspectival and such that they do not command universal endorsement. Perhaps, the most paradigmatic case is love. Reasons for love are distinctively special and partial (Frankfurt 2004). They are not meant to elicit universal endorsement. Reasons for love may be said to exhibit a very peculiar singularity, and a common proof for this claim is that lovers are invariably at loss to explain their love in principled terms. Giorgio may be able to cite some characteristic features of Budapest's baths as reasons for loving Hungary rather than Tuscany, but they would hardly be considerations for convincing anybody else that they should love Hungary rather than any other place exhibiting the same relevant features. Likewise, special commitments and personal bonds generate reasons that are not universal in scope and authority. Reasons that derive from personal commitments, political ideals, friendships, and loving relations seem insulated from the requirement of universal authority that applies in the case of objective reasons. Perhaps more importantly, to adequately account for protecting moral values such as love and friendship, it seems that, the independence of subjective reasons should be preserved.

5 Personal and Impersonal Reasons: Integrity and Authenticity

This is where the distinction between subjective and objective reasons intersects another important distinction between *personal* and *impersonal* reasons. Some moral reasons spring from special relations and are rooted in special concerns. Moral obligations we have to our families, friends, and fellow citizens are of this kind, and their compellingness and import do not seem to compare to obligations we have toward strangers. For some, these *special* obligations represent the core of morality; for others, morality in the narrow sense coincides instead with the obligations we have toward any other persons, groups or peoples, regardless of special bonds. While moral obligations are often born as burdensome constraints on our action, requests that undercut our projects and concerns, or external impositions that undermine our personality, special obligations are perceived as the very stuff of moral life and appreciated as crucial modes of expressing our integrity and authenticity. Moral decisions and personal preferences are the axes of morality. Personal decisions show that the action is authentically the agent's own action, rather than something imposed from an alien source of authority. This is an important aspect of autonomy, which distinguishes actions that the agent decided to perform because they are expressive of her character and integrity, from those in which she plays only or mostly a causal role, as it happens when the agent is coerced, threatened, or acts unwillingly to avert a greater evil.

To vindicate this aspect, it seems that one should resist the view that objectivity requires the moral agent to overcome the partialities of the personal stance as immoral biases. Furthermore, the presumption that the subjective stance on moral value is a defect to be corrected by broadening its scope and reach an objective standpoint is problematic. Some philosophers argue that the subjective and personal nature of these reasons matter more than their objective vindication (Williams 1981c, Wolf 1997). When moral reasons are completely alien to the agent's own deliberative set or undermine the agent's integrity by undercutting all her subjective reasons, they can hardly exert rational authority over them. This approach raises the question of the place of moral reasons in our life. Critics argue that to attribute overridingness to moral reasons impoverishes character and undermines the richness of personal relations, producing a dull and single-minded agent, or so the objection goes. In response to this objection, philosophers have insisted on the continuity between moral and practical reason (Annas 1993, Engstrom 2009).

6 Drawing the Boundaries of the Moral Domain

How to draw the boundaries of the moral domain is a very controversial philosophical question which has important implications. While we often refer to "common morality," it is arguable that this expression must be understood indexically, that is, as referred to the specific morality held at a specific time and place, by some society. This *descriptive use* of the term "morality" is prevalent in anthropology and comparative and evolutionary psychology (De Waal 1996, Sinnott-Armstrong 2008). The implication of this definition is that there is no universal and universally authoritative body of moral cognitions or moral norms. By contrast, on some *normative use* of the term "morality," it refers to a body of moral cognitions or norms that are available to and binding for all relevant agents. That is, ideally or under specified conditions, all relevant agents are guided by such moral norms. According to the normative understanding, moral norms are typically considered impartial, overriding non-moral considerations and universally authoritative and binding for all relevant agents (Kant 1967, Darwall 1983, O'Neill 1989). Moral theories differ as to the account of the source of authority and the contents of such moral norms. They also importantly differ as to the inclusiveness of the class of relevant agents. For some, moral norms apply to all rational beings. For others, moral norms apply only to human rational agents because of their peculiar and distinctive epistemic and practical limitations, such as fallibility, frailty, and mutual vulnerability. Supporters of morality hold that it is a cooperative enterprise, which generally favors human flourishing and represents the rationally best way to deal with problems such as distributing scarce resources (Baier 1958, Rawls 1971, Frankena 1973, Gauthier 1986). Debunkers hold that morality is an instrument in the service of some groups, which manipulate their partners in order to promote their own specific interests, or else it is a system of blame provided by natural selection. In any case, moral judgments do not have any special authority, but they are the expression of particular perspectives, visions, or projections In particular, moral judgments lack both ontological support and rational authority, so that their apparent inescapability or necessity does not sustain close investigation (Mackie 1977, Joyce 2001).

7 Moral Reasons and Moral Reasoning

Moral theories not only differ in their account of the contents of moral reasons, but also in their account of how moral reasons are produced or recognized. On the *recognitional view*, moral reasoning aims at recognizing what there is moral reason to do, to believe or to feel. It starts with some moral premises and ends with conclusions about what to do, to feel, or to believe. On the *constructivist* view, instead, reasoning builds up reasons according to some procedure, against the background of a conception of moral agency and relevant facts of the matter. A further important question is whether and how moral reasoning relates to practical reasoning. According to Aristotelian views, moral reasoning is part and parcel of a general account of practical reasoning, which amounts to the specification of ends. The wise is capable of recognizing the good ends because of their adequate upbringing, but the structure of their reasoning is not too different from the vicious persons. The wise and the vicious have different ends, but they determine their ends in a similar manner. However, the wise is capable of harmonious integrity because their ends fully cohere, while the vicious are always at risk of disintegration because their ends cannot integrate and push apart (Engstrom 2009). Kantians offer a unified account of practical reasoning, whose universal structure mirrors theoretical reasoning. However, there is an important distinction between *moral* and *prudential reasoning*. Prudential reasoning is hypothetical because it is conditional on some particular ends that the agent actually holds; hence, it does not produce unconditional obligations that apply to all rational beings. By contrast, moral reasoning does not depend on any specific end, but it is modeled by universal co-legislation. The difference between these two kinds of reasoning concerns their authority. While the conclusions of moral reasoning hold universally and unconditionally for all relevant agents who are represented as co-legislators, the conclusions of prudential reasoning hold for the narrow group of agents who share in their specific ends or interests (Korsgaard 1997). Furthermore, the authority of prudential or instrumental reasoning depends on the authority of moral reasoning. That is to say that instrumental reasoning carries normative significance only against the background of a general account of practical reasoning. Not all agree that moral reasons are unconditional and necessary, as Kantians do. Humeans hold that all moral reasons are hypothetical, and not structurally different than other sorts of practical reasons, such as etiquette (Foot 1978, McDowell 1978).

8 Moral Reasons in Conflict

Moral reasons are generally thought to exhibit a distinctive kind of gravity and importance. This is often explained with the philosophical claim that they represent an eminent domain of objects, e.g., values of a particularly dignified sort. Moral reasons are both unconditionally authoritative and rationally overriding. This view is widely challenged, as the fortune of Williams' argument about internal reasons shows. The issue of the normative force of moral reasons importantly relates to the phenomenon of moral conflict, which ramifies through personal and interpersonal dimensions.

In the intra-personal case, moral reasons may clash with non-moral or prudential reasons. As the Kantian case of the prudent merchant shows, moral reasons need not be always in contrast to prudential reasons: Sometimes they converge on recommending the same line of action. It might be prudent to price one's merchandize fairly in order to keep one's clients. In other cases, however, it might be more profitable to exploit, and these are cases of moral conflict. According to Kantian theories, moral reasons are rationally overriding in deliberation in that they always trump non-moral preferences, interests, and desires (Kant 1967, Korsgaard 1996). However, this is not to say that interests, desires, and preferences do not provide reasons for action. On the contrary, the claim is that they generally do, and when they clash with moral reasons, they fail to provide definitive reasons for action. Other moral theories admit of alternative ways to treat the relation among moral and non-moral reasons for action, which include overridingness, weakening, outweighing, annulling, defeating, and neutralizing (Nozick 1968, 30–35). When defeated or undermined, moral reasons leave a remainder which generates further moral reasons for compromise, reparation, and compensation. Furthermore, for some moral theories, moral learning through reasoning and experience allows for a variety of ways in which moral and non-moral reasons align and can be made cohere and integrate into time. Moral education is thus an important aspect of a theory of moral reasons in that it broadens and deepens the ways of coping with moral conflicts.

This diachronic dimension of moral reasoning is particularly relevant in the account of interpersonal moral conflicts, where the parties disagree not only because their interests clash but also because their values do. Conflicts of this kind are pervasive in a pluralistic society, and it is part of democratic agenda to understand how to legitimately address such conflict. If pluralism is a value to be protected and fostered, some conflicts and disagreements ought to be preserved (Williams 1981a). Yet it must be possible to find ways of accommodating such conflicts and disagreements without undermining the civic structure of the society (Rawls 1993, Nozick 1968). In such cases, norms of basic rationality help citizen to deal with conflict of moral reasons, without entering the underlying dispute about values. This view seems to imply that all rational agents share the same basic norms of rationality, which allows that to share a conception of "public reason" (Rawls 1993, Rawls 1999). However, it is arguable that norms of rationality do not guide us independently of commitment to any specific values. In fact, some hold that even norms of rationality deeply depend on more fundamental special identities, and therefore, they are conditional requirements that depend on sharing such special identities (MacIntyre 1988). To fruitfully address this crucial problem, it is helpful to distinguish claims of structural rationality, which depend on how we form reasons, and claims of substantive rationality, which vary according to specific normative theories (Scanlon 2007). This distinction allows us to identify different levels and dimensions of rational disagreements and investigate distinct modes of resolution.

9 Moral Reasons and Coordination

On a prominent view, morality is basically a cooperative enterprise aiming at coordination (Kant 1967, Hobbes 1994, Frankena 1973). This is the point of convergence between Hobbesian and Kantian traditions, insofar as they attempt to found moral obligations as rational requirements. However, Hobbesians take reason to be basically self-interested while Kantians hold that by undertaking reasoning one is also already committed to morality, in the minimal sense of a fundamental disposition to reason with others. Critics have argued that the Kantian solution is question-begging, since it assumes some basic moral commitment. But it is also, and more radically, open to debate that morality works as a coordination device, since moral differences and disagreements are pervasive and divisive. This latter worry can be addressed by distinguishing between some basic moral concern or moral dispositions (such as the requirement to reason with others) and the adoption of a full-fledged substantive morality or moral code. The claim that morality is a coordination device is compatible with the existence of different moral codes, traditions, and also with cultural change over time and institutional transition. Moreover, the emergence of moral codes and practices is amenable to evolutionary explanations (Gibbard 1990, Hauser 2006, Street 2006, Copp 2008, Fitzpatrick 2014). The basic point is that we exchange moral reasons with others in order to solve some coordination problems. To this effect, moral reasons importantly contribute to represent the problem at hand. These reasons concern not only actions to undertake, but also emotions to express and beliefs to endorse.

The importance of moral reasons as cooperative schemes in coordination problems may be taken as an aspect of a more general feature of reasoning. On some prominent views, reasoning is dialogical (Kant 1967, Habermas 1984). If we take moral reasoning to be characterized by universalization, then its structure is always *collective* as it is fundamentally reasoning among different personae. How to design the personae involved in collective moral reasoning is a philosophical question, which carries divisive theoretical and practical disagreements. The dialogical account of moral reasoning can be associated with different models of moral agents, with radically different results. This view importantly implies that there is no natural and trivial answer to the issue of the boundaries and the category of relevant agents involved; hence, the scope of applicability of moral obligations is not naturally defined.

10 Moral Reasons and Compliance

A significant problem related to the category of moral agency to which moral reasons apply concerns the issue of compliance. If moral obligations are rational requirements, is compliance with morality granted by rationality alone? The question is complex, and it ranges over two partially independent debates. The first debate concerns the source of authority of moral reasons and revolves around the distinction reason/sensibility; the second debate concerns motivational force of moral reasons and revolves around the internal/external reasons distinction. As for the debate concerning source of authority of moral reasons, we might distinguish two camps: rationalists and sentimentalists. On the rationalist view, moral reasons apply to all beings endowed with rationality and are universally binding and authoritative. Rational agents take moral reasons to be intrinsically authoritative and compelling. That is, they are guided by what is morally dutiful or virtuous independently of what is merely desirable or prudent. This is not to say that all rational agents always comply with morality, as they might be overpowered by external forces or interfered with. The point is that if all goes well, moral reasons are capable of driving rational action without further motivational aid. According to sentimentalists, instead, reasons are motivationally inert beliefs, which cannot compel action independently of the presence of desires. Alternatively, sentimentalists explain moral motivation in terms of moral sentiments, attitudes, dispositions, and desires that motivate the agent to act according to duty. This disagreement about the motivational import of moral reasons is thus rooted in a difference about the powers and scope of reason (Korsgaard 1996, Smith 2013). However, the question is not solved simply choosing between these two accounts of practical reason. To take morality as a requirement of rationality raises questions about how to make requirements compelling. On the externalist view, moral requirements are compelling insofar as they are combined with a desire to comply (or some other motivational force). This view makes the authority of moral reasons conditional on some external sanctions, such as blame and reprobation or some internal sanctions such as sense of guilt, remorse, or shame. On the internalist view, instead, moral requirements are motivating insofar as they are themselves normative (Nagel

1970). Partly, the question relates to whether reasons—understood as beliefs or principles—can generate desires or motivate action independently of desire (Parfit 1997, 2006). According to Williams (1981b, 101–113), all reasons are internal in that they are necessarily linked to the agent's actual deliberative set. Williams' argument that genuine reasons for action must be linked to what we care about challenges a too narrow and moralistic conception of practical reasoning (Williams 1981b, Frankfurt 2006). The intended effect of the argument is that there are no reasons that apply independently of the agent's special plans, motivations, and projects. However, there is always a question about how to integrate such projects and plans and coordinate with other relevant agents who are equally entitled to carry their own projects and plans. To adequately address such problems, it seems crucial to acknowledge the moral and political varieties of ends at stake in practical reasoning and refocus the debate on the resources for reasoning with others.

References

Annas, J. 1993. *The morality of happiness*. Oxford: Oxford University Press.

Anscombe, G.E.M. 1957. *Intention*. Oxford: Basil Blackwell.

Baier, K. 1958. *The moral point of view*. Ithaca, NY: Cornell University Press.

Copp, D. 2008. Darwinian skepticism about moral realism. *Philosophical Issues* 18: 186–206.

Darwall, S. 1983. *Impartial reason*. Ithaca, NY: Cornell University Press.

De Waal, F. 1996. *Good natured: The origins of right and wrong in humans and other animals*. Cambridge, Mass.: Harvard University Press.

Engstrom, S. 2009. *The form of practical knowledge*. Cambridge, MA: Harvard University Press.

Fitzpatrick, W. 2014. Morality and evolutionary biology. *Stanford Encyclopedia of Philosophy*. Available online at: https://plato.stanford.edu/entries/moralitybiology/.

Foot, P. 1978. Morality as a system of hypothetical imperatives. In *Virtues and vices*, ed. P. Foot. Berkeley, CA: University of California Press.

Frankena, W.K. 1958. Obligation and motivation in recent moral philosophy. In *Essays in moral philosophy,* ed. A.I. Melden, 40–81. Seattle, WA: University of Washington Press.

Frankena, W.K. 1973. *Ethics*. Englewood Cliffs, NJ: Prentice-Hall.

Frankfurt, H. 1988. *The importance of what we care about. philosophical essays*. Cambridge: Cambridge University Press.

Frankfurt, H. 1999. *Necessity, volition, and love*. Cambridge: Cambridge University Press.

Frankfurt, H. 2004. *The reasons of love*. Princeton, NJ: Princeton University Press.

Frankfurt, H. 2006. *Taking ourselves seriously and getting it right*. Stanford, CA: Stanford University Press.

Gauthier, D. 1986. *Morals by agreement*. Oxford: Oxford University Press.

Gibbard, A. 1990. *Wise choices, apt feelings*. Oxford: Clarendon Press.

Habermas, J. 1984. *Theory of communicative action,* trans. T. McCarthy. Boston, MA: Beacon Press. (1st ed., 1981).

Hauser, M. 2006. *Moral minds: How nature designed our universal sense of right and wrong*. New York, NY: Harper Collins.

Hobbes, T. 1994. *Leviathan,* ed. Edwin Curly. Indianapolis, IN: Hackett Publishing Company (1st ed., 1660).

Joyce, R. 2001. *The myth of morality*. Cambridge: Cambridge University Press.

Kant, I. 1967. *Groundwork of the metaphysics of morals.* New York, NY: Barnes & Noble. (1st ed., 1785).

Korsgaard, C.M. 1996. Skepticism about practical reason. In *Creating the Kingdom of Ends.* Cambridge: Cambridge University Press.

Korsgaard, C.M. 1997. The normativity of instrumental reason. In *Ethics and practical reason*, ed. Garrett Cullity, and Berys Gaut, 215–254. Oxford: Oxford University Press.

Korsgaard, C.M. 2008. *The constitution of agency*. Oxford: Oxford University Press.

MacIntyre, A. 1988. *Which justice? Whose rationality*. Notre Dame, Ind.: Notre Dame University Press.

Mackie, J.L. 1977. *Ethics: Inventing right and wrong*. London: Penguin.

McDowell, J. 1978. Are moral requirements hypothetical imperatives? *Proceedings of the Aristotelian Society Supplementary* 52: 13–29.

Mill, J.S. 1998. *Utilitarianism.* ed. Roger Crisp. New York: Oxford University Press. (1st ed., 1863).

Nagel, T. 1970. *The possibility of altruism*. Oxford: Clarendon Press.

Nagel, T. 1986. *The view from nowhere*. Oxford: Oxford University Press.

Nozick, R. 1968. Moral complications and moral structure. *Natural Law Forum* 13 (1): 1–50.

O'Neill, O. 1989. *Constructions of reason,* 206–218. Cambridge: Cambridge University Press.

Parfit, D. 1997. Reasons and motivation. *Proceedings of the Aristotelian Society Supplementary* 71: 99–130.

Parfit, D. 2006. Normativity. In *Oxford studies in metaethics,* ed. Russ Shafer-Landau, vol. 1, 325–380. Oxford: Clarendon Press.

Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.

Rawls, J. 1993. *Political liberalism*. New York, NY: Columbia University Press.

Rawls, J. 1999. The idea of an overlapping consensus. In John Rawls, *Collected Papers.* ed. S. Freeman, 421–448. Cambridge, MA: Harvard University Press. (1st ed., 1987).

Scanlon, T.M. 1998. *What we owe to each other*. Cambridge, MA: Harvard University Press.

Scanlon, T.M. 2007. Structural irrationality. In *Common minds: Themes from the philosophy of Philip Pettit*, ed. G. Brennan, R. Goodin, F. Jackson, and M. Smith, 84–103. Oxford: Oxford University Press.

Sinnott-Armstrong, W. (ed.). 2008. The evolution of morality: Adaptations and innateness. In *Moral psychology* 1. Cambridge, MA: MIT Press.

Smith, M. 2013. A constitutivist theory of reasons: Its promise and parts. *Law, Ethics, and Philosophy* 1, pp. 9–30.

Williams, B. 1981a. Conflicts of values. In *71-82*, ed. Moral Luck. Cambridge: Cambridge University Press.

Williams, B. 1981b. Internal and external reasons. *Moral luck*, 101–113. Cambridge: Cambridge University Press.

Williams, B. 1981c. Moral Luck. *Moral luck*, 29–37. Cambridge: Cambridge University Press.

Wolf, S. 1997. Moral saints. In *Virtue ethics,* ed. R. Crisp and M. Slote, 79–98. Oxford: Oxford University Press. (1st ed., 1986).

Legal Reasoning and Argumentation

Douglas Walton

Wigmore (1931) thought that there was a science of proof underlying legal reasoning. He thought this science of proof was inductive. Nowadays, there is much controversy and indeed much skepticism on the part of those in the legal profession about modeling legal reasoning as inductive using the Bayesian calculus to attach probability values to statements and evaluate legal reasoning using conditional probability (Tillers 1989). In the meantime, argumentation-based technology has provided qualitative methods that can be used to identify, analyze, and evaluate arguments. In particular, argumentation schemes, standardized argument patterns different from the familiar deductive and inductive models of reasoning, have proved useful for this purpose (Wyner and Bench-Capon 2007). These developments have lent support to Wigmore's view that there is a science of proof underlying legal reasoning different from deductive logic (Sartor 2005; Prakken 2005, 2006). In this chapter it is shown how recent advances in argumentation show the value of modeling legal reasoning in this new way.

In this chapter, legal reasoning is divided into two broad categories: (1) the kind of reasoning that applies rules to cases and (2) the kind of reasoning used to determine what the facts of a case are. In this chapter, it is shown how to apply several centrally important argumentation schemes to a procedural sequence of case-based reasoning in which there is a successive refinement of cases. This sequence has an opening stage where the ultimate *probandum* is stated, an argumentation stage where arguments on both sides are put forward, and a closing stage. It will also be shown that it is necessary to distinguish between explanation and argument to better appreciate the role of explanation in legal reasoning. As will be shown, inference to the best explanation is another type of reasoning commonly used in legal argumentation and important for understanding how it works (Josephson and Josephson 1994).

D. Walton (✉)
University of Windsor, Centre for Research in Reasoning,
Argumentation and Rhetoric (CRRAR), Windsor, ON, Canada
e-mail: waltoncrrar@gmail.com

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_3

The first section studies the kind of reasoning that applies rules to cases in law, including the following forms of reasoning: argument from an established rule, argument from a verbal classification, and argument from precedent. By using an example from a Supreme Court case summary, it is shown how all three kinds of reasoning can be combined into a chain of reasoning that represents the structure of the evidence used to support an ultimate conclusion to be proved in a case. In this section, it is shown how rule-based reasoning of this kind is more complex than it might initially seem. One reason is that reasoning from precedent depends on argument from analogy. Another reason is that as rules are applied to cases and analogies are made from a precedent case to a given case, the rules need to be continually modified as they are re-applied to the series of cases. The second section introduces the argumentation scheme for argument from analogy, shows how it is based on a notion of similarity between pairs of cases, and shows how case-based reasoning is based on a chained sequence of similarity reasoning by analogy in which cases are successively refined over the continuing sequence. The third section discusses the distinction between reasoning and argument. The following forms of reasoning are analyzed and discussed: practical reasoning, value-based practical reasoning, reasoning from lack of evidence, abductive reasoning, and argument from perception. The fourth section extends the analysis to two forms of reasoning that draw from inferences from sources, argument from witness testimony and argument from expert opinion. The fifth section shows how the structure of reasoning exhibited in the first four sections is that of defeasible logic. The sixth section shows how the notion of proof, including the notions of standard of proof and burden of proof, needs to be defined within a procedural context of argumentation that has the three main stages stated above. The final section contains the conclusion.

1 Forms of Reasoning by Applying Rules to Cases

Legal reasoning is often visibly based on a form of inference called argument from an established rule in the argumentation literature (Walton, Reed and Macagno 2008, 343).

> *Major Premise*: If carrying out types of actions including the state of affairs *A* is the established rule for *a*, then (unless the case is an exception), *a* must carry out *A*.
>
> *Minor Premise*: Carrying out types of actions including state of affairs *A* is the established rule for *a*.
>
> *Conclusion*: Therefore *a* must carry out *A*.

In this form of reasoning, *a* is a rational agent that is capable of carrying out goal-directed actions and recognizing the consequences of its actions. An agent also has the capability of feedback, that is, the capability of changing its actions depending on their perceived consequences. This form of reasoning also contains the assumption that the agent has a knowledge base containing a set of established rules. The old idea of mechanical jurisprudence considers the application of rules to cases as a straightforward application of deductive reasoning. The new approach of artificial intelligence and law sees the agent (judge or trier) as applying rules that can be defeated or overruled by exceptions as new evidence is introduced into its knowledge base.

Legal positivists, for example (Hart 1961), see law as consisting of two kinds of legal rules. The primary rules are the legal norms that regulate the activity of citizens and other persons. The secondary rules represent procedural norms that regulate the processes whereby the legislatures and courts put the primary legal rules into place and modify and apply them. But Hart recognized that both kinds of legal rules are inherently defeasible, meaning they admit of exceptions.

The term "defeasible" comes from medieval English contract law. It referred to a contract that has a clause in it that could defeat the contract in a case the circumstances of the case fit the clause. This meaning is now broadened to include the notion of a defeasible rule, a rule that is open to exceptions. Hart, in his famous paper "The Ascription of Responsibility and Rights" (1949, 1961), extended the usage of this term even further by writing about defeasible concepts. His most famous example is from *The Concept of Law* (1961). Consider the rule that no vehicles are allowed in the park. This rule could be defeated by special circumstances, for example during a parade, but it could also be defeated because of the open texture of the concept of a vehicle. Even though a car is classified as a vehicle, and would be excluded from the park, it may be debatable whether other objects such as a bicycle or a skateboard also fit into the same classification.

Consider the case of the drug-sniffing dog (Brewer 1996; Weinreb 2005). Suppose a trained dog sniffs luggage left in a public place and signals to the police that it contains drugs. Should this event be classified as a search according to the Fourth Amendment? If so, the evidence so obtained is not admissible as evidence. The problem is that the concept of a search is defeasible and law cannot define it by means of a set of necessary and sufficient conditions for closed to future revision because new cases may arise. Instead of providing closed essential definitions that give necessary and sufficient conditions, the best that can be done is to provide rules that may give necessary or sufficient conditions by indicating what types of things are included or excluded generally under the concept.

Weinreb (2005, 24) discussed two examples of such rules. One is the rule that if a police officer opens luggage and then observes something inside the luggage, the information collected is classified as a search. This is a narrow rule because it applies only to luggage, but it may still offer some helpful guidance in a case. The other is the rule that if a police officer obtains information about a person or thing in a public place without intrusion on the person or taking possession of or interfering with the use of the thing, it is not classified as a search. This rule is more general, but depends on the meaning of the prior concept of an intrusion on the person, as well as other concepts like that of taking possession of something or interfering with the use of something. These terms will, of course, also be open-textured and could possibly be subject to disputation. Questions arise quite often concerning how actions, events, and objects should be properly classified. In the example where the trained dog sniffs luggage left in a public place and signals to the police that it contains drugs, the question is whether this event should be classified as a search according to the Fourth Amendment.

Arguing from a legal classification is a special form of reasoning in its own right. The following scheme represents argument from verbal classification (Walton, Reed and Macagno 2008, 319). Here, the constant *a* represents an individual that can be an object of any kind, including an event, a physical object, an animal, or a human being.

> *Individual Premise*: *a* has property *F*.
>
> *Classification Premise*: For all *x*, if x has property *F*, then *x* can be classified as having property *G*.
>
> *Conclusion*: *a* has property *G*.

An ontology, a framework that specifies and organizes classes of concepts that can be used to represent the important features of cases (Ashley 2009, 8), can be brought forward to represent classifications of concepts to support legal reasoning about claims and issues. It includes representation of actual concepts like "animal," as well as legal concepts like "possession." In order to see how argumentation-based theories of legal reasoning in artificial intelligence would model the reasoning in a typical common law case, in addition to the scheme for argument from an established rule and the scheme for argument from verbal classification, we also need to take into account the working of a third scheme, called argument from precedent.

Most notably in common law countries, a ruling on a case is influenced by precedents. The most common type of argument from precedent used in legal reasoning applies to a current case and a prior case that has already been decided where the ruling can be applied to the current case (Schauer 1987). The argumentation scheme appropriate for this type of argument is the one for argument from precedent (Walton, Reed and Macagno 2008, 72).

> *Previous Case Premise*: The source case is a previously decided case.
>
> *Previous Ruling Premise*: In the source case, rule *R* was applied and produced finding *F*.
>
> *New Case Premise*: The target case is a new case that has not yet been decided.
>
> *Similarity Premise*: The target case is similar to the source case in relevant respects.
>
> *Conclusion*: Rule *R* should be applied to the target case and produce finding *F*.

This way of configuring argument from precedent makes it a species of argument from analogy (Macagno and Walton 2009). In the next section, we will see how argument from classification is an extension of argument from analogy typically used in many arguments from precedent.

The following example can be used to illustrate how forms of reasoning like argument from an established rule and argument from a verbal classification can be used to form a chain of reasoning in a legal case that has the claim at issue in the case as its ultimate conclusion to be proved. In this US Supreme Court case (CSX Transportation, Inc. v. Alabama Department of Revenue et al. certiorari to the US Court of Appeals for the eleventh circuit No. 09-520, decided February 22, 2011) CSX claimed that the State of Alabama had discriminated against them (http://www.supremecourt.gov/opinions/10pdf/09-520.pdf). The State taxes diesel fuel consumed by railroads but exempts interstate motor and water carriers. CSX claimed that this tax scheme discriminates against railroads in violation of the Railroad Revitalization and Regulatory Reform Act of 1976 which bars discriminatory taxation. The trial summary quoted below gives a concise account of the chain of reasoning used by the court to arrive at its decision.

> The key question thus becomes whether a tax might be said to "discriminate" against a railroad under subsection (b)(4) where the State has granted exemptions from the tax to other entities (here, the railroad's competitors). Because the statute does not define "discriminates," the Court again looks to the term's ordinary meaning, which is to fail to treat all persons equally when no reasonable distinction can be found between those favored and those not favored. To charge one group of taxpayers a 2% rate and another group a 4% rate, if the groups are the same in all relevant respects, is to discriminate against the latter. That discrimination continues if the favored group's rate goes down to 0%, which is all an exemption. To say that such a tax does not "discriminate" is to adopt a definition at odds with the word's natural meaning. This Court has repeatedly recognized that tax schemes with exemptions may be discriminatory. See, e.g., Davis v. Michigan Department of Treasury, 489 U. S. 803. And even Department of Revenue of Ore. v. ACF Industries, Inc., 510 U. S. 332, on which the Eleventh Circuit heavily relied in dismissing CSX's suit, made clear that tax exemptions "could be a variant of tax discrimination." Id., at 343. In addition, the statute's prohibition of discrimination applies regardless whether the favored entities are interstate or local. The distinctions drawn in the statute are not between interstate and local actors, as Alabama suggests, but between railroads and all other actors, whether interstate or local.

The central reasoning in this case can be visually represented as an inverted tree structure with the ultimate conclusion at the top, with the relevant arguments used to support that conclusion represented below in the argument diagram. The text boxes represent the statements that are the premises and conclusions of the arguments. The rounded nodes represent arguments, and the lines joining the nodes to text boxes represent inferences from premises to conclusions. In many instances, the conclusion of one argument becomes a premise in the next one, producing a chain of argumentation terminating in the root of the tree at the top.

We can see that in some instances an argument is represented as having one premise, while in other instances an argument has multiple premises. In some cases, we have two or more separate arguments going to support the same conclusion. For example, the two arguments for established rule just under the conclusion that attacks might be said to discriminate against a railroad under subsection, etc., each independently support it. This type of structure is sometimes called a convergent argument. Where we have two or more premises that are combined together to support the same conclusion so that each premise needs the others, this configuration is called a linked argument. The argument from the rule of definition at the top is a case in point.

To keep the argument diagram in Fig. 1 as clear and concise as possible at this point, no implicit premises have been represented. For example, three arguments just under the main conclusion have implicit premises, but these premises are not represented on the diagram. Argumentation schemes representing typical forms of defeasible reasoning are placed in all of the argument nodes in the diagram except one.

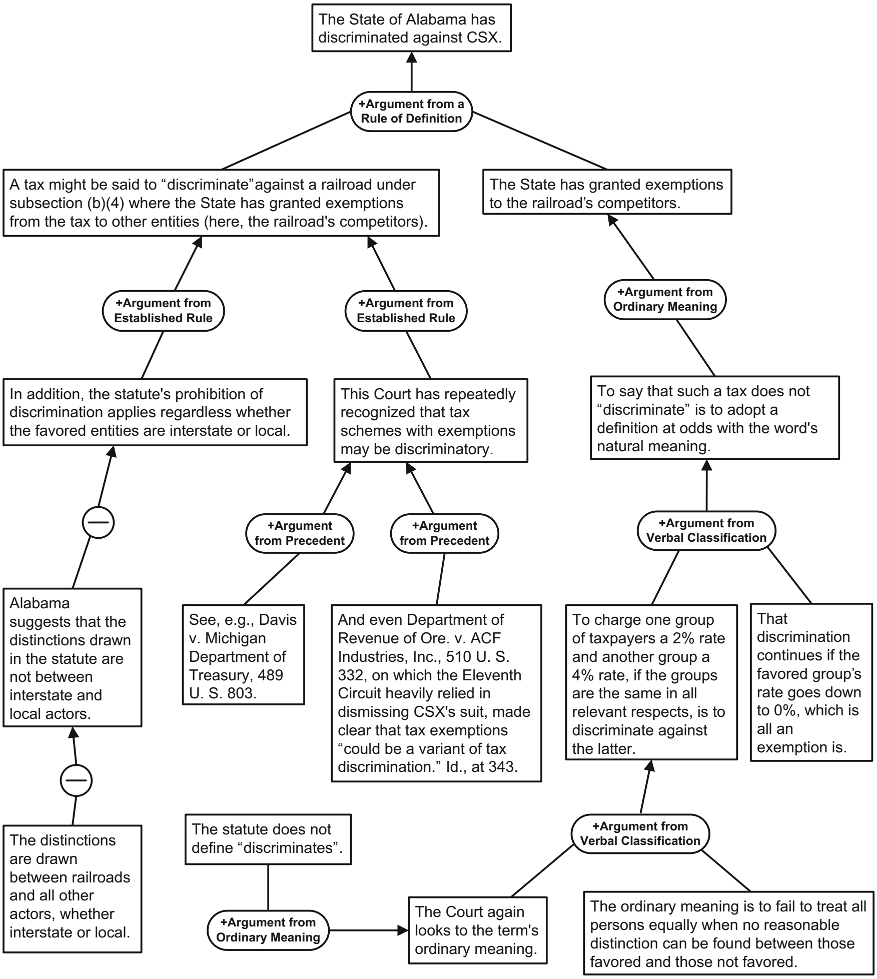

The State of Alabama has discriminated against CSX.
+Argument from a Rule of Definition
A tax might be said to "discriminate"against a railroad under subsection (b)(4) where the State has granted exemptions from the tax to other entities (here, the railroad's competitors).
The State has granted exemptions to the railroad's competitors.
+Argument from Established Rule
+Argument from Established Rule
+Argument from Ordinary Meaning
In addition, the statute's prohibition of discrimination applies regardless whether the favored entities are interstate or local.
This Court has repeatedly recognized that tax schemes with exemptions may be discriminatory.
To say that such a tax does not "discriminate" is to adopt a definition at odds with the word's natural meaning.
—
+Argument from Precedent
+Argument from Precedent
+Argument from Verbal Classification
Alabama suggests that the distinctions drawn in the statute are not between interstate and local actors.
See, e.g., Davis v. Michigan Department of Treasury, 489 U. S. 803.
And even Department of Revenue of Ore. v. ACF Industries, Inc., 510 U. S. 332, on which the Eleventh Circuit heavily relied in dismissing CSX's suit, made clear that tax exemptions "could be a variant of tax discrimination." Id., at 343.
To charge one group of taxpayers a 2% rate and another group a 4% rate, if the groups are the same in all relevant respects, is to discriminate against the latter.
That discrimination continues if the favored group's rate goes down to 0%, which is all an exemption is.
—
The distinctions are drawn between railroads and all other actors, whether interstate or local.
The statute does not define "discriminates".
+Argument from Ordinary Meaning
The Court again looks to the term's ordinary meaning.
+Argument from Verbal Classification
The ordinary meaning is to fail to treat all persons equally when no reasonable distinction can be found between those favored and those not favored.

Fig. 1 Chain of argumentation in the CSX case

Not all instances of argumentation have a known argumentation scheme that represents the type of argument, and its form, that enables the inference transition from the premises to conclusion. All the argumentation schemes in Fig. 1 are explained later in the chapter (except one, called "argument from ordinary meaning"). Pro-arguments, arguments supporting the conclusion, are indicated by a plus sign in the argument node. Contra-arguments, arguments against a conclusion, are indicated by a minus sign in the argument node. This argument map was drawn using the Carneades Argumentation System, a formal argumentation system that has an argument mapping tool specially designed to represent legal argumentation. How the system works will be briefly explained later in part four of the paper.

Representing the reasoning used in this case presented in the quoted Supreme Court summary above offers a nice way of visually grasping how legal reasoning works by applying rules to cases and by a process of classifying key terms. Once we realize that this kind of reasoning is defeasible and that it is based on terms that are open-textured and very susceptible to argumentation by the opposed side, it no longer seems as simple or straightforward as it might have been when we call it "applying rules to cases." Understanding how rules should be applied to cases takes us to the subject of reasoning from precedent cases.

Reasoning from a precedent depends on an underlying form of reasoning from analogy based on the similarity of the source case to the target case. On this model, rules are continually being modified as they are applied over and over again to a series of cases. By means of examining a number of examples of legal rulings that demonstrate how actual legal method in the common law systems works from examples that result in changes of legal rulings in successive trials, Levi (1949, 104) was led to the conclusion that "legal reasoning has a logic of its own." These examples revealed a contrast between "logic and actual legal method" (104). According to Levi's analyses of how legal decisions were arrived at in the extensive examples he provided, particular entities are classified as falling under general terms that occur in rules that are applied to cases and then modified when the new case is decided on in a different way. According to Levi (1949, 8), this process of legal reasoning has three stages. The first stage is the creation of a legal concept built up from cases. The second stage continues this process of reasoning by example by fixing the concept. The third stage is the breakdown of the concept. The example given by Levi (1949, 14) is the "inherently dangerous rule." In commercial transactions where one party sells something to another party and the second party is injured by using the product, differences in liability turn around the issue of whether a commercial product can be classified as "inherently dangerous" or not. This category was gradually expanded through a series of cases where one product was judged to be similar enough to another product that had already been classified as inherently dangerous so that the second product could also be placed in the same category.

2 Case-Based Reasoning from Analogy

The literature on argument from analogy in fields spanning logic, argumentation studies, computer science, and law, is enormous. Many proposals have been put forward to represent argument from analogy as a form of reasoning or argumentation scheme, and there is no space to try to summarize them here. We can only refer the reader to the multi-disciplinary bibliography of Guarini et al. (2009). However, the use of argument from analogy in case-based reasoning (Ashley 1988) is central. Here, we concentrate on two particular proposals to represent the structure of this argumentation scheme that provides a useful contrast to focus the discussion.

The simplest argumentation scheme for argument from analogy can be represented by this first version from (Walton, Reed and Macagno 2008, 315).

Similarity Premise: Generally, the source case is similar to the target case.

Base Premise: *A* is true (false) in the source case.

Conclusion: *A* is true (false) in the target case.

Let us call this scheme the basic scheme for argument from analogy. The assumption behind the basic scheme for argument from analogy is that there exists a similarity between two cases where *A* holds in the source case and can shift a weight of evidence to make it plausible that *A* holds in both cases. This kind of argument is defeasible, and it can in some instances even be misleading and fallacious, as the traditions of informal fallacies warn us (Hamblin 1970). It is an important kind of argument to study, because so much of our reasoning is based on it (Schauer 2009). But how can similarity be modeled in such a way that we have evidence that is useful to determine whether and how the source case is similar to the target case? An example is helpful.

Barry Bonds hit a valuable home run ball into the stands in the case of Popov v. Hayashi, a trial concerning the issue of which one of two fans could claim ownership rights to the ball. The reasoning in the trial partly turned on some historical precedent cases that concerned the hunting and fishing of wild animals. The trial became a classic example for study on how case-based reasoning can be applied from similar precedent cases to an analogous case at issue (Wyner et al. 2007; Bench-Capon 2009, 2012). The ball went into the upper portion of the webbing of a glove worn by a fan, Alex Popov, but as it entered his glove, he was thrown to the ground by a mob of fans trying to get the ball. While Popov was pinned to ground by the mob, a nearby fan, Patrick Hayashi, not part of the mob that had knocked Popov down, pocketed the loose ball. When the man making a videotape pointed the camera at Hayashi, he held the ball in the air for the others to see. Hayashi was not at fault for the assault on Popov. According to generally accepted rules of baseball, a fan has the right to keep the baseball he has caught in the stands. But such a catch only bestows this right when the fan has the ball in his hand or glove and the ball remains there after its momentum has ceased. If no one catches the baseball, any person in the stands may come to own it by picking it up. According to these accepted rules, it would appear that Hayashi had the right to ownership of the ball. However, Popov also claimed this right and took the case to court.

The contested issue in the trial that took place in the Superior Court of California City and County of San Francisco was about which party should properly be said to have possession of the ball, but the outcome could not be decided by simply applying the legal concept of possession (McCarthy 2002, 5). Some comparable historical precedent cases were presented where there was pursuit of an animal that the pursuer failed to catch because somebody or something intervened, and the issue was whether the pursuer could claim possession of the animal. In the case of Pierson v. Post (3Cai. R. 175; 1805 N.Y. LEXIS 311), Pierson was chasing a fox with hounds when Post captured and killed it, even though he was aware that it was being pursued. The court decided in favor of Post on the basis that mere pursuit did not give Pierson a right to the fox. In the case of Young v. Hitchens (6 Q.B.606 (1844)), Young was a fisherman who spread his net, but when it was almost closed, Hitchens went through the gap with a net and caught the fish. The court decided in favor of Hitchens.

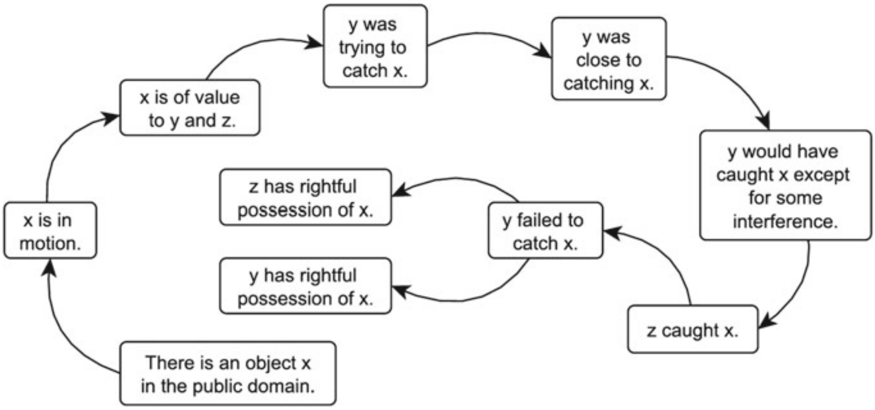

y was trying to catch x.
y was close to catching x.
x is of value to y and z.
y would have caught x except for some interference.
z has rightful possession of x.
y failed to catch x.
x is in motion.
y has rightful possession of x.
z caught x.
There is an object x in the public domain.

Fig. 2 Basis of similarity of the animal cases to the Popov case

In Keeble v. Hickeringill ((1707) 103 ER 1127), *P* owned a pond and made his living by shooting ducks lured onto it with decoys and selling them. *D* used guns to frighten the ducks away from the pond. This case was decided in favor of *P*. In Ghen v. Rich (8 F.159 D. Mass, 1881), Ghen harpooned a whale from his ship, but it was washed ashore found by another man who sold it to Rich. The generally accepted rule of whaling was that the party who finds the whale should report it and can then collect a fee. This case was decided in favor of Ghen.

In the end, after examining many arguments, Judge McCarthy (2002, 9) ruled that any award to one party would be unfair to the other and that each had an equal and undivided interest in the ball. The historical precedent cases are nevertheless interesting in their own right as part of the evidence in the case, because they raise questions about what is meant by "similarity" when a precedent case is seen to be analogous to a case at issue. In many respects, these wild animal cases are not similar to the baseball case at all. As Gray (2002, 1) observed, "a baseball at the end of its arc of descent is not at all like a fox racing across the commons, acting under its own volition, desperately attempting to evade death at the hands of its pursuers." But there is a general similarity underlying the pattern of action in all these cases. In general, they are all about an agent trying to catch something to possess it and about some kind of interference that prevented him from obtaining it, leading to a question of who has the right to possess it (Walton 2010). The similarity can be visualized as an abstract sequence of actions and events that hangs together in a pattern shown in Fig. 2, where *x* is a variable for an object and *y* and *z* are variables for agents. The open arrows represent transitions from one action or event to another in a story scheme (Pennington and Hastie 1993; Bex 2009), while the arrows in the middle represent inferences drawn to a conclusion.

On this view, legal reasoning is a sequence of steps based on similarity between pairs of cases in which a rule applied to one case can also be applied to a second case

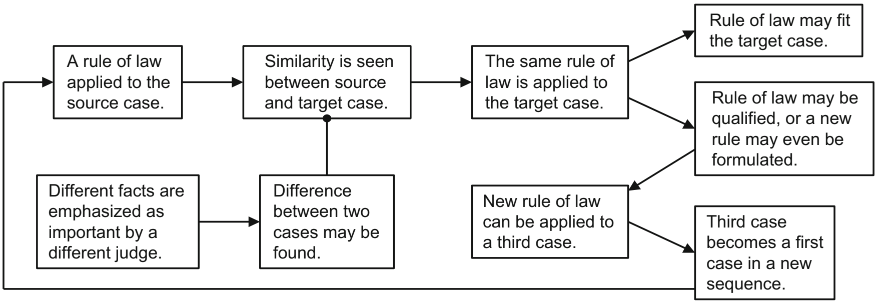

A rule of law applied to the source case.
Similarity is seen between source and target case.
The same rule of law is applied to the target case.
Rule of law may fit the target case.
Rule of law may be qualified, or a new rule may even be formulated.
Different facts are emphasized as important by a different judge.
Difference between two cases may be found.
New rule of law can be applied to a third case.
Third case becomes a first case in a new sequence.

Fig. 3 Sequence of case-based similarity reasoning from analogy

that is taken to be similar to the first one. Essentially, the sequence of reasoning is based on argument from analogy.

But that is not the end of the sequence. There are two possibilities. The argument from analogy may be defeated when a significant difference between the two cases is found (Ashley 2006). Such a difference arises because different facts are emphasized as important by a different judge. The other possibility is that the argument from analogy may be successfully applied to the second case, and this in turn may have two possible outcomes, shown in Fig. 3.

One outcome is that the rule may fit the second case, and in this instance, the same conclusion will be drawn in the second case as was drawn in the first case. The other is that the rule will be changed, by being qualified or otherwise modified, or it may even by being reconfigured by a different rule that replaces the earlier rule. When this outcome occurs, the new rule can be applied to a new case, the third case in the sequence, which starts the whole process over again, or establishes the new rule, and the new third case is applied to a fourth case on the basis of a similarity seen between the third and fourth cases.

Levi (1949, 3) sums up the process by not agreeing that it is a system of applying rules to facts, but rather by showing that it is a complex procedure in which "the rules are discovered in the process of determining similarity or difference." This kind of reasoning is dynamic, because there is a continual feedback process in which the new cases may lead to rulings that are inconsistent with the rules held in the old cases, in which the terms used in the new cases may result in different classifications from those in the old cases.

This process cycles on and on indefinitely as the rules of law are refined, creating precedents by being applied to new cases, as shown in Fig. 4.

The sequence of reasoning shown in Fig. 4 is a process of successive refinement based on analogies between cases. The best model we have of this procedure of reasoning to conclusions is that of case-based reasoning, a technology used to solve a problem posed in a given case by drawing on similar cases retrieved from a database of past cases (Ashley 2006). The solution to the problem posed in the given case is achieved by matching the given case against the retrieved cases by a process of analogy that selects similar cases. The HYPO system produces point–counterpoint

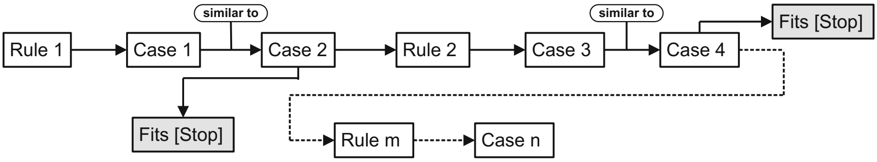


Fig. 4 Successive refinement of cases as a continuing sequence

arguments in trade secrets law, and the CATO system teaches law students how to create case-based arguments. HYPO analyzes a given case by retrieving similar cases from its knowledge base, and makes a judgment concerning which cases are most "on point." CATO has templates for argument moves, such as the one for argument from analogy, and rules that show how to attack a rule (Ashley and Rissland 2003, 41). It can generate an argument for one side while also producing counterarguments that support the other side. Case-based reasoning requires argumentation schemes, especially argument from analogy and argument from precedent (Ashley 2009), and combines these arguments into chains or trees of argumentation modeled on the chain of reasoning shown in Fig. 4.

3 Reasoning and Argument

From a logical, as opposed to a psychological point of view, reasoning may be defined as a series of steps of inference in which some propositions are inferred from others (Walton 1990a, b, 404). Reasoning is sequential and best visualized abstractly as an argument diagram where propositions are contained in text boxes. These text boxes are joined to other text boxes with arrows representing inferences from some propositions to others. Although it is possible to have (as in multi-conclusion logic) premises that lead by inferences to two or more distinct conclusions, it simplifies our view of reasoning if we exclude this possibility. On this more restricted view, reasoning can have several starting points (premises) but the inference drawn from them always leads to a single conclusion. On this view, however, reasoning can be sequential. That is, some premises can lead to a particular conclusion, and this conclusion can then become a premise in the next step of inference to a different conclusion. Such a configuration is called chaining of reasoning in artificial intelligence. If we look at an argument this way, its structure can often be represented as a type of tree referred to as a tree with a single root proposition as the ultimate conclusion. This tree branches outward to a series of connected inferences all leading away from the ultimate conclusion that is the root. An excellent example is the reasoning used in a Supreme Court case represented as an argument diagram using the Araucaria tool for argument visualization (Fig. 1). The reader can see that the diagram of the chain of reasoning as an inverted tree structure with one proposition designated as the ultimate probandum represented as the root of the tree.

To help clarify this definition of reasoning, it is helpful to draw a distinction between epistemic reasoning and practical reasoning. Epistemic reasoning is used to determine whether a proposition is true or false, or whether it is unknown to be either true or false, based on the knowledge of an agent. The simplest form of practical reasoning, called instrumental practical reasoning, has the following form (Walton et al. 2008, 323). The first-person pronoun "I" represents a rational agent that has goals, some (though possibly incomplete) knowledge of its circumstances, the capability of altering those circumstances, and the capability of perceiving the consequences of so acting.

Major Premise: I have a goal *G*.

Minor Premise: Carrying out this action *A* is a means to realize *G*.

Conclusion: Therefore, I ought (practically speaking) to carry out this action *A*.

Practical reasoning can move forward, from a goal to an action, as part of agent-based deliberation, but it can also be used backward by inference to the best explanation (see below) to reconstruct an agent's internal mental states such as motive or intent, based on an agent's known actions and words.

Practical reasoning can be undercut by citing possible negative consequences of the proposed action. Such a form of attack is a species of reasoning in its own right (Walton et al. 2008, 332).

Premise: If *A* is brought about, then bad consequences will occur.

Conclusion: *A* should not be brought about.

By its use of the word "bad," this form of reasoning is seen to be based on values that the agent may be presumed to have and hence it is a species of value-based reasoning (Bench-Capon 2003). However, arguments from positive or negative values can also operate as individual arguments in their own right (Bench-Capon 2003) independent of argument from consequences. The first argumentation scheme represents the argument from positive value.

Major Premise: If value *V* is positive, it supports commitment to goal *G*.

Minor Premise: Value *V* is positive as judged by agent *a*.

Conclusion: *V* is a reason for *a* to commit to goal G.

The negative counterpart is called argument from negative value.

Major Premise: If value *V* is negative, it attacks commitment to goal *G*.

Minor Premise: Value *V* is negative as judged by agent *a*.

Conclusion: *V* is a reason for *a* to retract commitment to goal G.

Argument from values is combined with instrumental practical reasoning to yield the scheme for value-based practical reasoning. This scheme was first formulated in the following form by Atkinson and Bench-Capon (2007, 861).

Circumstances Premise: S_1 is the case in the current circumstances.

Action Premise: Performing *A* in S_1 would bring about S_2.

Goal Premise: G would be realized in S_2

Value Premise: Achieving the goal G would promote the value V.

Conclusion: Action A should be performed.

The following critical questions match the scheme for value-based practical reasoning.

CQ_1: Is V is a legitimate value?

CQ_2: Is G is a worthy goal?

CQ_3: Is action A possible?

CQ_4: Does there exist an action that would bring about S_1 more effectively than A?

CQ_5: Does there exist an action that would realize the goal G more effectively than A?

CQ_6: Does there exist an action that would promote the value V more effectively than A?

CQ_7: Would performing A in S_1 have side effects which demote V or some other value?

An example of value-based practical reasoning (Atkinson et al. 2006, 82) is: I may diet to lose weight, with the goal of not being overweight, to promote the value of health. In the value-based scheme, the notion of a goal is separated into three elements: the state of affairs brought about by the action, the goal (the desired features in that state of affairs), and the value. The value is defined as the reason why those features are desirable. The structure is based on an Action-based Alternating Transition System (Wooldridge and van der Hoek 2005) in which an agent performs an action by moving from a current state of affairs to a new one with many differences that may make the new state of affairs better with respect to some value of the agent.

Practical reasoning is used in a situation of uncertainty and incomplete knowledge where an agent has to make a decision in a given situation that is constantly changing, based on its goals and its knowledge of that situation. This assumption that there is new incoming information to the agent because the situation is constantly changing is called the open world assumption in artificial intelligence. The conclusion is whether a particular course of action should be taken or not. Sometimes doing nothing (inaction) needs to be represented as a possible course of action, because doing nothing at all can often have negative consequences and can affect the agent's goals. Although the open world assumption is typical of practical reasoning of the kind that takes place in realistic decision-making, it is also possible in some instances to invoke what is called the closed world assumption. The closed world assumption rules that no further evidence will count as relevant because the knowledge already available can be regarded as exhaustive of all the relevant evidence for the conclusion.

In a common law criminal trial, the presumption of innocence is taken to shift the burden of proof onto the prosecution to prove its claim of guilt to the standard of beyond a reasonable doubt. All the defense has to do is to cast doubt on the argumentation put forward by the prosecution. This asymmetrical management of the burden of proof in a trial is evocative of the argument from lack of evidence, often called the argument from ignorance in the literature on informal fallacies in logic. However, when used in this context, the following version of the argument is reasonable: It has not been proved that the defendant is guilty; therefore, the defendant should be presumed to be innocent (not guilty). This kind of reasoning is reasonable, except for a complication in cases of Scottish jurisdiction, where the jury can return a verdict of not proven. The question of how burden of proof should be determined in different settings of argumentation is taken up in Sect. 6.

Invoking the closed world assumption has been identified as a form of epistemic reasoning called argument from lack of evidence (Walton, Reed and Macagno 2008, 327).

Major Premise: If A were true, then A would be known to be true.

Minor Premise: It is not the case that A is known to be true.

Conclusion: A is not true.

This form of reasoning depends on how far along the search for evidence has progressed in a given case and may therefore be regarded as defeasible, unless the knowledge base can be closed on the grounds that all the available evidence has been collected and processed. This ground is called the standard of proof (see Sect. 6). The following inference is an example of a typical instance of this kind of reasoning in history.

Minor Premise: There are no known instances of Romans being awarded medals for bravery in battle posthumously.

Major Premise: If there were instances of Romans being awarded medals for bravery in battle posthumously, we would know of them.

Conclusion: Therefore, the Romans did not award medals for bravery in battle posthumously.

To support the minor premise, the following statements might be offered as evidence: We would see evidence on tombstones or in written records of battles. Sometimes reasoning from lack of evidence can be provisionally acceptable, depending on the standard of proof, even though it is based on negative evidence.

If the closed world assumption cannot properly be made, there is the possibility of new information that can affect the outcome of practical reasoning. It is especially important for adequate practical reasoning that a rational agent be open to new incoming information and be flexible in taking this information into account in modifying its goals and actions accordingly.

Another form of inference that is very important in legal reasoning is abductive reasoning, or inference to the best explanation (Pardo and Allen 2008). According to Josephson and Josephson (1994, 14), abductive inference has the following form, showing its structure as inference to the best explanation. H is a hypothesis.

- D is a collection of data.
- H explains D.
- No other hypothesis can explain D as well as H does.
- Therefore, H is probably true.

An example quoted from (Wigmore 1940, 420) shows how he analyzed cases of legal evidence as instances of inference to the best explanation.

> The fact that *a* before a robbery had no money, but after had a large sum, is offered to indicate that he by robbery became possessed of the large sum of money. There are several other possible explanations - the receipt of a legacy, the payment of a debt, the winning of a gambling game, and the like. Nevertheless, the desired explanation rises, among other explanations, to a fair degree of plausibility, and the evidence is received.

The evidence put forward in this example has the form of an inference to the best explanation where the conclusion was arrived at by means of a choice among several competing explanations of given facts.

Another important form of reasoning in law is the drawing of an inference based on perception (Pollock 1995, 41).

> *Premise 1*: Person *P* has a φ image (an image of a perceptible property).
>
> *Premise 2*: To have a φ image (an image of a perceptible property) is a *prima facie* reason to believe that the circumstances exemplify φ.
>
> *Conclusion*: φ is the case.

Pollock (1995, 41) offered the following argument as an example.

> *Minor Premise*: This object looks red to me.
>
> *Major Premise*: When an object looks red, then (normally, but subject to exceptions) it is red.
>
> *Conclusion*: This object is red.

This argument is defeasible, as Pollock pointed out, since even objects that are not red can look red when illuminated by a red light. It is a species of defeasible reasoning that can give a reason to accept its conclusion, provided there is no reason to think that the situation is exceptional.

The most important forms of legal reasoning in law are defeasible (Hart 1949, 1961). Other good examples are instances of drawing inferences from sources, like reasoning from witness testimony and expert opinion. Such forms of reasoning are not well modeled by a deductive logic based on the major premise that what an expert says is always true. Fitting reasoning from expert opinion testimony into a deductive model would render it into a fallacious form of reasoning by making it intolerably rigid. Instead, we need to look to defeasible reasoning.

4 Reasoning by Drawing Inferences from Sources

Drawing an inference from perception clearly represents a kind of reasoning we utilize all the time, not only in law but also in scientific reasoning and in everyday conversational reasoning. As skeptics have often noted, this kind of reasoning is defeasible. Unfortunately, however, in many situations where we have to draw a conclusion on what to do or what to accept as a hypothesis, the reasoner himself has no direct access to data through perception. In such cases, we have to rely on information derived from sources. Source-based reasoning is vitally important in law, because many of the supposed facts happened in the past. One of the most important forms of source-based reasoning and law is inference from witness testimony. The argumentation scheme for this form of reasoning is given in Walton (2008, 60). It has three premises.

Position to Know Premise: Witness *W* is in a position to know whether *A* is true or not.

Truth-Telling Premise: Witness *W* is telling the truth (as *W* knows it).

Statement Premise: Witness *W* states that *A* is true (false).

Conclusion: Therefore (defeasibly) *A* is true (false).

In argument from witness testimony, two key premises are the position to know premise and the truth-telling premise. It is assumed that the witness is basing what she says on her genuine knowledge of some real situation or true set of facts. Moreover, it is assumed that she is telling the truth about those facts, as she saw or knows what she witnessed. These assumptions pose some constraints on witness testimony as a form of argument. It needs to be assumed that the account the witness has presented is internally consistent, and is consistent with known facts of the case that can be verified by independent objective evidence. These matters can be tested by using the following critical questions (Walton 2008, 61) matching the scheme for argument from witness testimony.

Internal Consistency Question: Is what the witness said internally consistent?

Factual Consistency Question: Is what the witness said consistent with the known facts of the case (based on evidence apart from what the witness testified to)?

Consistency with Other Witnesses Question: Is what the witness said consistent with what other witnesses have (independently) testified to?

Trustworthiness Question: Is the witness personally reliable as a source?

Plausibility Question: How plausible is the statement A asserted by the witness?

Bias Question: Is there a bias that can be attributed to the account given by the witness?

If the witness was really in a position to know the facts of a case and is giving an honest and accurate report of what she saw or heard, this should produce an account that is internally inconsistent. Or if it does not appear to be consistent in some points, the apparent inconsistency should be able to be explained or resolved. But consistency can only be tested by probing into the account given by the witness and seeing if her story "stands up" under questioning during examination. This procedure is analyzed in the next section.

Another form of argument that is important in legal reasoning is that of argument from expert opinion. Epistemic reasoning to a conclusion based on expert opinion testimony as an admissible form of evidence requires that the source be qualifiable as an expert. For example, ballistics experts and DNA experts are often used to give expert testimony as evidence in trials, but they must qualify as experts. The most basic version of the form of reasoning from expert opinion is modified from the one in Walton Reed and Macagno (2008, 310) as follows.

Major Premise: Source *E* is an expert in field *F*.

Minor Premise: *E* asserts that proposition *A* is true (false).

Second Minor Premise: *A* is within *F*.

Conclusion: *A* is true (false).

It is not helpful to treat this form of reasoning as deductive, for that would amount to taking an expert as an infallible source of knowledge. Taking that approach makes argumentation susceptible to many serious problems known to be associated with the fallacious misuse of argument from expert opinion. According to the contrasting approach of Walton (1997, 223), an argument from expert opinion should be evaluated by the asking of six basic critical questions.

Expertise Question: How credible is *E* as an expert source?

Field Question: Is *E* an expert in the field *F* that *A* is in?

Opinion Question: What did *E* assert that implies *A*?

Trustworthiness Question: Is *E* personally reliable as a source?

Consistency Question: Is *A* consistent with what other experts assert?

Backup Evidence Question: Is *E*'s assertion based on evidence?

According to Walton (1997), if a respondent asks any one of the six critical questions, the original argument defaults until the question has been answered adequately.

A problem with using critical questions to evaluate cases where expert opinion is used as a source of evidence is that we can no longer use an argument diagram to summarize, analyze or evaluate the basic evidence in a case and display its structure as a sequence of reasoning. The reason is that everything that appears in the text box on a standard argument diagram needs to be a statement, a proposition that is either true or false. It is harder to analyze the structure of questions, even though they are certainly very important as devices in both everyday and legal argumentation, for example in examining a witness. Using critical questions definitely takes us outside the realm of reasoning to the realm of argument, where claims are made and subjected to doubt by the asking of critical questions by an opponent.

We can see the problem more explicitly if we ask what happens when a critic asks a critical question. If the critic asks the opinion question, in other words if he asks the arguer who has appealed to argument from expert opinion to quote the specific statement that the expert made that supposedly implies *A*, then the arguer certainly has to respond to this reasonable question by presenting the critic with a specific proposition, for example by quoting exactly what the experts said. If the arguer fails to carry out this reasonable kind of request, this argument from expert opinion will surely fail to be persuasive. However, suppose the critic asks the trustworthiness question: is *E* personally reliable as a source? It could be perfectly reasonable for the arguer to shift the burden of disproof back to the questioner by replying, "of course she is personally reliable, for after all she is an expert, and if you wish to make any allegation to the effect that she is not personally reliable, you had better prove that." This kind of reply would certainly be sufficient as an adequate answer to the question. In other words, there are two kinds of critical questions. When the first kind of critical question is asked, any failure to answer inappropriately will defeat the argument. When the second kind of critical question is asked, the burden shifts from the arguer to the questioner to support the critical question with further argument. Because of these crucial differences between the critical questions, it seems impossible to represent them in any straightforward way as statements that are additional assumptions of the argument.

Fortunately, however, there is a way that we can represent critical questions on an argument diagram by treating them as additional premises that need to be added to the given premises in the argumentation scheme (Walton and Gordon 2005). Artificial intelligence has found a way to do this by using the Carneades system (Gordon 2010). Carneades is a mathematical and computational model that defines mathematical properties of arguments that are used to identify, analyze, and visualize real arguments. By applying argumentation schemes, Carneades analyzes and evaluates the acceptability of arguments, based on proof standards, for example preponderance of the evidence. Carneades takes the approach that the way critical questions are modeled depends on the individual argumentation scheme, by distinguishing three kinds of premises. Ordinary premises are just the regular premises of an argumentation scheme that are explicitly given in the scheme itself. But there are two additional kinds of premises not stated in the scheme. Assumptions are to be acceptable unless called into question. Exceptions are modeled as premises that are not assumed to be acceptable and that can defeat an argument as it proceeds. Ordinary premises of an argument, like assumptions, are assumed to be acceptable, but they must be supported by further arguments in order to be judged acceptable.

In Fig. 5 the conclusion of the argument from expert opinion is represented by the text box at the far left stating that *A* is true. The argument from expert opinion is represented in the node with a plus sign in front of its name, indicating it is a pro-argument supporting the conclusion that *A* is true. In the list of premises on the right of the node, the first four are the ordinary premises of the scheme for argument from expert opinion. Hence, they are labeled as being in assumptions, statements that are assumed to hold, but are subject to critical questioning. Since there are no arguments against them, or critical questions directed to them, they are shown as acceptable in Fig. 5 by darkening the text boxes in which they appear. The next two premises are also classified as assumptions. Asking a critical question challenging either one of these for assumptions will shift the burden of proof onto the proponent of the argument from expert opinion and temporarily defeat it until the proponent replies to the question appropriately. Since they have not been questioned or attacked either, they are also shown as accepted. The two critical questions at the bottom, the trustworthy question and the “consistency with other what experts say” question, are represented as exceptions. These two critical questions are shown on the diagram as undercutters, because they are displayed as contra-arguments, indicated by the minus signs in their nodes. An undercutter is an argument that attacks an argument node rather than attacking a premise or conclusion (a proposition).

The “consistency with other what experts say” question is an undercutter with a premise displayed in white text box indicating that it is not accepted. Notice however that the other one of these premises, the trustworthiness premise, has an argument from bias supporting the statement that the expert is not trustworthy. Since this undercutter has been backed up by evidence to support it, in the form of the argument from bias, it successfully defeats the argument from expert opinion. Hence, despite the fact that the other four premises above it are assumed to hold, the argument

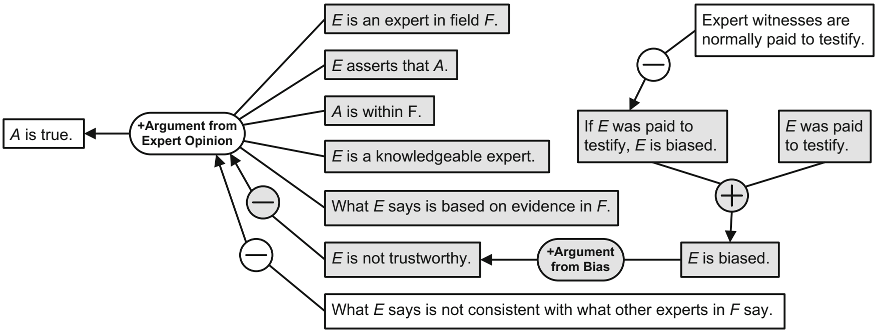

A is true.
+Argument from Expert Opinion
E is an expert in field F.
E asserts that A.
A is within F.
E is a knowledgeable expert.
What E says is based on evidence in F.
E is not trustworthy.
What E says is not consistent with what other experts in F say.
+Argument from Bias
E is biased.
Expert witnesses are normally paid to testify.
If E was paid to testify, E is biased.
E was paid to testify.

Fig. 5 Argument from expert opinion with undercutter in a Carneades Map

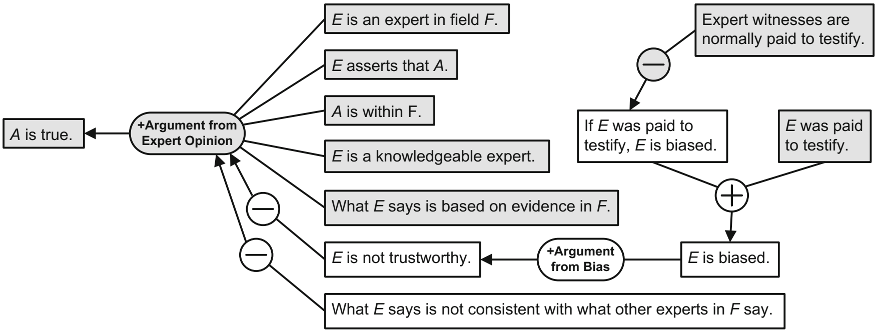

A is true.
+Argument from Expert Opinion
E is an expert in field F.
E asserts that A.
A is within F.
E is a knowledgeable expert.
What E says is based on evidence in F.
E is not trustworthy.
What E says is not consistent with what other experts in F say.
+Argument from Bias
E is biased.
Expert witnesses are normally paid to testify.
If E was paid to testify, E is biased.
E was paid to testify.

Fig. 6 Counterattack to the undercutter in an argument from expert opinion

from expert opinion fails. Notice however that one of the premises of the argument supporting the claim that *E* is biased is attacked by a counterargument stating that expert witnesses are normally paid to testify. In other words, the argument is that since expert witnesses are normally paid to testify, if a particular expert is paid to testify, that should not be taken as evidence that he is biased. However, the way the argumentation is represented in Fig. 5 to proposition that expert witnesses are normally paid to testify is not accepted, nor is it supported by a further argument.

But what would happen if the proposition that expert witnesses are normally paid to testify were to be accepted? This new evidential situation is shown in Fig. 6.

Notice at the top right of Fig. 6 a contra-argument based on the premise that expert witnesses are normally paid to testify has been used to attack one of the premises of the argument from bias. On the basis of this attack, one of the premises supporting the conclusion that *E* is biased is no longer accepted. Hence, the argument from bias is defeated by this premise attack, and the topmost undercutter now fails to defeat the argument from expert opinion. Since there is no supporting evidence behind the other undercutter displayed in the text box at the bottom of Fig. 6, the argument from expert opinion is restored. In this situation, Carneades will automatically display the ultimate conclusion that is true as "accepted," once the user has input the information that all the other premises of the argument from expert opinion are assumed to hold, as shown in Fig. 6. So we can see that Carneades is a dynamic system that takes into account new arguments brought forward from its knowledge base, so that in some instances an argument that was formerly used to successfully refute an argument can be defeated by a new argument.

We have shown how Carneades incorporates defeasible logic and builds on it to provide a computational tool that not only enables us to do argument mapping, but to represent the critical questions matching defeasible argumentation scheme on an argument map, and to track new relevant evidence coming in from its knowledge base.

5 Defeasible Logic

Defeasible logic is a logical system (Nute 1994) that models reasoning used to derive provisional conclusions from partial and sometimes conflicting information. Using this kind of reasoning, a conclusion can be tentatively accepted, subject to new evidence that may come to be known at some point in the future of an investigation. This new evidence may require the retraction of the conclusion that was formerly accepted, based on the evidence that was available at that earlier point. The use of this kind of defeasible reasoning is highly appropriate during an investigation or sequence of argumentation where the inflow of new evidence may be closed off, even though it may later be reopened, for example in an appeal. Once the investigation has been closed off, the conclusion of the reasoning can be accepted as final, for the purposes of the investigation. But before that point, where there is argumentation on both sides, and all the arguments are not in, the reasoning needs to be regarded as defeasible.

The basic units of defeasible logic are called facts and rules. Facts are statements that are accepted as true within the confines of a discussion. Statements, also called propositions, are denoted by letters, A, B, C, …, using subscripts if necessary. There are two kinds of rules in defeasible logic, called strict rules and defeasible rules. Strict rules are absolutely universal in the sense that they do not admit of exceptions. An example of a strict rule would be the universal generalization “All penguins are birds.” In defeasible logic, a strict rule has the form of a material conditional with a conjunctive antecedent of the following form: A_1, A_2, A_n, …, $\rightarrow B$. In this kind of conditional, it is not possible for all the A_i to be true and the B false. Defeasible rules are rules that are subject to exceptions, and that may fail if an exception is shown to exist in a given case. An example of a defeasible rule would be the statement “Birds fly,” meaning that birds generally fly or that birds normally fly, but not implying that all birds fly without any exception being allowed. A defeasible rule has the form A_1, A_2, A_n …, => B, where each of the A_i is called a prerequisite. The set of all the A_i taken together is called the antecedent, and the statement B is called the consequent. For example, suppose that Tweety is a bird, but we also know that he is a penguin, and that we know that penguins cannot fly. In light of our knowledge of this exception, the conclusion that Tweety flies cannot be inferred and has to be retracted. The rule still holds, but the inference itself fails to support the conclusion any longer. Defeasible logic is the best way to represent the structure of reasoning of argumentation schemes of the kind most commonly used in law, for as shown above, this type of reasoning is defeasible.

It is a problem that the forms of reasoning that we are familiar with from deductive logic and from the kind of inductive logical reasoning used in the Bayesian rules do not appear to fit argumentation schemes of the defeasible kind illustrated above. However, Verheij (2001, 232) showed that these defeasible argumentation schemes fit a form of argument he called *modus non excipiens*, which has the following form.

As a rule, if *P* then *Q*

P

It is not the case that there is an exception to the rule that if *P* then *Q*

Therefore *Q*

Verheij showed how this form of argument can be used for evaluating defeasible inferences like the Tweety argument: If Tweety is a bird, Tweety flies; Tweety is a bird; therefore, Tweety flies. This form of argument was called defeasible *modus ponens* (DMP) by Walton (2002). A version of an example from (Copi and Cohen 1998, 363) can be used to illustrate DMP: If he has a very good defense lawyer, he will be acquitted; Bob has a very good defense lawyer; therefore, he will (likely) be acquitted. This argument is clearly defeasible, for even though Bob has a good lawyer, he may not be acquitted.

Using a concept from defeasible logic called defeasible implication, or the defeasible conditional, represented by the symbol =>, we can represent DMP is having the following form.

Major Premise: *A* => *B*

Minor Premise: *A*

Conclusion: *B*

The first premise is the conditional, "If *A* is true then generally, but subject to exceptions, *B* is true." In some instances, the argumentation schemes above, for example the argument from negative evidence, explicitly have the DMP form. In other instances, the scheme for argument from expert opinion for example, the scheme can be put into the DMP form by adding an implicit premise as an additional assumption.

To see how this works using an example, let us consider the following expanded version of the argument from expert opinion scheme.

Major Premise: Source *E* is an expert in subject domain *S* containing proposition *A*.

Minor Premise: *E* asserts that proposition *A* (in domain *S*) is true (false).

Conditional Premise: If source *E* is an expert in a subject domain *S* containing proposition *A*, and *E* asserts that proposition *A* is true (false), then *A* may plausibly be taken to be true (false).

Conclusion: *A* may plausibly be taken to be true (false).

If you look at this version of the scheme, you can see that the argument from expert opinion has a *defeasible modus ponens* structure as an inference.

> *Major Premise*: (E is an expert and E says that A) => A
>
> *Minor Premise*: E is an expert and E says that A
>
> *Conclusion*: A

This form of argument is not exactly the same as DMP because the conditional in the major premise has a conjunctive antecedent. The scheme has this form: (A & B) => C, A & B, therefore C. Nevertheless, it is a substitution instance of the DMP form. It is fair to say that in its general outline it has the structure of the DMP form of inference.

The analysis so far, however, does not take into account the critical questions for the argument from expert opinion. There was a suggestion made by (Reed and Walton 2003, 202) that the conditional premise could be expanded to take the critical questions into account in a still more fully expanded version of the scheme. This proposal is now easily carried out using the Carneades system of treating the critical questions as additional assumptions or exceptions in the scheme. This form of the argument can now be seen as fitting DMP.

6 Reasoning, Argument, and Proof

Reasoning is included in argument, but argument is a wider notion. Argument is reasoning used to try to resolve some central issue that is unsettled (Prakken and Sartor 2006). As stated above, in reasoning there is always a set of propositions called start points (premises) and a single endpoint (conclusion). In an argument, the conclusion is always the claim made by one party that is doubted or is open to doubt by the other party. The other party may be a single person or an audience composed of more than one person, for example a jury. In argument, the conclusion is always unsettled, or open to doubt. Indeed, that is the whole point of using an argument. If there is no doubt about a proposition, and everybody accepts it as true, there is no reason for arguing either for or against it. Thus, the speech act of putting forward an argument is different from the speech act of putting forward an explanation. An explanation is only appropriate if the proposition to be explained is taken as an accepted fact by all parties. An argument is only appropriate if the proposition at issue is not taken as an accepted fact by all parties, so that there is some doubt about whether it is true or not. The definition of the notion of an argument put forward here is dialectical, implying that an argument is only appropriate in a setting where two or more parties take part in trying to use reasoning to examine and evaluate the evidence on both sides of a disputed issue. The term “argumentation” is appropriately used here, because an argument, defined in this way, always needs to be evaluated within a procedural setting. There needs to be an opening stage, in which the ultimate issue needs to be specified, an argumentation stage where the arguments on both sides are

Table 1 Seven basic types of dialogue

Dialogue type	Initial situation	Participant's goal	Communal goal
Persuasion	Conflict of opinions	Persuade other party	Resolve issue
Inquiry	Need to have proof	Verify evidence	Prove hypothesis
Discovery	Need an explanation	Find a hypothesis	Support hypothesis
Negotiation	Conflict of interests	Get what you want	Settle issue
Information	Need information	Acquire information	Exchange information
Deliberation	Practical choice	Fit goals and actions	Decide what to do
Eristic	Personal conflict	Hit out at opponent	Reveal deep conflict

put forward so that they can be critically questioned by the opposing side, and the closing stage where it can be determined which side had the stronger argument.

One of the best examples of argumentation that can be given is a common law trial in which one side has made a claim that the other side disputes. The claim made by the first part is the ultimate *probandum*, the ultimate conclusion to be proved, while the other party has the job of casting doubt on the first party's argument. In this setting, there are three parties involved, the two primary parties and a third-party trier, who may be a judge or jury. Procedural rules determine what sort of evidence is admissible, and the two primary parties are supposed to use this evidence to try to prove their own contentions and cast doubt on the contentions of the other side. The trier makes a decision at the closing stage by examining the arguments on both sides and determining whether one side or the other successfully carried out its central task. The overall framework of a common law trial is that of a persuasion dialogue, because the party who has made the initial claim that defines the issue to be settled has a burden of persuasion. It has the burden to prove its claim by sufficient evidence or it loses by default and the other party wins. This burden is often called the presumption of innocence in criminal trials.

Although persuasion dialogue is a very important type of dialogue, so much so that argumentation is often mainly associated with it, there are other types of dialogue that are important from a legal reasoning point of view. Six types were distinguished in Walton and Krabbe (1995, 66), and the seventh type was also later recognized. Each type has a communal goal that needs to be distinguished from the individual goals of the participants. The defining characteristics of each type are shown in Table 1.

Each type of dialogue has an opening stage, an argumentation stage, and a closing stage. A dialogue is formally defined as an ordered 3-tuple $\{O, A, C\}$ where O is the opening stage, A is the argumentation stage, and C is the closing stage (Gordon and Walton 2009, 5). Procedural rules define what types of moves are allowed by the parties during the argumentation stage (Walton and Krabbe 1995). At the opening stage, the participants agree to take part in some type of dialogue that has an identified collective goal. The initial situation is framed at the opening stage, and the dialogue moves through the argumentation stage toward the closing stage. This way of abstractly modeling argumentation in a dialogue setting is normative, meaning that it prescribes how the dialogue should ideally go if the participants aim to use reasoning to resolve the issue that was framed at the opening stage. In any real case, the actual order of events may be quite different.

In these dialogue systems, the simplest cases are those with only two participants. To test out the systems, Lodder (1999, 99) considered a short sample dialogue of legal argumentation arising out of an actual case. In this case, a gang member named Tyrell was searched during a football game, and he was found to be in possession of illegal drugs. The issue was whether this evidence had been obtained illegally. The two participants in the dialogue are called Bert and Ernie, but any names could be chosen, like Proponent and Respondent, or Black and White.

Bert: It was not allowed to search Tyrell.

Ernie: Why do you think so?

Bert: Only if someone is a suspect may he be searched, and Tyrell was not a suspect.

Ernie: I agree, but Tyrell was on probation, and had to allow a search at any time.

Bert: You are right, a search was allowed.

What is especially interesting about the argumentation in this dialogue is that it is based on defeasible reasoning, based on generally accepted legal rules that are subject to exceptions. Two of these general rules can be identified as follows.

Search Rule: Someone may be searched only if he is a suspect.

Probation Rule: Someone who is on probation can be searched at any time.

The reason behind the second rule is that one of the conditions of probation is to allow a search at any time. Turning to the dialogue, the line of argumentation can be followed along, based on how these defeasible rules are applied. First, Bert made the claim that it was not allowed to search Tyrell. This assertion can be taken to mean that Bert has advocated a viewpoint, to use the language of the critical discussion. In other words, he has made a statement and committed himself to the truth or acceptability of it. In the critical discussion model, it follows that Bert is obliged to defend his claim, if challenged, or give it up, if the other party can give a convincing argument against it. Ernie then questions Bert's claim, at the next move in the dialogue. Bert then fulfills his obligation by presenting an argument. His argument cites the claimed fact that Tyrell was not a suspect, along with the search rule as an additional premise. The two premises function together as an argument to support argument against Ernie's prior argument. Also, it should be noted that the search rule is a defeasible rule that is subject to exceptions. Thus Bert's argument, although it appears to be reasonable as matters stand, could be defeated by new evidence that might come into the dialogue. And in fact, that is what happens in the dialogue, when Ernie cites the presumed fact that Tyrell was on probation, and uses it along with the probation rule to put forward a new defeasible argument that defeats Bert's prior argument.

Information-seeking dialogue is, at first, hard to grasp as a framework of legal argumentation. The most familiar framework is that of persuasion dialogue, which is highly adversarial, in that its basic structure is based on opposed arguments. Prakken (2006) has investigated a formal framework for the study of dialogue games for argumentation in which each dialogue move either attacks or surrenders to some earlier move of the other participant. This framework assumes an explicit reply structure on dialogues and imposes strict protocols on turn taking and relevance. However, Prakken has also explored less rigorous and more permissive formal dialogues in which these strong conditions are relaxed. In these more permissive types of dialogue, alternative replies to the same turn are allowed, and some dialogues do not have to adhere to the rule that each move must have a bearing on the outcome of the dialogue. Prakken (2006, 28) presented the following example dialogue.

Witness: Suspect was at home with me that day.

Prosecutor: Are you a student?

Witness: Yes.

Prosecutor: Was that day during summer holiday?

Witness: Yes.

Prosecutor: Aren't all students away during summer holiday?

Prakken cites this example as a case where the cross-examination of a witness has the goal of revealing an inconsistency in the testimony. He offered the example as a typical case in which a line of questioning does not indicate from the start where it is aiming. But clearly it is as species of information-seeking dialogue. It can also be identified as falling under the category of examination dialogue, a subspecies of information-seeking dialogue. It should probably be seen as taking place at the beginning stage of the line of questioning, in which the prosecutor is asking questions of a kind that attempt to pin down the commitments of the witness in such a way that they can be subsequently be critically examined. One such strategy of cross-examination that the prosecutor may have in mind is that of later trapping the witness into committing himself to an inconsistency, or to some proposition that would appear to be implausible to the judge or jury in the trial. On this model even though the dialogue is a fragmentary one, and very little context is given, it can be straightforwardly categorized as an instance of a kind of examination dialogue that falls under the information-seeking category. As seen from a viewpoint of persuasion dialogue, it may appear to be aimless, but as seen from a viewpoint of examination dialogue, it can be analyzed as part of a dialogue that has a goal.

So far there has been a lot of discussion about both reasoning and argument, but now we also need to consider the concept of proof. A proof may be defined as an argument in which the premises furnish sufficient evidence to reasonably accept the conclusion. A burden of proof is a requirement set on one side or the other to meet a standard of proof in order for the argument of that side to be judged successful as a proof (Gordon et al. 2007). The problem in legal reasoning, and indeed in all kinds of reasoning generally, is that it is rarely if ever possible to be able to prove a conclusion beyond all doubt. Hence, the question of when a burden of proof is met by a sequence of argumentation in a given case depends on the proof standard that is required for a successful argument in that case. What proof standard is required depends on the type of dialogue. In inquiry dialogue, for example an investigation into the cause of an air disaster has a very high standard of proof. In another dialogue setting, the standard may not need to be set as high. Proof standards need to be set at the opening stage of a dialogue.

According to the scintilla of evidence standard, an argument is taken to be a proof even if there is only a small amount of evidence supporting it (Gordon and Walton 2009). The preponderance of evidence proof standard is met by an argument that is stronger than the opposed argument in the case, even if it is only slightly stronger. The clear and convincing evidence standard is higher than that of the preponderance standard, but not as high as the highest standard, called proof beyond reasonable doubt. The beyond reasonable doubt standard is the strongest one, and it is applicable in criminal cases. The beyond reasonable doubt standard is often equated with the presumption of innocence in criminal cases (Tillers and Gottfried 2006).

Burden of proof is a slippery and vague notion in law, and so far it has resisted any precise definition that has been unanimously accepted in law (Prakken and Sartor 2009). However, recent work in artificial intelligence has constructed logical models based on defeasible logic that are helpful in clarifying how the notion of burden of proof works in legal argumentation by distinguishing different kinds of burden of proof that can appear at different stages in a dialogue (Prakken and Sartor 2003). At the opening stage, when a person makes a claim at the first point in the sequence described above, he has a right to a legal remedy if he can bring forward facts that are sufficient to prove that he is entitled to some remedy. This is called the burden of claiming. The second type of onus is the burden of questioning. If one party makes an allegation by claiming that some proposition is true during the process of the argumentation, and the other party fails to present a counterargument, or even to deny the claim, then that claim is taken to be implicitly conceded. This type of burden of proof is called the burden of questioning because it puts an obligation on the other party to question or contest a claim made by the other side, by asking the other side to produce arguments to support its claim. The third burden is called the burden of production or the burden of producing evidence. It is the burden to respond to a questioning of one's claim by producing evidence to support it (Prakken and Sartor 2009).

The fourth type of burden of proof is called the burden of persuasion in law. It is set by law at the opening stage of the trial and determines which side has won or lost the case at the end of the trial once all the arguments have been examined. The burden of persuasion works differently in a civil proceeding than in a criminal one. In a civil proceeding, the plaintiff has the burden of persuasion for all the claims he has made as factual, while the defendant has the burden for any exceptions that he has pleaded. In criminal law, the prosecution has the burden of persuasion for all facts of the case. These include not only the elements of the alleged crime, but also the burden of disproving defenses. For example, let us say that in a murder case in a particular jurisdiction, the prosecution has to prove that there was a killing and that it was done with malice aforethought. If the defendant pleads self-defense, the prosecution has to prove that there was no self-defense (Prakken and Sartor 2009).

The fifth type of burden is called the tactical burden of proof. It applies during the argumentation stage of the trial, when a lawyer pleading a case has to make strategic decisions on whether it is better to present an argument or not. This is a hypothetical assessment made only by the advocates on the two sides, and it can shift back and forth during a sequence of argumentation.

7 Conclusions

Some important general theoretical questions for argumentation theory underlie this chapter. Although it is generally appropriate to call the inferential structures studied in the chapter argumentation schemes, in many instances they could equally appropriately be called reasoning schemes. Discussion of this question in the chapter has led to another high-level question for argumentation theory: What is the difference between reasoning and argument? It appears that at least in some instances, argumentation schemes apply to cases as forms of inference used to generate conclusions by inference from a set of assumptions. In such cases, the conclusion may be drawn by the inference of a solitary agent and may not be disputed or doubted by a second party. In such cases, it may seem appropriate to better describe the argumentation scheme by calling it a reasoning scheme. So why are they called argumentation schemes at all, if the process of generating conclusions by inference would seem to be appropriately described by using the term “reasoning”? The answer given in this chapter is that they are forms of reasoning, and as such they can be used to identify instances of known kinds of reasoning, a useful task, but there is also an important reason why it is more generally appropriate to call them argumentation schemes. The reason is that the arguments fitting them are evaluated using sets of critical questions matching a particular scheme. This is a process in which one party, the proponent, puts forward an argument and another party, the respondent, asks critical questions in a dialogue format. In the simplest case, where two parties are involved, one of them is casting doubt on the other’s argument. Here, the use of the term “argument” is highly appropriate, because for there to be an argument there has to be a claim that is unsettled.

References

Ashley, K. 1988. Arguing by analogy in law: A case-based model. In *Analogical reasoning*, ed. D.H. Helman, 205–224. Dordrecht: Kluwer.

Ashley, K. 2006. Case-based reasoning. In *Information technology and lawyers,* ed. A.R. Lodder and A. Oskamp, 23–60. Berlin: Springer.

Ashley, K. 2009. Ontological requirements for analogical, teleological and hypothetical reasoning. In *Proceeding of ICAIL 2009: 12th international conference on artificial intelligence and law*, 1–10. New York, N.Y.: Association for Computing Machinery.

Ashley, K., and E. Rissland. 2003. Law, learning and representation. *Artificial Intelligence* 150: 17–58.

Atkinson, K., and T. Bench-Capon. 2007. Practical reasoning as presumptive argumentation using action based alternating transition systems. *Artificial Intelligence* 171: 855–874.

Atkinson, K., T. Bench-Capon, and P. McBurney. 2005. Arguing about cases as practical reasoning. In *Proceedings of the 10th international conference on artificial intelligence and law,* ed. G. Sartor, 35–44. New York, N.Y.: ACM Press.

Atkinson, K., T. Bench-Capon, and P. McBurney. 2006. Computational representation of practical argument. *Synthese* 152: 157–206.

Bench-Capon, T. 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation* 13: 429–448.

Bench-Capon, T. 2009. Dimension based representation of Popov v Hayashi. In *Modelling legal cases,* ed. K. Atkinson, 41–52. Barcelona: Huygens Editorial.

Bench-Capon, T. 2012. Representing Popov vs. Hayashi with Dimensions and Factors. *Artificial Intelligence and Law* 20: 15–35.

Bex, F. 2009. Analysing stories using schemes. In *Legal evidence and proof: statistics, stories, logic*, ed. H. Kaptein, H. Prakken and B. Verheij, 93–116. Farnham: Ashgate.

Brewer, S. 1996. Exemplary reasoning: Semantics, pragmatics and the rational force of legal argument by analogy. *Harvard Law Review* 923–1038.

Copi, I.M., and C. Cohen. 1998. *Introduction to logic*, 10th ed. Upper Saddle River: Prentice Hall.

Gordon, T.F. 2010. The Carneades argumentation support system. In *Dialectics, dialogue and argumentation*, ed. C. Reed, and C.W. Tindale. London: College Publications.

Gordon, T.F., and D. Walton. 2009. Proof burdens and standards. In *Argumentation and Artificial Intelligence*, ed. I. Rahwan, and G. Simari, 239–260. Berlin: Springer.

Gordon, T.F., H. Prakken, and D. Walton. 2007. The Carneades model of argument and burden of proof. *Artificial Intelligence* 171: 875–896.

Gray, B.E. 2002. Reported and recommendations on the law of capture and first possession: Popov v. Hayashi. In *Superior of the State of California for the City and County of San Francisco.* Case no. 400545, November 6, 2002. http://web.mac.com/graybe/Site/Writings_files/Hayashi%20Brief.pdf. Accessed 24 May 2009.

Guarini, M., A. Butchart, P. Simard Smith, and A. Moldovan. 2009. Resources for research on analogy: A multi-disciplinary guide. *Informal Logic* 29: 84–197.

Hamblin, C. 1970. *Fallacies*. London: Methuen.

Hart, H.L.A. 1949, 1951. The ascription of responsibility and rights. *Proceedings of the Aristotelian Society* 49: 171–194. Reprinted in *Logic and language,* ed. A. Flew. 145–166. Oxford: Blackwell, 1951.

Hart, H.L.A. 1961. *The concept of law*. Oxford: Oxford University Press.

Josephson, J.R., and S.G. Josephson. 1994. *Abductive inference: Computation, philosophy, technology*. New York, N.Y.: Cambridge University Press.

Levi, E.H. 1949. *An introduction to legal reasoning*. Chicago, Ill: University of Chicago Press.

Lodder, A.R. 1999. *Dialaw: On legal justification and dialogical models of argumentation*. Dordrecht: Kluwer.

Macagno, F., and D. Walton. 2009. Argument from analogy in law, the classical tradition, and recent theories. *Philosophy & Rhetoric* 42: 154–182.

McCarthy, K.M. 2002. Statement of decision. Case of Popov v. Hayashi #4005545. *Superior Court of California.* www.findlaw. Accessed 12 Dec 2002.

Nute, D. 1994. Defeasible logic. In *Handbook of logic in artificial intelligence and logic programming*, vol. 3. *Nonmonotonic reasoning and uncertain reasoning*, ed. D.M. Gabbay, C.J. Hogger, and J.A. Robinson, 353–395. Oxford: Oxford University Press.

Pardo, M.S., and R.J. Allen. 2008. Juridical proof and the best explanation. *Law and Philosophy* 27: 223–268.

Pennington, N., and R. Hastie. 1993. The story model for juror decision making. In *Inside the juror: The psychology of juror decision making*, ed. R. Hastie, 192–221. Cambridge: Cambridge University Press.

Pollock, J. 1995. *Cognitive Carpentry*. Cambridge, Mass.: MIT Press.

Prakken, H. 2005. AI & law, logic and argument schemes. *Argumentation* 19: 303–320.

Prakken, H. 2006. Models of persuasion dialogue. http://www.cs.uu.nl/groups/IS/archive/henry/argbookhp.pdf. (Originally published as Formal systems for persuasion dialogue. *The Knowledge Engineering Review* 21: 163–188, 2006.).

Prakken, H., and G. Sartor. 2006. A dialectical model of assessing conflicting arguments in legal reasoning. *Artificial Intelligence and Law* 4: 331–368.

Prakken, H., and G. Sartor. 2009. A logical analysis of burdens of proof. In *Legal evidence and proof: Statistics, stories, logic*, ed. H. Kaptein, H. Prakken, and B. Verheij, 223–253. Farnham: Ashgate Publishing.

Reed, C., and D. Walton. 2003. Diagramming, argumentation schemes and critical questions. In *Anyone who has a view: Theoretical contributions to the study of argumentation*, ed. F.H. van Eemeren, et al., 195–211. Dordrecht: Kluwer.

Sartor, G. 2005. *Legal reasoning: A cognitive approach to the law*. Berlin: Springer.

Schauer, F. 1987. Precedent. *Stanford Law Review* 39: 571–605.

Schauer, F. 2009. *Thinking like a lawyer*. Cambridge, Mass.: Harvard University Press.

Tillers, P. 1989. Webs of things in the mind: A new science of evidence. *Michigan Law Review* 87: 1225–1258.

Tillers, P., and J. Gottfried. 2006. Case comment—United States v. Copeland, 369 F. Supp. 2d 275 (E.D.N.Y. 2005): A collateral attack on the legal maxim that proof beyond a reasonable doubt is unquantifiable? *Law, Probability and Risk* 5: 135–157.

Twining, W., and D. Miers. 2010. How to do things with rules. Cambridge: Cambridge University Press.

Verheij, B. 2001. Legal decision making as dialectical theory construction with argumentation schemes. In *The 8th international conference on artificial intelligence and law: Proceedings of the conference,* 225−236. New York Association for Computing Machinery. http://www.ai.rug.nl/~verheij/publications.htm.

Walton, D. 1990a. What is reasoning? What is an argument? *Journal of Philosophy* 87: 399–419.

Walton, D. 1990b. *Practical reasoning*. Savage, Md.: Roman and Littlefield.

Walton, D. 1997. *Appeal to expert opinion*. University Park, Penn.: Penn State Press.

Walton, D. 2002. Are some modus ponens arguments deductively invalid? *Informal Logic* 22: 19–46.

Walton, D. 2008. *Witness testimony evidence: Argumentation, artificial intelligence and law*. Cambridge: Cambridge University Press.

Walton, D. 2010. Similarity, precedent and argument from analogy. *Artificial Intelligence and Law* 18: 217–246.

Walton, D., and E.C.W. Krabbe. 1995. *Commitment in dialogue*. Albany, Texas: SUNY Press.

Walton, D., and T.F. Gordon. 2005. Critical questions in computational models of legal argument. In *IAAIL workshop series, international workshop on argumentation in artificial intelligence and law*, ed. P.E. Dunne, and T.J.M. Bench-Capon, 103–111. Nijmegen: Wolf Legal Publishers.

Walton, D., C. Reed, and F. Macagno. 2008. *Argumentation schemes*. Cambridge: Cambridge University Press.

Weinreb, L.L. 2005. *Legal reason: The use of analogy in legal argument*. Cambridge: Cambridge University Press.

Wigmore, J.H. 1931. *The principles of judicial proof*, 2nd ed. Boston, Mass.: Little, Brown and Company.

Wigmore, J.H. 1940. *Evidence in trials at common law*. Boston, Mass.: Little, Brown & Co.

Wooldridge, M., and W. van der Hoek. 2005. On obligations and normative ability: Towards a logical analysis of the social contract. *Journal of Applied Logic* 3: 396–420.

Wyner, A., and T. Bench-Capon. 2007. Argument schemes for legal case-based reasoning. In *Legal knowledge and information systems (JURIX 2007)*, ed. A. Lodder, and L. Mommers, 139–149. Amsterdam: IOS Press.

A. Wyner, T.J.M. Bench-Capon, and K. Atkinson. 2007. Arguments, values and baseballs: Representation of Popov v. Hayashi. In *Legal knowledge and information systems (JURIX 2007)*, eds. A. Lodder and L. Mommers, 151−160. Amsterdam: IOS Press.

Norms in Action: A Logical Perspective

Emiliano Lorini

1 Introduction

A theory of action is fundamental for legal theory, as the law is meant to direct behaviour: it influences the behaviour of agents who can understand the law's prescriptions and act accordingly. A connection between law and action is assumed by the most diverse approaches to the law; when no reference is made to this connection it is since it appears to be an obvious truism. Let us list just a few examples where this connection appears most clearly.

The Digest of Justinian, while not explicitly linking law to action, distinguishes the way in which the law influences human action, by affirming that the law is meant to "command, forbid and punish" (D 1.7). Similarly, Cicero affirmed that the (just or rational) law "enjoins what ought to be done and forbids the opposite," i.e., that it has the function to "enjoin the right action and forbid wrong-doing" (Cicero 1998, n. 18).

The connection between law and action is also explicitly included in Aquinas's definition of the law as "a certain rule and measure of acts whereby man is induced to act or is restrained from acting" (Aquinas 1947, q90, a1).

Similarly, Grotius affirmed that "the law of nature is a dictate of right reason, which points out that an act, according as it is or is not in conformity with rational nature, has in it a quality of moral baseness or moral necessity; and that, in consequence, such an act is either forbidden or enjoined" (Grotius [1625] 1925, Book I, chap. I.x.1).

Leibniz also affirmed the connection between law and action (and between virtues and obligations of natural law) affirming that the legal obligation concerns "what it

The introduction to this chapter has been written by Emiliano Lorini and Giovanni Sartor.

E. Lorini (✉)
IRIT-CNRS Toulouse University, Toulouse, France
e-mail: Lorini@irit.fr

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_4

necessary for a good man to do," while permissibility covers "what it possible for a a good man to do" (Leibniz [1671] 1930, 431).

Moving from natural law to nineteenth-century positivism, we can find the connection between law and action in Jeremy Bentham's definition of law: "A law may be defined as an assemblage of signs declarative of a volition conceived or adopted by the sovereign in a state, concerning the conduct to be observed in a certain case by a certain person or class of persons" (Bentham [1872] 1970, 1).

The connection between law and action can also be found in Hans Kelsen's statement that "the norms of an order regulate [...] always human behavior—only it can be regulated by norms" (Kelsen 1967, 14–15). According to Kelsen, "the behavior regulated by a normative order is either a definite action or the omission (nonperformance) of such an action." In Kelsen's perspective sanctions are connecting to actions in two ways: they are meant to induce the omission of the punished actions, and they are established by authorizing "coercive action" (i.e., the imposition of sanctions) by state authorities.

The connection between law and action, action being the matter of legal regulations, is also affirmed by H. L. A Hart. For this author, both primary and secondary norms are concerned with actions: primary duty-imposing norms are meant to direct human actions and, on the other hand, secondary power-conferring norms also provide for the institutional effects of the actions through the exercise of such powers.

> Under rules of the one type, which may well be considered the basic or primary type, human beings are required to do or abstain from certain actions, whether they wish to or not. Rules of the other types are in a sense parasitic upon or secondary to the first; for they provide that human beings may by doing or saying certain things introduce new rules of the primary type, extinguish or modify old ones, or in various ways determine their incidence or control their operations. Rules of the first type impose duties; rules of the second type confer powers, public or private (Hart 1994).

In legal theory, a necessary connection between law and action is also affirmed by those approaches that view legal norms, and in particular duty-imposing norms as "reasons for action," at least when such norms are issued by a legitimate authorities (Raz 1979).

Besides being discussed or assumed in legal theory, the connection between law and action, has been the object of a vast doctrinal discussion in different domains of the law. In private law, it is often addressed in connection with the distinction of different kinds of triggers for legal effects. In particular, merely natural facts are traditionally distinguished from human acts, the latter consisting of behaviour being controlled, or at least controllable, by the agent. Within human acts, declarations of will (also called juristic acts) are often distinguished: they are acts consisting in declarations of legal outcomes (obligations or transfers between the parties) which are meant to produce such outcomes. Contracts are the most notable class of the latter acts. Criminal law focuses on the structure of criminal action, which includes discussions of the connection between the behavioural components (the so-called actus reus) and the corresponding mental states or attitudes (mens rea).

While the rich legal tradition of legal theory and legal doctrine provide many ideas for formal analysis of action—which cannot be addressed in this contribution—very few logical accounts of action have been so far provided by legal scholarships. Only a few contribution—at the interface of legal theory, philosophical logic, and, more recently, computing—have attempted at formalizing the action component of legal norms.

In particular, the concept of an act is at the core of the logical analysis of legal prescriptions developed by Von Wright (1963), the founder of modern deontic logic. According to this author, the law consists of behavioural prescriptions, i.e., of "commands, permissions, and prohibitions, which are given or issued to agents concerning their conduct." Correspondingly, legal norms are modelled by applying deontic modal operators for obligation and permission to action descriptions. Von Wrights's logic of action is connected to the idea of a transition between the states of affairs existing before and after the act: "acts may quite appropriately be described as the bringing about or effecting (at will?) of a change." Two types of acts are distinguished, actions and forbearances, which consist in bringing about, or in refraining from bringing about such changes.

The logical connection between laws and action was also addressed by Alchourrón and Bulygin (1971), who postulate that a set of possible actions is available (the universe of action), whose truth-functional compounds provide possible contents for deontic operators.

A critical analysis of Von Wright's theory was provided by Ota Weinberger (1998), a leading legal theorist and logician, who argued that the law requires a different approach to action, which he calls a "formal-finalistic action theory." According to this approach, an action is viewed as the outcome of a choice based on an information process, whose input includes both factual and normative information. This process involves the solution of an optimization problem, which may require taking into account multiple goals as well as constraints over means. Unfortunately, Weiberger did not provide a formal model of its theory of action.

A formally developed logic of norm and action, meant to address legal content, was developed by Kanger (1972). After characterizing "a system of law" as "a set of rules which has the purpose of regulating human action under certain conditions of law," Kanger modelled the norms of such a system by combining deontic logic and action logic. The basic idea of Kanger's approach to action is that an agent brings about a result—she does the action of bringing it about—when the agent's behaviour is both a necessary and a sufficient condition for the result to be produced. Action description of this kind can then be the object of deontic operators. Kanger's approach has been refined and developed in particular by Pörn (1977), who defined a logic of "bringing it about that" (BIAT), and developed a modal semantics for it. According to this logic, agent i brings about a state of affairs φ if and only if two conditions hold: it is necessary for everything that i does that φ, and but for what i does it might be the case that $\neg\varphi$. In other terms given i's behaviour necessarily φ is the case, and without that behaviour $\neg\varphi$ might be the case. For instance, it may be said that I close the door if, given my behaviour, necessarily the door is closed and, without my behaviour, it might be open. Further developments of the BIAT logic have been proposed by a number of authors, among which Elgesem (1997), see also Governatori and Rotolo (2005). A very rich traditions, which we cannot examine in this paper concerns the combination of the BIAT action logic and Hohfeldian modalities, on which see (Lindahl 1977; Jones and Sergot 1996).

Only recently logics of action have been provided that are sufficiently general to represent norm-related concepts such as causality, responsibility, and influence. Such logics are needed to capture aspects of legal agency that so far have only been addressed by informal doctrinal accounts. In fact, to describe social interaction in a formal way, it is necessary to have a representation language that allows to describe, at the same time, the causal relations between the actions and their effects, the agents' action repertoires and capabilities as well as the effects of the joint actions of agents. Actions occur in time and have a duration. Thus, a logical theory of interaction requires a clear understanding of the relationship between actions and time.

One of the logics that has been used to represent the concepts of individual and joint actions is propositional dynamic logic (PDL) (Harel and Tiuryn 2000). PDL has been introduced in theoretical computer science about thirty years ago in order to represent the concept of a (computer) program and the basic operations on programs (e.g., sequential composition, non-deterministic choice, iteration, test). Consequently, the use of PDL in modelling interaction between agents (see, e.g., Schmidt et al. 2004) works under the assumption that actions can be conceived as programs executed by the agents in the system. PDL's semantics is based on the concept of labelled transition system, that is to say, a graph whose vertices represent possible states of the system and whose edges are labelled with actions of agents. These edges represent transitions between states that are determined by the execution of an atomic program. PDL has been also shown to be a valuable formal language for representing normative concepts such as obligation, prohibition and permission (Meyer 1988; van der Meyden 1996). In PDL atomic programs are abstract entities in the sense that their semantics is just specified in terms of state transitions.

A concrete variant of PDL, called DL-PA (Dynamic Logic of Propositional Assignments) (Balbiani et al. 2013; Tiomkin and Makowsky 1985), has also been proposed. Differently from PDL, in DL-PA atomic programs are concrete. Specifically, they are assignments of propositional variables (i.e., an atomic program consists in setting to either $\top$ or $\bot$ the value of a given propositional variable p). The DL-PA notion of action, viewed as a propositional assignment, is compatible with Von Wright's idea of viewing an action as bringing about or effecting (at will) of a change. It is also shared with other formal systems proposed in the recent years in artificial intelligence (AI) such as boolean games and the Coalition Logic of Propositional Control (CL-PC). CL-PC was introduced by van der Hoek and Wooldridge (2005) as a formal language for reasoning about capabilities of agents and coalitions in multi-agent environments. In this logic, the notion of capability is modelled by means of the concept of *control*. In particular, it is assumed that each agent i is associated with a specific finite subset Atm_i of the finite set of all atomic propositions Atm. Atm_i is the set of propositions *controlled* by the agent i. That is, the agent i has the ability to assign a (truth) value to each proposition Atm_i but cannot affect the truth values of the propositions in $Atm \setminus Atm_i$. It is also assumed that control over propositions is exclusive, that is, two agents cannot control the same proposition (i.e., if $i \neq j$ then $Atm_i \cap Atm_j = \emptyset$). Moreover, it is assumed that control over propositions is complete, that is, every proposition is controlled by at least one agent (i.e., for every $p \in Atm$ there exists an agent i such that $p \in Atm_i$).

Boolean games (Harrenstein et al. 2001; Bonzon et al. 2006) share with CL-PC the idea that an agent's action consists in affecting the truth values of the variables she controls. They are games in which each player wants to achieve a certain goal represented by a propositional formula: they correspond to the specific subclass of normal form games in which agents have binary preferences (i.e., payoffs can be either 0 or 1). They have been proved to provide a useful and natural abstraction for reasoning about social interaction in multi-agent systems.

An alternative approach to the logical formalization of action and of the connection between actions and norms in multi-agent domains is the logic STIT (the logic of "seeing to it that"), which was introduced for the first time in the philosophical area (Belnap et al. 2001; Horty 2001; Horty and Belnap 1995) and has become popular in AI in the recent years. This logic is well-suited to represent the concept of causality (whether an agent brings about a certain state of affairs as a result of her current choice) as well as social concepts such as the concepts of responsibility, guilt, delegation, and social influence that are of primary importance in modelling social relations between human and artificial agents. Two variants of STIT have been studied in the literature which differ at the syntactic level: an atemporal version and a temporal version. The temporal version of STIT is a combination of temporal operators of temporal logic for expressing temporal properties of facts (e.g., whether a given fact φ will be true in the future) and operators of agency that allow to express the consequences of the choice of an agent or group of agents. The language of atemporal STIT is nothing but the language of temporal STIT restricted to the agency operators which does not include temporal operators. A central assumption of STIT is that agents' choices are independent, in the sense that an agent can never be deprived of choices due to the choices made by other agents. This distinguishes STIT from the BIAT approach in which agents' choices are not necessarily independent (see Sergot 2014 for a discussion on this point).

Other logical systems have been proposed in the recent years which move from the concept of action to the game-theoretic concept of strategy. Informally speaking, a strategy for a certain agent specifies, for every state of the system characterized by a tree or by a transition system, what the agent is expected to do at this state of the system. The most representative example of strategy logics is alternating-time temporal logic (ATL) (Alur et al. 2002) which can be seen as the strategic variant of Coalition Logic (CL) by Pauly (2002) and which allows to formally represent the consequences of the strategy of a certain agent or coalition of agents.[1]

Plan of the chapter The aim of the present chapter is to show how logic can be used for formalizing legal concepts in which the notion of action plays a fundamental role. Given the diversity of existing logics of action, we have decided to focus on STIT logic. There are two main motivations behind this decision. First of all, STIT offers a general framework for modelling action and time and for representing the

[1] Differently from STIT, CL can only represent the consequences of the choice of a certain player or coalition players, a choice being the restriction of a strategy to the current state of the system.

consequence of an agent's choice. As we will show, the latter captures a fundamental aspect of agent causation that is relevant for legal theory. Secondly, the formal semantics of STIT is extremely elegant and intuitive. Moreover, it is directly connected with the formal representation of action and action-related concepts (e.g., power and capability) used in game theory. From this perspective, STIT has a high-level generality, as it can be seen as the prototypical logic of action based on a game-theoretic semantics.

The paper is organized as follows. Section 2 is devoted to illustrate in a rather informal way the semantics of STIT as well as its formal language. Section 3 is devoted to illustrate the use of STIT for the formalization of responsibility and influence, two concepts that are highly relevant for legal theory. Finally, Sect. 4 presents a simple extension of STIT by the concept of "obligation to do" based on a utilitarian view of norms.

2 A Logic for Reasoning About Choices, Actions, and Time

STIT logic (the logic of *seeing to it that*) by Belnap et al. (2001) is one of the most prominent formal accounts of agency. It is the logic of sentences of the form "the agent i sees to it that φ is true." Different semantics for STIT have been proposed in the literature (Belnap et al. 2001; Broersen 2011; Wölf 2002; Schwarzentruber 2012). The original semantics of STIT by Belnap et al. (2001) is defined in terms of **BT+AC** structures: branching time structures (**BT**) augmented by agent choice functions (**AC**). A **BT** structure is made of a set of moments and a tree-like ordering over them. An **AC** for an agent i is a function mapping each moment m into a partition of the set of histories passing through that moment, a history h being a maximal set of linearly ordered moments and the equivalence classes of the partition being the possible choices for agent i at moment m.

In Lorini (2013), a Kripke-style semantics for STIT has been proposed. On the conceptual side, the main difference between this Kripke semantics for STIT and Belnap et al.'s **BT+AC** semantics is that the former takes the concept of *world* as a primitive instead of the concept of *moment* and defines: (i) a *moment* as an equivalence class induced by a certain equivalence relation over the set of worlds, (ii) a *history* as a linearly ordered set of worlds induced by a certain partial order over the set of worlds, and (iii) an agent i's set of *choices* at a moment as a partition of that moment. The main advantage of the Kripke semantics for STIT over Belnap et al.'s original semantics in terms of **BT+AC** structures is that the former is a standard multi-relational semantics commonly used in the area of modal logic (Blackburn et al. 2001), whereas the latter is non-standard.

It is worth noting that, at the semantic level, temporal STIT can be conceived as a logic of action interpreted over infinitely repeated games. This highlights the connection between STIT and game theory.

The Kripke semantics of STIT is illustrated in Fig. 1, where each moment m_1, m_2, and m_3 consists of a set of worlds represented by points. For example, moment

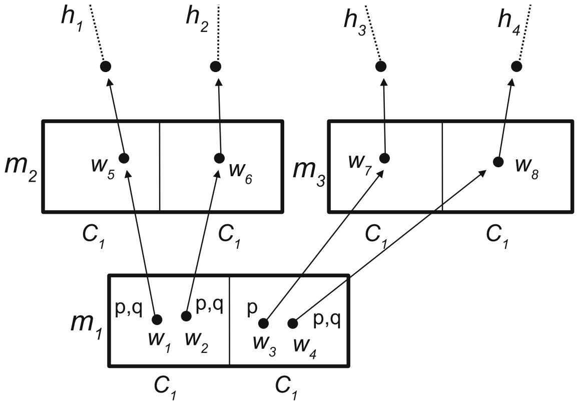


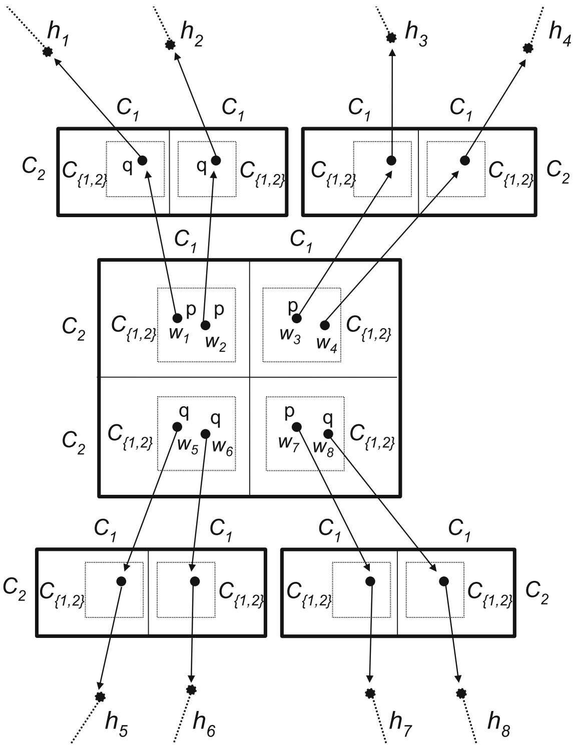
Fig. 1 Illustration of Kripke semantics of STIT

m_1 consists of the set of worlds $\{w_1, w_2, w_3, w_4\}$. Moreover, for every moment there exists a set of histories passing through it, where a history is defined as a linearly ordered set of worlds. For example, the set of histories passing through moment m_1 is $\{h_1, h_2, h_3, h_4\}$. Finally, for every moment, there exists a partition which characterizes the set of available choices of agent 1 in this moment. For example, at moment m_1, agent 1 has two choices available, namely $\{w_1, w_2\}$ and $\{w_3, w_4\}$. Note that an agent's set of choices at a certain moment can also be seen as a partition of the set of histories passing through this moment. For example, we can identify the choices available to agent 1's at m_1 with the two sets of histories $\{h_1, h_2\}$ and $\{h_3, h_4\}$. Following Horty (2001), the Kripke semantics for STIT also account for collective choices of groups of agents. Specifically, the choice of a group coincides with the intersection of the choices of the agents in the group. For instance, in Fig. 2, the individual choices of agents 1 and 2 are, respectively, $\{w_1, w_2, w_5, w_6\}$ and $\{w_1, w_2, w_3, w_4\}$, while the collective choice of group $\{1, 2\}$ is $\{w_1, w_2\}$.

Clearly, for every moment m in a Kripke semantics for STIT, one can identify the set of histories passing through it by considering all histories that contain at least one world in the moment m. Moreover, an agent i's set of choices available at m can also be seen as a partition of the set of histories passing through m. At first glance, an important difference between Belnap et al.'s semantics and Kripke semantics for STIT seems to be that in the former the truth of a formula is relative to a moment-history pair m/h, also called *index*, whereas in the latter it is relative to a world w. However, this difference is only apparent, because in the Kripke semantics for STIT there is a one-to-one correspondence between worlds and indexes, in the sense that: (i) for every index *m/h* there exists a unique world w at the intersection between m and h, (ii) and for every world w there exists a unique index *m/h* such that the intersection between m and h includes w.

In the Kripke semantics for STIT the concept of world should be understood as a "time point" and the equivalence class defining a moment as a set of alternative concomitant "time points." In this sense, the concept of moment captures a first aspect of indeterminism, as it represents the alternative ways the *present* could be.

Fig. 2 Kripke semantics of STIT with groups



A second aspect of indeterminism is given by the fact that moments are related in a (tree-like) branching time structure. In this sense, the *future* could evolve in different ways from a given moment. In the Kripke semantics for STIT these two aspects of indeterminism are related, as illustrated in Fig. 1. Indeed, if two distinct moments m_2 and m_3 are in the future of moment m_1, then there are two distinct worlds in m_1 (w_1 and w_3) such that a successor of the former (w_5) is included in m_2 and a successor of the latter (w_7) is included in m_3.

In Horty (2001), the following language of temporal STIT (TSTIT), i.e., the variant of STIT with tense operators, is considered:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid [J \ \mathsf{stit}]\varphi \mid \Box\varphi \mid \mathsf{G}\varphi \mid \mathsf{H}\varphi$$

where p ranges over an infinite set of atomic propositions Atm and J ranges over a set of groups of agents $X \subseteq 2^{Agt} \setminus \{\emptyset\}$, where Agt is a finite set of agents whose elements are denoted by $i, j, \ldots$ For notational convenience, we write $[i \ \mathsf{stit}]$ instead of $[\{i\} \ \mathsf{stit}]$ with $i \in Agt$.

When $X = 2^{Agt} \setminus \{\emptyset\}$, the previous TSTIT language allows us to talk about time, the consequences of all agents' individual choices as well as the consequences of all groups of agents' collective choices.

Let us discuss the meaning of the different modal operators. The formal language of TSTIT includes the future tense operator G and the past tense operator H, where $\mathsf{G}\varphi$ and $\mathsf{H}\varphi$, respectively, stand for "φ will always be true in the future" and "φ has always been true in the past." For example, the formula $\mathsf{G}\neg p$ is true at world w_1 in Fig. 1. Indeed, it is the case that p is false at all future worlds of w_1. Moreover, the formula $\mathsf{H}p$ is true at world w_5 since it is the case that p is true at all past worlds of w_5. In Lorini and Sartor (2016), a variant of temporal STIT with discrete time is studied. It includes the "next" operator X (where $\mathsf{X}\varphi$ stands for "φ is going to be true in the next world") and the "yesterday" operator Y (where $\mathsf{Y}\varphi$ stands for "φ was true in the previous world").

Moreover, the previous TSTIT language also includes the so-called historical necessity operator $\Box$ which allows us to represent those facts that are necessarily true, in the sense of being true at every point of a given moment or, equivalently, at every history passing through a given moment. For example, the formula $\Box p$ is true at world w_1 in Fig. 1 since p is true at every point of moment m_1 including world w_1.

The "historical possibility" operator $\Diamond$ is defined to be the dual operator of $\Box$, $\langle J\ \mathsf{stit}\rangle$ is defined to be the dual operator of $[J\ \mathsf{stit}]$, while F and P are defined to be the dual operators of G and H. That is:

$$\begin{aligned}\Diamond\varphi &\stackrel{\text{def}}{=} \neg\Box\neg\varphi\\ \langle J\ \mathsf{stit}\rangle\varphi &\stackrel{\text{def}}{=} \neg[J\ \mathsf{stit}]\neg\varphi\\ \mathsf{F}\varphi &\stackrel{\text{def}}{=} \neg\mathsf{G}\neg\varphi\\ \mathsf{P}\varphi &\stackrel{\text{def}}{=} \neg\mathsf{H}\neg\varphi\end{aligned}$$

STIT logic provides for different concepts of agency, all characterized by the fact that an agent acts only if she sees to it that a certain state of affairs is the case. In the previous TSTIT language, the so-called Chellas STIT operator $[i\ \mathsf{stit}]$, named after its proponent (Chellas 1992), is taken as a primitive. According to the STIT semantics, an agent i Chellas-sees-to-it that φ, denoted by formula $[i\ \mathsf{stit}]\varphi$, at a certain world w if and only if, for every world v, if w and v belong to the same choice of agent i then φ is true at world v. For example, in Fig. 1, agent 1 Chellas-sees-to-it that p at world w_1 because p is true both at world w_1 and at world w_2. The previous TSTIT language generalizes the Chellas STIT operator to groups of agents. For example, suppose $Agt = \{1, 2\}$. Then, in Fig. 2, it is the case that group $\{1, 2\}$ Chellas-sees-to-it that p, denoted by formula $[\{1, 2\}\ \mathsf{stit}]p$, at world w_1, because $\{w_1, w_2\}$ corresponds to the collective choice of group $\{1, 2\}$ at w_1, and p is true both at world w_1 and at world w_2.

A more sophisticated operator of agency is the deliberative STIT (Horty and Belnap 1995) which is defined as follows by means of the Chellas STIT operator and the historical necessity operator $\Box$:

$$[i\ \mathsf{dstit}]\varphi \overset{\mathtt{def}}{=} [i\ \mathsf{stit}]\varphi \wedge \Diamond\neg\varphi$$

In other words, deliberative STIT satisfies the same positive condition as Chellas STIT *plus* a negative condition: an agent i deliberately sees to it that φ, denoted by formula $[i\ \mathsf{dstit}]\varphi$, at a certain world w if and only if: (i) agent i Chellas-sees-to-it that φ at w, that is to say, agent i's current choice at w ensures φ, and (ii) at w agent i could make a choice that does not necessarily ensure φ.[2] Notice that the latter is equivalent to say that there exists a world v such that w and v belong to the same moment and φ is false at v. For example, in Fig. 1, agent 1 deliberately sees to it that q at world w_1 because q is true both at world w_1 and at world w_2, while being false at world w_3. In other terms, while the truth of $[i\ \mathsf{stit}]\varphi$ only requires that i's choice ensures that φ, the truth of $[i\ \mathsf{dstit}]\varphi$ also requires that i had the opportunity of making an alternative choice that would not guarantee that φ would be the case. Deliberative STIT captures a fundamental aspect of the concept of action, namely the idea that for a state of affairs to be the consequence of an action (or for an action to be the cause of a state of affairs), it is not sufficient that the action is a sufficient condition for that state of affairs to hold, it is also required that, without the action, the state of affairs possibly would not hold. In this sense, while $[i\ \mathsf{stit}]\varphi$ at w is consistent with (and is indeed entailed by) the necessity of φ at w, $[i\ \mathsf{dstit}]\varphi$ at w is incompatible with the necessity of φ at w, since it requires that at w also $\neg\varphi$ was an open possibility. Consequently, the deliberative STIT is more appropriate than the Chellas STIT to describe the consequences of an agent's action, as *incompatibility with necessity* is a requirement for any reasonable concept of action.[3]

In Lorini (2013), a sound and complete axiomatization for the fragment of the previous TSTIT language in which $X = \{\{i\} \mid i \in Agt\} \cup \{Agt\}$, with respect to the STIT Kripke semantics, is provided. It is summarized in Fig. 3.

This includes all tautologies of classical propositional calculus **(PC)** as well as modus ponens **(MP)**. Moreover, we have all principles of the normal modal logic S5 for every operator $[i\ \mathsf{stit}]$, for the operator $[Agt\ \mathsf{stit}]$ and for the operator $\Box$, all principles of the normal modal logic KD4 for the future tense operator G and all principles of the normal modal logic K for the past tense operator H. That is, we have Axiom K for each operator: $(\blacksquare\varphi \wedge \blacksquare(\varphi \rightarrow \psi)) \rightarrow \blacksquare\psi$ with $\blacksquare \in \{\Box, \mathsf{G}, \mathsf{H}, [Agt\ \mathsf{stit}]\} \cup \{[i\ \mathsf{stit}] \mid i \in Agt\}$. We have Axiom D for the future tense modality G: $\neg(\mathsf{G}\varphi \wedge \mathsf{G}\neg\varphi)$. We have Axiom 4 for $\Box$, G, $[Agt\ \mathsf{stit}]$ and for every $[i\ \mathsf{stit}]$: $\blacksquare\varphi \rightarrow \blacksquare\blacksquare\varphi$

[2] We shall not consider here the achievement STIT operator by Belnap and Perloff (1988) which provide a more sophisticated account of agency but whose interpretation is considerably more complicated than the semantics of the deliberative STIT.

[3] The classical argument against the use of Chellas STIT for modelling action is that, according to Chellas STIT, an agent brings about all tautologies and that it is counterintuitive to say that a tautology is a consequence of an agent's action.

PC	All tautologies of classical propositional calculus
S5(i)	All S5-principles for the operators $[i\ \mathsf{stit}]$
S5($\Box$)	All S5-principles for the operator $\Box$
S5(Agt)	All S5-principles for the operator $[Agt\ \mathsf{stit}]$
KD4(G)	All KD4-principles for the operator G
K(H)	All K-principles for the operator H
($\Box \to i$)	$\Box\varphi \to [i\ \mathsf{stit}]\varphi$
($i \to Agt$)	$([1\ \mathsf{stit}]\varphi_1 \wedge \ldots \wedge [n\ \mathsf{stit}]\varphi_n) \to [Agt\ \mathsf{stit}](\varphi_1 \wedge \ldots \wedge \varphi_n)$
(AIA)	$(\Diamond[1\ \mathsf{stit}]\varphi_1 \wedge \ldots \wedge \Diamond[n\ \mathsf{stit}]\varphi_n) \to \Diamond([1\ \mathsf{stit}]\varphi_1 \wedge \ldots \wedge [n\ \mathsf{stit}]\varphi_n)$
(Conv$_{\mathsf{G},\mathsf{H}}$)	$\varphi \to \mathsf{GP}\varphi$
(Conv$_{\mathsf{H},\mathsf{G}}$)	$\varphi \to \mathsf{HF}\varphi$
(Connected$_{\mathsf{G}}$)	$\mathsf{PF}\varphi \to (\mathsf{P}\varphi \vee \varphi \vee \mathsf{F}\varphi)$
(Connected$_{\mathsf{H}}$)	$\mathsf{FP}\varphi \to (\mathsf{P}\varphi \vee \varphi \vee \mathsf{F}\varphi)$
(NCUH)	$[Agt\ \mathsf{stit}]\mathsf{G}\varphi \to \mathsf{G}\Box\varphi$
(MP)	$\dfrac{\varphi, \varphi \to \psi}{\psi}$
(IRR)	$\dfrac{(\Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p)) \to \varphi}{\varphi}$, provided p does not occur in φ

Fig. 3 Axiomatization of for the TSTIT with agency operators $[i\ \mathsf{stit}]$ and $[Agt\ \mathsf{stit}]$

with $\blacksquare \in \{\Box, [Agt\ \mathsf{stit}], \mathsf{G}\} \cup \{[i\ \mathsf{stit}] \mid i \in Agt\}$. Furthermore, we have Axiom T for $\Box$, $[Agt\ \mathsf{stit}]$ and for every $[i\ \mathsf{stit}]$: $\blacksquare\varphi \to \varphi$ with $\blacksquare \in \{\Box, [Agt\ \mathsf{stit}]\} \cup \{[i\ \mathsf{stit}] \mid i \in Agt\}$. We have Axiom B for $\Box$, $[Agt\ \mathsf{stit}]$ and for every $[i\ \mathsf{stit}]$: $\varphi \to \blacksquare\neg\blacksquare\neg\varphi$ with $\blacksquare \in \{\Box, [Agt\ \mathsf{stit}]\} \cup \{[i\ \mathsf{stit}] \mid i \in Agt\}$. Finally we have the rule of necessitation for each modal operator: $\frac{\varphi}{\blacksquare\varphi}$ with $\blacksquare \in \{\Box, [Agt\ \mathsf{stit}], \mathsf{G}, \mathsf{H}\} \cup \{[i\ \mathsf{stit}] \mid i \in Agt\}$.

($\Box \to i$) and **(AIA)** are the two central principles in Xu's axiomatization of the Chellas's STIT operators $[i\ \mathsf{stit}]$ (Xu 1998). According to Axiom ($\Box \to i$), if φ is true regardless of what every agent does, then every agent sees to it that φ. In other words, an agent brings about those facts that are inevitable.[4] According to Axiom ($i \to Agt$), all agents bring about together what each of them brings about individually.

We have principles for the tense operators and for the relationship between time and action. **(Connected$_{\mathsf{G}}$)** and **(Connected$_{\mathsf{H}}$)** are the basic axioms for the linearity of the future and for the linearity of the past (Goldblatt 1992). **(Conv$_{\mathsf{G},\mathsf{H}}$)** and **(Conv$_{\mathsf{H},\mathsf{G}}$)** are the basic interaction axioms between future and past of minimal tense logic according to which "what is, will always have been" and "what is, has always been going to be."

[4] Xu considers a family of axiom schemas (**AIA**$_k$) for independence of agents of the form $(\Diamond[1\ \mathsf{stit}]\varphi_1 \wedge \ldots \wedge \Diamond[k\ \mathsf{stit}]\varphi_k) \to \Diamond([1\ \mathsf{stit}]\varphi_1 \wedge \ldots \wedge [k\ \mathsf{stit}]\varphi_k)$ that is parameterized by the integer k. As pointed out by (Belnap et al. 2001), (**AIA**$_{k+1}$) implies (**AIA**$_k$). Therefore, as Agt is finite, in OPDL the family of axiom schemas can be replaced by the single axiom **(AIA)**.

Axiom **(NCUH)** corresponds to so-called property of “no choice between undivided histories” which is implicit in the Kripke semantics for STIT illustrated above: if in some future world φ will be possible then the actual collective choice of all agents will possibly result in a state in which φ is true.

(IRR) is a variant of the well-known Gabbay’s irreflexivity rule that has been widely used in the past for proving completeness results for different kinds of temporal logic in which time is supposed to be irreflexive (see, e.g., Gabbay et al. 1994; Zanardo 1996; Reynolds 2003; von Kutschera 1997). The idea is that the irreflexivity for time, although not definable in terms of an axiom, can be characterized in an alternative sense by means of the rule **(IRR)**. This rule is perhaps more comprehensible if we consider its contrapositive: if p does not occur in φ and φ is TSTIT consistent, then $\Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p) \wedge \varphi$ is TSTIT consistent.

3 Formalization of Responsibility and Influence

This section is devoted to illustrate the application of STIT to the logical formalization of the concepts of responsibility (Sect. 3.1) and social influence (Sect. 3.2) which requires a comprehensive theory of the relationship between action and time.

3.1 Responsibility

The concept of responsibility is highly relevant not only for the legal domain but also for AI. Specifically, it has been proved to be useful in the domain of autonomous agents and multi-agent systems (MASs). For instance, autonomous vehicles should be endowed with the capability of reasoning about their own responsibility and that of others. This kind of capability allows agents to identify those actions that might be blameworthy, because they do not conform to legal norms, and therefore refrain from performing them. Moreover, an intelligent virtual agent interacting with a human can be designed to recognize humans’ emotions such as guilt or pride and to act consequently. This specific capacity can be achieved by endowing the agent with the more general capability of reasoning about humans’ responsibility and humans’ beliefs about her own and others’ responsibility.

In Lorini et al. (2014) a formalization of the concept of responsibility in STIT is proposed. This focuses on both the consequences that the agent’s actual choices have for other agents and the consequences of the actions that the agent could have chosen to perform and did not. More generally, it considers both the active (the agent’s seeing to it that φ is the case) and passive (the agent’s preventing φ from happening) dimensions of responsibility. The former is also called *responsibility for action*, while the latter is also called *responsibility for omission*. In particular:

An agent i is actively responsible for ensuring that a certain fact is true if and only if, i sees to it that the fact is true, regardless of what the others have decided to do.

An agent i is passively responsible for ensuring that a certain fact is true if and only if, the fact is actually true and i could have prevented it from being true, regardless of what the others have decided to do.

The previous concept of active responsibility is captured by the deliberative STIT operator introduced in Sect. 2. This justifies the following abbreviation:

$$[i\ \mathsf{aresp}]\varphi \stackrel{\text{def}}{=} [i\ \mathsf{dstit}]\varphi$$

where $[i\ \mathsf{aresp}]\varphi$ has to be read "agent i is *actively* responsible for ensuring φ." The previous concept of passive responsibility is captured by the following abbreviation:

$$[i\ \mathsf{presp}]\varphi \stackrel{\text{def}}{=} \varphi \wedge \langle Agt\setminus\{i\}\ \mathsf{stit}\rangle[Agt\ \mathsf{stit}]\neg\varphi$$

where $[i\ \mathsf{presp}]\varphi$ has to be read "agent i is *passively* responsible for ensuring φ." According to this definition, agent i is passively responsible for ensuring φ if and only if, φ is actually true (φ) and i could have prevented φ from being true, regardless of what the others have decided to do ($\langle Agt\setminus\{i\}\ \mathsf{stit}\rangle[Agt\ \mathsf{stit}]\neg\varphi$).

We have already illustrated the meaning of the operator $[i\ \mathsf{dstit}]$. Let us illustrate the meaning of the passive responsibility operator $[i\ \mathsf{presp}]$ via the example of STIT model in Fig. 2. At world w_1 agent 2 is passively responsible for making p true. Indeed, p is true at w_1. Moreover, it is the case that, agent 2 could have prevented p from being true, regardless of what agent 1 has decided to do. The latter condition captures the fundamental counterfactual aspect of the concept of passive responsibility.

3.2 *Influence*

In Lorini and Sartor (2016), a STIT logic analysis of the concept of social influence is proposed. It starts from a general view about the way rational agents make choices. Specifically, the assumption is that an agent might have several choices or alternatives *available* defining her *choice set* at a given moment, and that what the agent does is determined by her *actual* choice, which is in turn determined by the agent's *choice context* including her preferences and beliefs and the composition of her choice set. The analysis of social influence expands this view by assuming that the agent's choice context determining the agent's actual choice might be determined by external causes. Specifically, the external conditions in which an agent finds herself or the other agents with whom the agent interacts may provide an input to the agent's decision-making process in such a way that a determinate action should follow. Specifically, influence consists in *determining* the voluntary action of an agent by modifying her *choice*

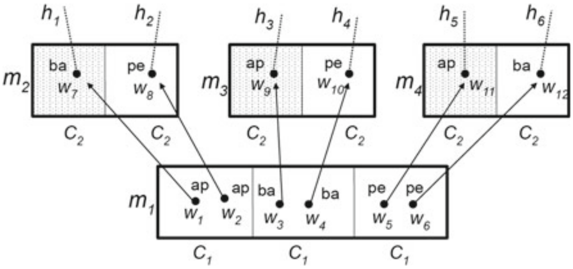


Fig. 4 Example of influence via choice restriction

context, so that a different choice becomes preferable to the influencee on comparison to what would be her preferred option without this modification. This may happen, for instance:

- by expanding the available choices (influence via choice set expansion), or
- by restricting the available choices (influence via choice set restriction) or
- by changing the payoffs associated with such choices, as when rewards or punishments are established (influence via payoff change).

The interesting aspect of STIT is that it is capable of: (i) capturing the temporal aspect of influence, namely the fact that the influencer's choice must precede the influencee's action,[5] and (ii) addressing the strategic aspect of influencing relationships through extensive form games.

To illustrate the concept of social influence, let us consider an example about influence via choice set restriction. The example is illustrated in Fig. 4. It represents a situation where there are three fruits on a table, an apple, a banana and a pear. The actions at issue consist in bringing about that the apple is eaten (*ap*), the banana is eaten (*ba*) or the pear is eaten (*pe*). Let us assume that agent 2 has certain preferences that remain constant along the tree structure. In particular, at all moments agent 2 prefers eating apples to bananas to pears. Let us also assume that 2 is rational, in the minimal sense that she acts in such a way as to achieve the outcome she prefers. Rational choices of agent 2 are depicted in grey. By choosing to eat the apple at w_1, 1 generates a situation where, given her preferences, 2 will necessarily eat the banana, rather than the pear. Indeed, although at moment m_2, 2 has two choices available, namely the choice of eating the banana and the choice of eating the pear, only the former is rational, in the sense of being compatible with 2's preferences. In this sense, by deciding to eat the apple at w_1 and removing this option from 2's choice set, 1 influences 2 to decide to eat the banana at w_7.

This example leads us to the following informal definition of social influence:

> An agent i influences another agent j to perform a certain (voluntary) action if and only if, i sees to it that that every rational choice of j will lead j to perform the action.

[5] The term "influencer" refers to the agent who exerts influence, whereas the term "influencee" refers to the agent being influenced.

In order to formalize the previous concept of social influence, in Lorini and Sartor (2016) STIT logic is extended by special "rational" STIT operators of the form $[i\ \mathsf{rdstit}]$. The formula $[i\ \mathsf{rdstit}]\varphi$ has to be read "if agent i's current action is the result of a rational choice of i, then i deliberately sees to it that φ." A minimal concept of rationality is adopted: it is assumed that the choices of an agent are ranked according to the agent's preferences, and an agent is rational as long as she implements her preferred choices. The $[i\ \mathsf{rdstit}]$ operator is interpreted relatively to STIT branching time structures, like the ones illustrated in Sect. 2. Specifically, the formula $[i\ \mathsf{rdstit}]\varphi$ is true at a certain world w if and only if, *if* the actual choice to which world w belongs is a rational choice of agent i *then*, at world w agent i deliberately sees to it that φ, in the sense of deliberative STIT discussed in Sect. 2. For example, at the world w_7 in Fig. 4, the formula $[2\ \mathsf{rdstit}]ba$ is true since the actual choice to which world w_7 belongs is a rational choice of agent 2 *and* at w_7 agent 2 deliberately sees to it that ba is the case.

To capture the idea of social influence, the following social influence operator based on the concept of deliberative STIT is introduced:

$$[i\ \mathsf{sinfl}\ j]\varphi \overset{\text{def}}{=} [i\ \mathsf{dstit}]\,\mathsf{X}\,[j\ \mathsf{rdstit}]\varphi$$

In other words, we shall say that an agent i influences another agent j to make φ true, denoted by $[i\ \mathsf{sinfl}\ j]\varphi$, if and only if i deliberately sees to it that if agent j's current choice is rational then j is going to deliberately see to it that φ. The reason why the operator $[i\ \mathsf{dstit}]$ is followed by the temporal operator X is that influence requires that the influencer's choice precedes the influencee's action. On the contrary, $[j\ \mathsf{rdstit}]$ is not required to be followed by X since in STIT the concept of action is simply captured by the deliberative STIT operator which does not necessarily need to be followed by temporal modalities. In order to illustrate the meaning of the influence operator, let us go back to the example of Fig. 4. Since agent 2 prefers eating bananas to pears, her only rational choice at moment m_2 is $\{w_7\}$. From this assumption, it follows that formula $[1\ \mathsf{sinfl}\ 2]ba$ is true at world w_1. Indeed, at world w_1 agent 1 deliberately sees to it that, in the next world, if agent 2's choice is rational then 2 deliberately sees to it that ba is the case.

Note that $[i\ \mathsf{sinfl}\ j]\varphi$ just says that the influencee i would realize φ is she were choosing rationally, but it does not assume that i chooses rationally, and therefore it does not entail that φ would be realized. The notion of *successful* influence also requires that in the next world along the actual history, the influencee chooses rationally, as specified by the following abbreviation:

$$[i\ \mathsf{succsinfl}\ j]\varphi \overset{\text{def}}{=} [i\ \mathsf{sinfl}\ j]\varphi \wedge \mathsf{X}\ \mathsf{ratCh}_i$$

where $[i\ \mathsf{succsinfl}\ j]\varphi$ has to be read "agent i *successfully* influences agent j to make φ true." The expression ratCh_i means that agent i's current choice is rational. It is an abbreviation, adopted for notational convenience, of $\neg[i\ \mathsf{rdstit}]\bot$, a formula that is satisfied only when i acts rationally in the current word.

This operator of successful influence clearly implies that in the next world along the actual history the influencee performs the action for which she has been influenced. This is expressed by the following valid formula:

$$[i \ \mathsf{succsinfl} \ j]\varphi \rightarrow \mathsf{X}\,[j \ \mathsf{dstit}]\varphi$$

In Lorini and Sartor (2016) is was also provided a complete axiomatization for STIT logic of social influence.

3.3 The Relationship Between Influence and Responsibility

The connection between the concept of influence and the concept of responsibility is tight and particularly relevant for legal theory. As legal theorists have emphasized (see, e.g., Hart and Honoré 1985; Kadish 1985), there exists a form of responsibility which consists in inducing another agent to violate a certain norm. In this case, the influencer becomes indirectly responsible for the violation of the norm, thereby being subject to a sanction. This captures the core of the idea of indirect (also called secondary or accomplice) responsibility in legal systems. In private law, there may be a "contributory liability" when someone with knowledge of the infringing activity, induces, causes, or materially contributes to the tort performed by another. In criminal law the idea of secondary responsibility concerns the connection between author of the crime, who performed the "actus reus" punished by the law, and his accomplices, who contributed to the performance of the "actus," without being part of it. For instance, the author of a robbery is the person, who breaks into a bank, threatens the clerks with a weapon and steals the money. Accomplices may have provided the weapons for the robbery, performed preliminary inspection of the places, or acted as lookout.

In Lorini and Sartor (2015), a formal theory based on STIT of the connection between influence and responsibility is provided. The relevance of such a theory for artificial intelligence (AI) lies in the possibility of exploiting it for automatic verification of secondary responsibility. Indeed, as highlighted above, the notion of influence is required to direct blame and sanctions not only against those who directly perform damaging acts, but also against those who have induced the authors to perform such acts. In the regulation of a society of artificial and/or human agents these responsibilities must also be introduced and checked, to effectively target cooperation aimed at socially obnoxious activities. This is clear for future society in which human agents will delegate tasks to autonomous agents and robots and will induce them to perform certain actions. Since responsibility can only be ascribed to humans, in case of violation of a norm by such artificial entities, humans agents will have to be sanctioned on the basis of their secondary responsibility for the violation.

As highlighted in Sect. 3.1, two basic forms of responsibility shall be distinguished, active responsibility and passive responsibility. Consequently, two forms of secondary responsibility shall be considered, namely active secondary responsibility and passive secondary responsibility. Active secondary responsibility consists in an agent being actively responsible for the action of another agent and is captured by the following abbreviation:

$$[i \ \mathsf{saresp} \ j]\varphi \overset{\text{def}}{=} [i \ \mathsf{aresp}]\, \mathsf{X}\, [j \ \mathsf{rdstit}]\varphi \wedge \mathsf{X} \ \mathsf{ratCh}_i$$

where $[i \ \mathsf{saresp} \ j]\varphi$ has to be read "agent i is *actively secondarily* responsible for ensuring φ via agent j." This means that agent i is actively responsible for ensuring that every rational choice of agent j will result in φ true ($[i \ \mathsf{aresp}]\, \mathsf{X}\, [j \ \mathsf{rdstit}]\varphi$), under the condition that in the next world agent j will choose rationally ($\mathsf{X} \ \mathsf{ratCh}_i$). As the following valid formula highlights, it clearly coincides with the concept of successful influence defined in Sect. 3.2:

$$[i \ \mathsf{saresp} \ j]\varphi \leftrightarrow [i \ \mathsf{succsinfl} \ j]\varphi$$

Passive secondary responsibility consists in an agent being passively responsible for the action of another agent and is captured by the following abbreviation:

$$[i \ \mathsf{spresp} \ j]\varphi \overset{\text{def}}{=} [i \ \mathsf{presp}]\, \mathsf{X}\, [j \ \mathsf{rdstit}]\varphi \wedge \mathsf{X} \ \mathsf{ratCh}_i$$

where $[i \ \mathsf{spresp} \ j]\varphi$ has to be read "agent i is *passively secondarily* responsible for ensuring φ via agent j." This means that agent i is passively responsible for ensuring that every rational choice of agent j will result in φ true ($[i \ \mathsf{presp}]\, \mathsf{X}\, [j \ \mathsf{rdstit}]\varphi$), under the condition that in the next world agent j will choose rationally ($\mathsf{X} \ \mathsf{ratCh}_i$).

4 Deontic Extension

In Sect. 3.2, we have shown how STIT can be extended by a simple notion of rational (or preferred) choice. The idea is that at any given moment an agent i has a set of available choices. Only a subset of her available choices are rational and are denoted by the logical symbol ratCh_i. In particular, ratCh_i means that agent i's actual choice is rational.

In similar way, STIT can be extended by a simple notion of ideal choice denoted by the symbol idlCh_i. As for rational choices, the set of agent i's ideal choices at a given moment is a subset of agent i's available choices at this moment. An agent's ideal choices are those choices that are the best for the society, as they conform to its norms. Clearly, an agent's set of ideal choices does not necessarily coincide with her set of rational choices. For example, it might be ideal for an agent to help disadvantaged people in the society, even though it might be rational for her not to do it.

In the rest of the section, we explain how the selection of ideal choices out of the set of available choices is determined by the their ideality values. In particular, ideal choices are those choices maximizing the ideality value. This utilitarian view of norms is commonly accepted in the deontic logic area to provide a formal semantics for normative concepts such as obligation "to be" (Føllesdal and Hilpinen 1971; Anderson 1957), obligation "to do" (Horty 2001; Kanger 1972; Sergot 1999) and dyadic or conditional obligation (Hansson 1969; Prakken and Sergot 1997). For example, the formal semantics for standard deontic logic (SDL) is based on the general concept of *ideality* according to which the set of ideal worlds (or situations) is a subset of the set of possible worlds (for a discussion of SDL, see Hilpinen 1982; Hilpinen and McNamara 2013). This is nothing but a special case of a more general formal semantics for deontic logic according to which possible worlds should be ranked in terms of their ideality values.

Ideality values of histories and ideality values of choices We enrich the STIT language with special atomic formulas of type $\mathsf{valCh}_{i,\geq x}$ and $\mathsf{valHis}_{\geq x}$, where i ranges over the set of agents Agt and x ranges over the set $\mathbb{N} \cup \{\omega\}$. $\mathbb{N}$ is the set of natural numbers and ω is the lowest transfinite ordinal number that is larger than all finite numbers in $\mathbb{N}$. The reason why we include ω is that we do not want to exclude histories which have assigned an *infinite* ideality value. The constant $\mathsf{valCh}_{i,\geq x}$ has to be read "agent i's actual choice has an ideality value at least x," while the constant $\mathsf{valHis}_{\geq x}$ has to be read "the actual history has an ideality value at least x."

The basic properties of these constants are captured by the following eight logical axioms:

(**ValCh**$_0$)	$\mathsf{valCh}_{i,\geq 0}$
(**ValHis**$_0$)	$\mathsf{valHis}_{\geq 0}$
(**ValCh**$_>$)	$\mathsf{valCh}_{i,\geq x} \rightarrow \mathsf{valCh}_{i,\geq y}$ if $x > y$
(**ValHis**$_>$)	$\mathsf{valHis}_{\geq x} \rightarrow \mathsf{valHis}_{\geq y}$ if $x > y$
(**ChDetValCh1**)	$\mathsf{valCh}_{i,\geq x} \rightarrow [i\ \mathsf{stit}]\mathsf{valCh}_{i,\geq x}$
(**ChDetValCh2**)	$\neg\mathsf{valCh}_{i,\geq x} \rightarrow [i\ \mathsf{stit}]\neg\mathsf{valCh}_{i,\geq x}$
(**TimeDetValHis1**)	$\mathsf{valHis}_{\geq x} \rightarrow (\mathsf{G}\ \mathsf{valHis}_{\geq x} \wedge \mathsf{H}\ \mathsf{valHis}_{\geq x})$
(**TimeDetValHis2**)	$\neg\mathsf{valHis}_{\geq x} \rightarrow (\mathsf{G}\neg\mathsf{valHis}_{\geq x} \wedge \mathsf{H}\neg\mathsf{valHis}_{\geq x})$

The first and second axioms state that every choice and every history have at least ideality value 0, as we assume that every choice and every history are identified with a non-negative integer.[6] The third and fourth axioms state that if $x > y$ and a choice/history has an ideality value at least x, then it has an ideality value at least y. According to the fifth and sixth axioms, the ideality value of a choice is choice-determinate. Finally, according to the seventh and eighth axioms, the ideality value of a history is history-determinate.

The following two abbreviations captures the exact ideality value of a choice and a history:

[6] For simplicity, we do not consider negative utilities.

$$\mathsf{valCh}_{i,x} \overset{\text{def}}{=} \mathsf{valCh}_{i,\geq x} \wedge \neg\mathsf{valCh}_{i,\geq x+1}$$
$$\mathsf{valHis}_{x} \overset{\text{def}}{=} \mathsf{valHis}_{\geq x} \wedge \neg\mathsf{valHis}_{\geq x+1}$$

where $\mathsf{valCh}_{i,x}$ and valHis_{x} have to be read, respectively, "agent i's actual choice has an ideality value equal to x," and valHis_{x} has to be read "the actual history has an ideality value equal to x." By convention, we assume that $\mathsf{valCh}_{i,\geq\omega+1}$ and $\mathsf{valHis}_{\geq\omega+1}$ are equivalent to $\bot$. Thus, $\mathsf{valCh}_{i,\omega} \overset{\text{def}}{=} \mathsf{valCh}_{i,\geq\omega}$ and $\mathsf{valHis}_{\omega} \overset{\text{def}}{=} \mathsf{valHis}_{\geq\omega}$.

Connection between ideality values of choices and ideal choices As we explained above, the atomic formula idlCh_i is aimed to identify agent i's ideal choices, that is, agent i's best choices according to the norms and standards of the society.

The following axiom captures the idea that at every moment an agent has at least one ideal choice:

$$(\textbf{AtLeastOneIdl}) \quad \Diamond\mathsf{idlCh}_i$$

It is justified by the assumption that there exists at least one choice that maximizes the ideality value.

The connection between ideal choices and ideality values of choices is captured by the following logical axiom:

$$(\textbf{ValChIdl}) \quad (\mathsf{idlCh}_i \wedge \mathsf{valCh}_{i,x}) \rightarrow \Box\neg\mathsf{valCh}_{i,\geq x+1}$$

The axiom means that if agent i's actual choice is an ideal choice and its ideality value is equal to x, then every available choice of agent i has an ideality value at most x.

Connection between ideality values of histories and ideality values of choices An important issue we have not yet addressed is the connection between ideality values of histories and ideality values of choices. There are different ways to represent this connection. One way is to assume that the ideality value of a choice corresponds to the *minimal* ideality value of a history passing through this choice. This is specified by the following logical axiom:

$$(\textbf{ValChValHis}_{min}) \quad \mathsf{valCh}_{i,x} \leftrightarrow (\langle i\ \mathsf{stit}\rangle\mathsf{valHis}_{x} \wedge [i\ \mathsf{stit}]\mathsf{valHis}_{\geq x})$$

Another way is to assume that the ideality value of a choice corresponds to the *maximal* ideality value of a history passing through this choice. This is specified by the following logical axiom:

$$(\textbf{ValChValHis}_{max}) \quad \mathsf{valCh}_{i,x} \leftrightarrow (\langle i\ \mathsf{stit}\rangle\mathsf{valHis}_{x} \wedge [i\ \mathsf{stit}]\neg\mathsf{valHis}_{\geq x+1})$$

The two Axioms ($\textbf{ValChValHis}_{min}$) and ($\textbf{ValChValHis}_{max}$) should be conceived as alternative criteria for selecting, among an agent's available choices, her ideal choices. In particular, Axiom ($\textbf{ValChValHis}_{min}$) together with Axiom (**ValChIdl**) correspond to the *maxmin* criterion of selecting those choices whose worst possible outcome is better than the least possible outcome of all other available choices. Axiom (**ValChValHis**$_{max}$) together with Axiom (**ValChIdl**) correspond to the *maxmax* criterion of selecting those choices whose best possible outcome is better than the best possible outcome of all other available choices. Such criteria have been extensively studied in the area of qualitative decision theory (see, e.g., Brafman and Tennenholtz 1996, 2000; Goldszmidt and Pearl 1996).

Other criteria for selecting choices on the basis of ideality values of histories passing through them have been studied in the literature. For instance, Horty (2001) considers a selection criterion based on the concept of *dominance*. The idea is that an agent i's choice should be selected if and only if, there is no other choice of the agent that dominates it. It is said that agent i's choice A dominates choice B if and only if (i) for every possible choice of the other agents, choosing A is at least as good as choosing B, and (ii) there exists a possible choice of the others such choosing A is better than choosing B.

Obligation "to do" Following Horty (2001), we are finally able to formally define a concept of obligation "to do" (or "ought to do"):

$$\mathsf{Obg}_i\varphi \overset{\text{def}}{=} \Box[i\ \mathsf{idstit}]\varphi.$$

where:

$$[i\ \mathsf{idstit}]\varphi \overset{\text{def}}{=} \mathsf{idlCh}_i \rightarrow [i\ \mathsf{dstit}]\varphi$$

$\mathsf{Obg}_i\varphi$ has to be read "agent i is obliged to ensure φ," while $[i\ \mathsf{idstit}]\varphi$ has to be read "if agent i's current action is the result of an ideal choice of i, then i deliberately sees to it that φ." According to the previous definition, agent i has the obligation to ensure φ, denoted by $\mathsf{Obg}_i\varphi$, if and only if every ideal choice of agent i guarantees that agent i deliberately sees to it that φ.

Note that this concept of "ought to do" is essentially different from the concept of "ought to do" as formulated in the so-called Kanger and Lindhal's tradition (Lindahl 1977; Kanger 1972; Sergot 1999) by combining the "ought to be" operator of standard deontic logic (SDL) of the form O with the deliberative STIT operator $[i\ \mathsf{dstit}]$. As shown by Horty (2001), defining the normative sentence "agent i is obliged to ensure φ" by $\mathsf{O}[i\ \mathsf{dstit}]\varphi$ leads to counterintuitive consequences. For example, suppose agent i is a military general during a war. He has to decide whether (i) to bomb a enemy placement with the possibility destroying it without killing any civilian but with the potential risk of destroying it and killing civilians, or (ii) to refrain from bombing the enemy. The concept of "ought to do" represented by the combined operator $\mathsf{O}[i\ \mathsf{dstit}]$ tells us unambiguously that in this situation agent i has to bomb, as the only ideal situation is the one in which the enemy placement is destroyed and no civilian is killed. This conclusion is unsatisfactory, as it does not explain why the agent should unilaterally prefer one of the two options.

On the contrary, the previous concept of "ought to do" represented by the modal operator Obg_i is more flexible, as it tells us that the decision whether to bomb or not depends on the decision criterion adopted by the agent. In particular, it tells us that agent i should bomb if she adopts the *maxmax* criterion specified by Axiom (**ValChValHis**$_{max}$), whereas she should refrain from bombing if she adopts the *maxmin* criterion specified by Axiom (**ValChValHis**$_{min}$). Let us illustrate this in more detail. We present two different STIT models corresponding to these two different situations. Let proposition p denote the fact that "the enemy placement is destroyed" and proposition q denote the fact that "civilians are killed."

In the first model, represented in Fig. 5, ideality values of choices are determined via the maximality criterion formally represented by Axiom (**ValChValHis**$_{max}$) and ideal choices are determined via the corresponding *maxmax* criterion. In the second model, represented in Fig. 6, ideality values of choices are determined via the minimality criterion formally represented by Axiom (**ValChValHis**$_{min}$) and ideal choices are determined via the corresponding *maxmin* criterion.

Each history and each choice are identified with corresponding ideality values. In particular, history h_1 has an ideality value equal to 2 (formula valHis_2 is true at world w_1), history h_2 has an ideality value equal to 0 (formula valHis_0 is true at world w_2),

Fig. 5 Ideal choices determined via *maxmax* criterion

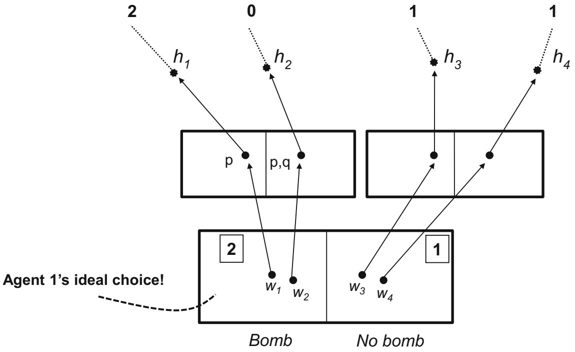
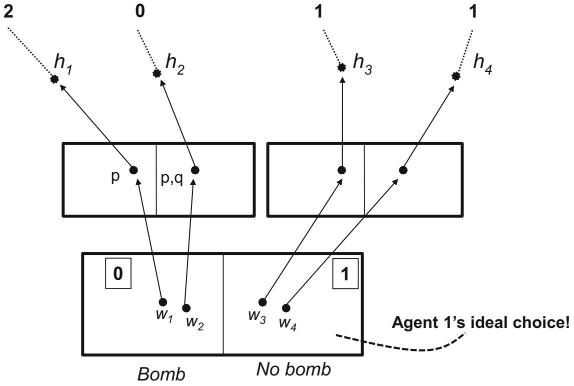


Fig. 6 Ideal choices determined via *maxmin* criterion



history h_3 has an ideality value equal to 1 (formula valHis_1 is true at world w_3), and history h_4 has an ideality value equal to 1 (formula valHis_1 is true at world w_4). We just assign arbitrary ideality values satisfying the following constraints: the situation in which the enemy placement is destroyed and no civilian is killed is better than the situation in which the enemy placement is not destroyed and no civilian is killed, and the situation in which the enemy placement is not destroyed and no civilian is killed is better than the situation in which the enemy placement is destroyed and civilians are killed.

Let us consider the model of Fig. 5. Agent 1's left choice in the initial moment containing world w_1 has an ideality value equal to 2 (formula $\mathsf{valCh}_{1,2}$ is true at worlds w_1 and w_2), while agent 1's right choice in the initial moment containing world w_1 has an ideality value equal to 1 (formula $\mathsf{valCh}_{1,1}$ is true at worlds w_3 and w_4). The ideality value of a choice corresponds to the *maximal* value of ideality of a history passing through it. It follows that idlCh_1 is true at worlds w_1 and w_2 and false at worlds w_3 and w_4. Indeed only agent 1's left choice is ideal. Consequently, agent 1 has the obligation to make p true (agent 1 has the obligation to destroy the enemy placement), as it is the case that p is true at every world included in agent 1's ideal choice. In particular, formulas $\mathsf{Obg}_1 p$ is at worlds w_1, w_2, w_3, and w_4. Furthermore, there exists a world in which p is false, which guarantees the satisfaction of the negative condition of the deliberative STIT formula $[1\ \mathsf{dstit}]p$.

On the contrary, in the model of Fig. 6, the ideality value of a choice corresponds to the *minimal* value of ideality of a history passing through it. It follows that idlCh_i is true at worlds w_3 and w_4 and false at worlds w_1 and w_2. Indeed, in this situation agent 1's right choice is the ideal one. Consequently, agent 1 has the obligation to make p false (agent 1 has the obligation to refrain from destroying the enemy placement), as it is the case that p is false at every world included in agent 1's ideal choice. Furthermore, there exists a world in which p is true, which guarantees the satisfaction of the negative condition of the deliberative STIT formula $[1\ \mathsf{dstit}]\neg p$.

5 Conclusion

Let us sum up what we have discussed so far. We have started with a concise presentation of the STIT formal language and semantics. Then, we have illustrated the use of STIT for the logical formalization of responsibility and influence, two concepts that are relevant for legal theory. Finally, we have presented a deontic extension of the STIT framework by a concept "ought to do," whose formal representation is based on the connection between the ideality value of a choice and the ideality value of a history passing through it.

An aspect we have not considered is the link between ideality values of histories and personal utilities of histories, where personal utilities are just the expressions of what the agents prefer. This is a fundamental component of norms of fairness and distributive justice. There are different ways of linking the two notions. For instance, Harsanyi's theory of fairness (Harsanyi 1982) provides support for an utilitarian interpretation of ideality according to which the ideality value of a history coincides with the sum of the individual utilities of this history for the agents. An alternative to Harsanyi's utilitarian view of ideality is Rawls' view (1971). In response to Harsanyi, Rawls proposed the *maximin* criterion of making the least happy agent as happy as possible: the ideality value of a history coincides with the minimal utility of this history for the agents.

Another aspect we have not considered is the relationship between norms and agents' preferences. In Sect. 3.2, we have discussed an extension of STIT by a concept of rational (or preferred) choice. In real situations, an agent's rational choices may differ from her ideal choices. The aim of a normative system is to reduce the gap between the agents' rational choices and her ideal choices, by finding the appropriate balance of rewards and punishments.

We believe these two perspectives are promising, as they would complement the present study of the relationship between norm and action with an analysis of the relationship between norm and mind by considering (i) how what is ideal for the society may depend on the agents' personal utilities, and (ii) how norms may influence an agent's decision-making process.

References

Alchourrón, C.E., and E. Bulygin. 1971. *Normative systems*. Cambridge: Springer.

Alur, R., T. Henzinger, and O. Kupferman. 2002. Alternating-time temporal logic. *Journal of the ACM* 49: 672–713.

Anderson, A.R. 1957. The formal analysis of normative concepts. *American Sociological Review* 22: 9–17.

Aquinas, T. 1947. *Summa theologica*. Allen, Tex.: Benzinger Bros.

Balbiani, P., A. Herzig, and N. Troquard. 2013. Dynamic logic of propositional assignments: A well-behaved variant of PDL. In *Proceedings of the 2013 28th annual ACM/IEEE symposium on logic in computer science (LICS 2013)*, 143–152. Amsterdam: Morgan Kaufmann Publishers.

Belnap, N., M. Perloff, and M. Xu. 2001. *Facing the future: Agents and choices in our indeterminist world*. New York, N.Y.: Oxford University Press.

Belnap, N., and M. Perloff. 1988. Seeing to it that: A canonical form for agentives. *Theoria* 54: 175–199.

Bentham, J. [1872] 1970. *Of laws in general*. London: Athlone.

Blackburn, P., M. de Rijke, and Y. Venema. 2001. *Modal logic*. Cambridge: Cambridge University Press.

Bonzon, E., J. Lang, M.-C. Lagasquie-Schiex, and B. Zanuttini. 2006. Boolean games revisited. In *Proceedings of the 17th European conference on artificial intelligence*, ed. A.P.G. Brewka, S. Coradeschi, and P. Traverso, 265–269. ACM

Brafman, R.I., and M. Tennenholtz. 1996. On the foundations of qualitative decision theory. In *Proceedings of the thirteenth national conference on artificial intelligence (AAAI'96)*, 1291–1296, Palo Alto. California: AAAI Press.

Brafman, R.I., and M. Tennenholtz. 2000. An axiomatic treatment of three qualitative decision criteria. *Journal of the ACM* 47 (3): 452–482.

Broersen, J. 2011. Deontic epistemic stit logic distinguishing modes of mens rea. *Journal of Applied Logic* 9 (2): 137–152.

Chellas, B.F. 1992. Time and modality in the logic of agency. *Studia Logica* 51 (3–4): 485–518.

Cicero, M.T. 1998. The laws. In *The republic and the laws*. Oxford: Oxford University Press.

Elgesem, D. 1997. The modal logic of agency. *Nordic Journal of Philosophical Logic* 2: 1–46.

Føllesdal, D., and R. Hilpinen. 1971. Deontic logic: An introduction. In *Deontic logic: Introductory and systematic reading*, ed. R. Hilpinen. Dordrecht: Reidel.

Gabbay, D.M., I. Hodkinson, and M.A. Reynolds. 1994. *Temporal logic: Mathematical foundations and computational*, vol. 1. Oxford: Clarendon Press.

Goldblatt, R. 1992. *Logics of time and computation*. Lecture Notes, Stanford, 2nd ed. California: CSLI Publications.

Goldszmidt, M., and J. Pearl. 1996. Qualitative probability for default reasoning, belief revision and causal modeling. *Artificial Intelligence* 84: 52–112.

Governatori, G., and A. Rotolo. 2005. On the axiomatisation of elgesem's logic of agency and ability. *Journal of Philosophical Logic* 34 (4): 403–431.

Grotius, H. [1625] 1925. *On the law of war and peace*, vol. 2. Oxford: Clarendon Press.

Hansson, B. 1969. An analysis of some deontic logics. *Mind* 3: 373–398.

Harel, D., D. K., and J. Tiuryn. 2000. *Dynamic logic*. Cambridge: MIT Press.

Harrenstein, P., J.-J. Meyer, W. van der Hoek, and C. Witteveen. 2001. Boolean games. In *Proceedings of the 8th conference on theoretical aspects of rationality and knowledge (TARK)*, 287–298. Amsterdam: Morgan Kaufmann Publishers.

Harsanyi, J. 1982. Morality and the theory of rational behaviour. In *Utilitarianism and beyond*, ed. A. Sen, and B. Williams. Cambridge: Cambridge University Press.

Hart, H.L.A., and T. Honoré. 1985. *Causation in the law*, 2nd ed. Oxford: Clarendon Press.

Hart, H.L.A. 1994. *The concept of law*, 2nd ed. Oxford: Oxford University Press.

Hilpinen, R., and P. McNamara. 2013. *Deontic logic: A historical survey and introduction*. London: College Publications.

Hilpinen, R. 1982. Deontic logic. In *The Blackwell guide to philosophical logic*, L. Goble ed. chap. 8, 159–182. Cambridge: Blackewell.

Horty, J.F., and N. Belnap. 1995. The deliberative STIT: A study of action, omission, and obligation. *Journal of Philosophical Logic* 24 (6): 583–644.

Horty, J.F. 2001. *Agency and deontic logic*. Oxford: Oxford University Press.

Jones, A.J., and M.J. Sergot. 1996. A formal characterisation of institutionalised power. *Logic Journal of the IGPL* 4: 429–445.

Kadish, S. 1985. Causation and complicity: A study in the interpretation of doctrine. *California Law Review* 73: 323–410.

Kanger, S. 1972. Law and logic. *Theoria* 38: 105–132.

Kelsen, H. 1967. *The pure theory of law*. Berkeley. California: University of California Press.

Leibniz, G.W. [1671] 1930. *Elementa Juris Naturalis*, vol. 1. Berlin: Akademie-Verlag.

Lindahl, L. 1977. *Position and change: A study in law and logic*. Dordrecht: Reidel.

Lorini, E., and G. Sartor. 2015. Influence and responsibility: A logical analysis. In *Proceedings of the twenty-eighth annual conference on legal knowledge and information systems (JURIX 2015)*. Frontiers in Artificial Intelligence and Applications, vol. 279, 51–60. Amsterdam: IOS Press.

Lorini, E., and G. Sartor. 2016. A STIT logic for reasoning about social influence. *Studia Logica* 104 (4): 773–812.

Lorini, E., D. Longin, and E. Mayor. 2014. A logical analysis of responsibility attribution : Emotions, individuals and collectives. *Journal of Logic and Computation* 24 (6): 1313–1339.

Lorini, E. 2013. Temporal STIT logic and its application to normative reasoning. *Journal of Applied Non-Classical Logics* 23 (4): 372–399.

Meyer, J.-J.C. 1988. A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic* 29 (1): 109–136.

Pauly, M. 2002. A modal logic for coalitional power in games. *Journal of Logic and Computation* 12 (1): 149–166.

Pörn, I. 1977. *Action theory and social science: Some formal models. synthese library*, vol. 120. Dordrech: Reidel.

Prakken, H., and M.J. Sergot. 1997. Dyadic deontic logics and contrary-to-duty obligations. In *Defeasible deontic logic*, ed. D. Nute, 223–262. Dordrecht: Kluwer.

Rawls, J. 1971. *A theory of justice*. Cambridge, Mass: Harvard University Press.

Raz, J. 1979. *The authority of law: Essays on law and morality*. Oxford: Clarendon Press.

Reynolds, M.A. 2003. An axiomatization of prior's ockhamist logic of historical necessity. In *Advances in modal logic*, vol. 4, 355–370. Amsterdam: King's College Publications.

Schmidt, R.A., D. Tishkovsky, and U. Hustadt. 2004. Interactions between knowledge, action and commitment within agent dynamic logic. *Studia Logica* 78 (3): 381–415.

Schwarzentruber, F. 2012. Complexity results of STIT fragments. *Studia Logica* 100 (5): 1001–1045.

Sergot, M. 1999. Normative positions. In *Norms, logics and information systems*, ed. P. McNamara, and H. Prakken, 289–308. Amsterdam: IOS Press.

Sergot, M. 2014. Some examples formulated in a 'seeing to it that' logic: Illustrations, observations, problems. In *Nuel Belnap on indeterminism and free action*, ed. T. Muller, 223–256. Berlin: Springer.

Tiomkin, M.L., and J.A. Makowsky. 1985. Propositional dynamic logic with local assignments. *Theoretical Computer Science* 36: 71–87.

van der Hoek, W., and M. Wooldridge. 2005. On the logic of cooperation and propositional control. *Artificial Intelligence* 164 (1–2): 81–119.

van der Meyden, R. 1996. The dynamic logic of permission. *Journal of Logic and Computation* 6 (3): 465–479.

von Kutschera, F. 1997. T × W completeness. *Journal of Philosophical Logic* 26 (3): 241–250.

Weinberger, O. 1998. *Alternative action theory. Simultaneously a critique of Georg Henrik von Wright's practical philosophy*. Berlin: Springer.

Wölf, S. 2002. Propositional Q-logic. *Journal of Philosophical Logic* 31: 387–414.

Wright, G.H.V. 1963. *Norm and action*. London: Routledge and Kegan.

Xu, M. 1998. Axioms for deliberative STIT. *Journal of Philosophical Logic* 27: 505–552.

Zanardo, A. 1996. Branching-time logic with quantification over branches: The point of view of modal logic. *Journal of Philosophical Logic* 67 (1): 143–166.

Of Norms

Jaap Hage

1 Terminology and Overview

Norms play a central role in practical reasoning, in law as well as in morality. An understanding of the nature of norms is therefore desirable for anyone who is theoretically engaged in practical reasoning, but such an understanding is not easy to achieve. The first challenge one encounters when trying to give an account of norms and their nature is that the terminology around norms is rich, to state it mildly. Von Wright starts his classic *Norm and Action* with the remark "The word 'norm' in English, and the corresponding word in other languages, is used in many senses and often with an unclear meaning" (Von Wright 1963, 1).

Ullmann-Margalit describes a social norm as "a prescribed guide for conduct or action which is generally complied with by the members of a society" (Ullmann-Margalit 1977, 12). However, she adds in a footnote that the term "norm" tends to be used by authors with a continental background, where authors with an Anglo-Saxon background prefer the terms "law" and "rule" (ibid.).

Kelsen writes at the very beginning of his posthumous study *Allgemeine Theorie der Normen* that the word "norm" denotes in the first place a command (*Gebot*), a prescription (*Vorschrift*), or an order (*Befehl*). He hastens to add, however, that ordering is not the only function of norms, but that empowering, allowing, and derogating are also functions of norms (Kelsen 1979, 1).

Although norms both play a role in law and in morality, the term "norm" has gained more popularity in the former field than in the latter. The term is also used more frequently in research from countries with a Roman law or Scandinavian background than in research from countries in the common law tradition. The reader may notice that these biases are reflected to some extent in the present contribution.

J. Hage (✉)
Faculty of Law, Maastricht University, Maastricht, The Netherlands
e-mail: jaap.hage@maastrichtuniversity.nl

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_5

Norms are both related to normativity and to rules. However, the notions of normativity and rules are as hard to pin down as the notion of a norm is. For example, the words “norm” and “rule” are sometimes used interchangeably for something that is normative. To clear out matters, this contribution will sharply distinguish between two oppositions: normative—non-normative and rule—description, and it will propose to use the term “norm” for normative rules only. In doing so, it must inevitably deviate to some extent from standard word use if there is something such as standard word use on these issues. The two distinctions will in the following be used as a framework for discussing not only norms as they are defined here, but also related phenomena that historically also have been called “norms.”

The main distinctions will be explained in Sects. 2 and 3 of this contribution: Sect. 2 deals with the nature of normativity, while Sect. 3 focuses on the nature of rules. Norms are often opposed to facts, because the former would be normative while the latter would not. It will be argued that the distinction normative—non-normative—is not the proper basis to distinguish between norms and facts and to that purpose Sects. 4 and 5, respectively, discuss different kinds of facts and more in particular deontic facts such as the existence of duties and obligations. Section 6 returns to rules and distinguishes three different kinds of rules. The distinction is then used to identify norms in the strict sense defined here and to discuss the related phenomenon of rules that confer competence and other forms of legal status. The contribution will be summarized in Sect. 7.

2 Normativity

Norms[1] are basically used for two purposes. The one purpose is to evaluate states of affairs and acts, and the other is to guide human behavior.[2] In this contribution, the emphasis will be on the second function, but in order to avoid possible confusions it is useful to say a little here about the evaluative function of norms.

2.1 Ought-to-Be and Ought-to-Do

It has become customary to distinguish between norms of the ought-to-do type and the ought-to-be type.[3] A norm of the ought-to-do type tells us what to do, while a norm of the ought-to-be type informs us what should ideally be the case, without

[1] For now, the term “norm” will be used in a broad sense. The more specific use will be introduced in Sect. 6.

[2] It is also possible that norms guide behavior of non-human entities, such as computer programs and robots, but here we will not pay special attention to these possibilities.

[3] The distinction is already quite old. Von Wright (1963, 14) mentions in a footnote the work of Max Scheler (1954) and Nicolaï Hartmann (1962). In the *Handbook of Deontic Logic and Normative Systems*, Hilpinen and McNamara (2013, 97) refer to Castañeda, H.-N. Castañeda (1972).

specifying that somebody should do something. An example of an ought-to-do norm is the norm that house owners should clear away the snow from the pavement before their houses. This norm specifies that something should be done, and also indicates who should do it. An example of an ought-to-be norm would be that letters ought to be stamped. This norm does not specify that some action ought to be undertaken, let alone who is responsible for undertaking this action.[4]

While ought-to-be norms indicate what should ideally be the case, there is no logical connection between what ought to be done and what is ideally the case. If there is to be a connection between what ought to be done and what is ideal, this connection must be created by some perfectionist theory of practical reasoning, such as utilitarianism.

Since ought-to-be norms do not specify what ought to be done, they have no use in guiding behavior; they can only be used to evaluate states of affairs as right (in accordance with the norm) or wrong (in violation of the norm). Norms of the ought-to-do type, on the contrary, can both be used to guide behavior and to evaluate it. The norm that house owners should clear away the snow from the pavement before their houses directs house owners to clear away snow. Looking backward, it can be used to evaluate the snow-clearing act of a house owner as right. Looking forward, it can be used in justifying the judgment that it would be wrong if the house owners in a particular street would not clear away the snow. Notice that although both ought-to-be and ought-to-do norms can be used to evaluate, the former will be used to evaluate states of affairs, while the latter will be used to evaluate behavior.[5] Ought-to-be and ought-to-do norms have in common that they underlie binary evaluations in terms of right and wrong, and not grading evaluations in terms of better and worse.

Because the emphasis of this contribution will be on norms that guide behavior and since only norms of the ought-to-do type can guide behavior, ought-to-be norms will be left out of consideration from here on.

2.2 Influencing and Guiding Behavior

One function of norms is to guide human behavior. To guide behavior is not the same thing as to influence behavior, although there is an important connection between the two. If Adrian influences the behavior of Bernadette, Adrian does something that exerts a causal influence on what Bernadette does. For instance, because the traffic is heavy, Adrian clutches his six-year-old daughter Bernadette to prevent her from crossing the street. In this way, Adrian influences his daughter's behavior by making it impossible. Bernadette has no choice whether she will cross the street.

[4] It should be noted that, in particular in law, formulations that suggest an ought-to-be norm because they do not specify that something ought to be done can nevertheless stand for ought-to-do norms, because it is clear from the context who is responsible for bringing about the right state of affairs.

[5] Attempts to define ought-to-do norms in terms of states of affairs that ought to be the case (see Hilpinen and McNamara 2013, 97–112 for an overview) are in the eyes of the present author a major source of problems in formal deontic logic. See Hage (2001).

Another way to withhold Bernadette from crossing would be to warn her. If Adrian warns his daughter not to cross the street, he leaves the choice to Bernadette, but tries to influence the choice that she will make. This influence is a causal relation between performing the speech act of warning and the motivation of Bernadette.

It would be quite similar if Adrian got frightened when he saw his daughter approaching the busy street and yells "stop" to her. Again, this is a speech act aimed at exerting a causal influence on Bernadette's motivation. The type of the speech act might be described as "giving an order," but it should be noted that this order should not be seen as the imposition of a duty on Bernadette not to cross, but *merely* as an attempt to causally influence Bernadette's behavior.

Let us assume that Adrian, being Bernadette's father, has some authority over his daughter and that he can impose duties on her. Suppose moreover that Adrian exercises this power and forbids Bernadette to cross the street. The *ultimate* purpose of this prohibition is to withhold Bernadette from crossing the street, and in this respect the speech act of prohibiting is similar to the issuing of a warning or a mere order. However, there is an important difference between on the one hand the prohibition and on the other hand the order (and the warning, for that matter). The order is merely an attempt to causally influence Bernadette's behavior, with the causal relation being between the performance of the speech act and the motivation to refrain from crossing the street. The prohibition is a way to impose a duty upon Bernadette, and this duty also exists if Bernadette is not motivated to comply and crosses the street nevertheless. Moreover, the relation between the speech act of prohibiting and its consequence, the existence of a duty, is conventional, not causal. The causal influence between the prohibition and Bernadette's behavior, if it exists, goes via Bernadette's knowledge that she has a duty not to cross the street to her being motivated not to cross. Duties themselves do not motivate, but the awareness of an existing duty may.

The existence of a duty not to cross the street is a reason that applies to Bernadette—and in that sense is a reason for Bernadette—not to cross.[6] Reasons directly guide behavior by telling what is the right thing to do, and one can indirectly guide behavior by creating reasons. In our example, Adrian would guide the behavior of Bernadette indirectly by prohibiting her to cross the street. In doing this, Adrian creates a duty and therewith a reason for Bernadette to refrain from crossing and it is this reason that directly guides Bernadette's behavior. Notice that this guidance by the reason is not a causal influence. It is still possible that Bernadette ignores the reason and is not at all motivated by it. If that happens, the existence of the reason does not causally influence Bernadette. Still the guidance exists; it consists in an indication of what is the right or the good thing to do, and if Bernadette does not act on the reason most likely, she did something wrong.[7]

[6] Reason terminology has become dominant in ethical theory. See, for instance, Williams (1981), Alvarez (2010), and Broome (2013). In legal theory, it still lacks the popularity it deserves, despite the efforts of Raz (1975), Hage (1997, 2005), and Bertea (2009) to promote it.

[7] This is not the place to discuss the different functions of reasons for acting, and the distinction between guiding and explanatory reasons. However, it is worthwhile to point out that even if guiding

2.3 Guidance by Norms: The Second-Person Point of View

Norms are not the same things as reasons for action, but they are closely related. Suppose that the norm exists that pedestrians are not allowed to cross the street if the traffic lights for pedestrians are red. In that case, the fact that the traffic lights are red is a reason for Adrian not to cross the street. The norm generates reasons in all cases to which it applies, and in that sense the norm guides behavior in a general way.

The presence of reasons for action is typically expressed by the use of "normative" words, such as "shall," "should," "must," "obliged," "obligated," "duty," "obligation," "forbidden," "prohibited," "permitted," "allowed," and "ought." Some of these words express a situation in which not only an agent should do something, but also somebody (else) is entitled to claim from the agent that he[8] acts in a particular way. The entitlement to such a claim, which can sometimes be enforced, is characteristic for moral and legal norms.

Norms are characterized by what Darwall called "the second-person standpoint." Darwall described this second-person standpoint as "the perspective you and I take up when we make and acknowledge claims on one another's conduct and will" (Darwall 2006, p. 3). This standpoint is characteristic for both legal and moral norms, but seems to be lacking for many prudential reasons. For example, if Bertie is thirsty, she has a reason to take a drink, but—barring exceptional circumstances—nobody, not even she herself, can claim from her that she takes a drink. Law, morality, and prudence all provide agents with reasons for action, but law and morality are normative in a sense in which prudence typically is not.

The normativity of norms does not only involve that norms indicate *what* should be done, but also that they can indicate *how* things should be done. Assume by way of example that there is a norm prescribing that one should eat asparagus with one's fingers, rather than with fork and knife. Of course, there is no duty to eat asparagus, but somebody who eats asparagus should do so with his fingers. If he uses fork and knife, he does something that is wrong. These "how-to norms" should be distinguished from the "technical norms" that specify how something should be done in order to succeed in bringing about a particular result. Somebody who does not eat asparagus with his fingers still succeeds in eating asparagus, but somebody who tries to make a last will without witnesses will normally not succeed in making last will. We will return to "how-to norms" in Sect. 5.4.

reasons do not need to exert a causal influence on the person to whom they apply, the very notion of a guiding reason would not make sense if people in general would not be motivated by the awareness that a guiding reason applied to them. See also Sect. 5.1.

Legal philosophers will recognize the parallel with the relation between a legal system's efficacy and the validity of the rules that belong to the system. Validity cannot be derived from efficacy, but it makes little sense to speak of the validity of norms that belong to a system that is completely inefficacious (Kelsen 1945, 42).

[8] This contribution adheres to the convention that authors should use the pronouns for their own gender to refer to persons whose gender is not important for the argument or otherwise determined by the text.

2.4 Norms and Facts

Perhaps this is the right moment to briefly address the distinction that is traditionally made between norms and facts, a distinction that is often traced back to the work of Hume. A typical use of norms, in particular ought-to-do norms, is to evaluate acts. For example, the norm that house owners are to clean away the snow before their houses can be used to evaluate the cleaning as right or correct. Norms can only fulfill this function if they are somehow different from the acts that they are used to evaluate. The fact that house owners tend to clear their pavements does not coincide with the duty for house owners to do so, and it does not even have to be evidence for the existence of such a duty. If this distinction between the norm and the behavior that does or does not conform to this norm is meant by the distinction between norm and fact, it is obvious that the distinction between fact and norm exists.

However, it sometimes seems that the distinction between norm and fact is considered to be much more prominent, as if it were a major ontological divide between what is "out there" independent of human beings and their minds, and what is added by human minds to what "really" exists. This major ontological difference does not exist, or—if it is assumed nevertheless (see the discussion of "objective facts" in Sect. 4.1)—it is of limited importance. The distinction between deontic (normative) and non-deontic (non-normative)[9] is real, but has no major ontological relevance.[10] It is comparable to the epistemic difference between what is certainly the case and what is the case. However, in Sect. 3, a different distinction will be made which does have major ontological relevance. This is the threefold distinction between facts, constraints on facts, and descriptions of facts. This distinction, which has nothing to do with the distinction between deontic and non-deontic, is highly relevant for a proper understanding of rules and therefore also of norms as deontic rules.

3 Rules as Soft Constraints on Possible Worlds

Often the notion of a rule is connected to the guidance of behavior: Rules would indicate what we should do. On this interpretation, the meanings of the terms "rule" and "norm" would practically coincide. And yet there are "rules" whose primary function does not seem to be to guide behavior and which can therefore only be "followed" in a broad interpretation of that term. Examples would be rules that

[9] From here on, we will follow the custom among logicians and use the terms "deontic" and "non-deontic" for the distinction between normative and non-normative.

[10] The importance that is attached to the distinction between is and ought may be explained from the function of critical morality, that is, to evaluate existing moral practices critically. The social practice of critically moralizing can only exist if the fact that some norms are actually used does not count as sufficient evidence for the claim that these norms should be used. Ignoring the difference between the norms, people actually use and the norms people should use make critically moralizing impossible. So, there is a practical relevance to distinguishing between is and ought, but this relevance does not justify that the real difference is blown up to an ontological gap.

confer competences and rules which make that something also count as something else.[11] A proper account of the nature of rules would explain why some rules can be complied with and other rules cannot. Such an account would also clarify the nature of norms as a special kind of rules. The first step in providing such an account is to go into some detail concerning "directions of fit."

3.1 Directions of Fit

Perhaps the best way to introduce the distinction between directions of fit is by means of an example of Anscombe (1976, 56). Suppose that Elisabeth makes a shopping list, which she uses in the supermarket to put items in her trolley. A detective follows her and makes a list of everything that she puts in her trolley. After Elisabeth is finished, the list of the detective will be identical to her shopping list. However, the lists had different functions. If Elisabeth uses the list correctly, she places exactly those items in her trolley that are indicated on the list. Her behavior is to be adapted to what is on her list. In the case of the detective, it is just the other way round; the list should reflect Elisabeth's shopping behavior. The two different functions of the list with regard to Elisabeth's behavior reflect the two different directions of fit that we are looking for.

The two items involved in Anscombe's example are a linguistic one, the list of items and the world. The directions of fit distinction can also be applied to other items than purely linguistic ones, but let us focus on the purely linguistic case first.

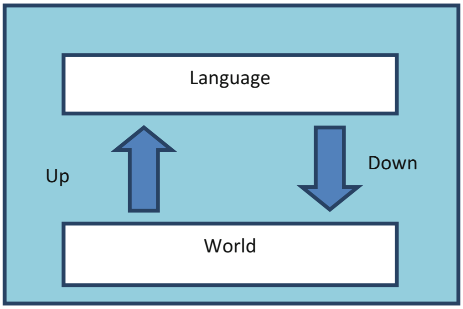
The relation between language and the world goes in two directions. If the linguistic entities are to be adapted to the world, as when the detective writes down which groceries are in the trolley, the fashionable expression is "word-to-world direction of fit" (Searle 1979, 1–30) . If the world is to be adapted to the linguistic entities, as when Elisabeth puts those items in her trolley that are mentioned in her shopping list, the fashionable expression is "world-to-word direction of fit."

However, expressive as these expressions "world-to-word direction of fit" and "word-to-world direction of fit" may be, they are also difficult to keep apart. Therefore, it is proposed to use the different expressions, "down" and "up" (see Fig. 1). The basic idea is that descriptive sentences consist of words that aim to fit the world. The propositions expressed by them are true, and the speech acts in which they are used are successful in the sense of "truthful," if and only if the facts in the world correspond to ("fit"), what these propositions express. This is the up direction of fit.

For the down direction of fit, we must distinguish between three kinds. For all three kinds holds that somehow the facts in the world are adapted, in order to "fit" what is expressed by the words. One case is when the words function as a *directive*, as

[11] It is possible to construct these rules as elements of more complicated rules that do guide behavior, and that is why it was written that it is not their *primary* function to guide behavior. However, it is difficult to disagree with Hart (2012, 35–42), who wrote that the construction of such rules as parts of mandatory rules would be a distortion. Still there is a sense in which, for example, power-conferring rules can be followed, and in Sect. 4.2 an example will be discussed.

Fig. 1 Directions of fit



when Adrian shouts "Bernadette, stop!" when he fears that Bernadette will cross the busy street. This order aims at making its addressee stop, and if the order is successful in the sense of "efficacious," Bernadette will stop and the facts in the world fit the content of the order. In this case, the relation between the utterance of the order (the performance of the speech act) and the facts in the world is causal by nature. We might therefore speak of the "causal down direction of fit."

A second case concerns constitutive speech acts, such as "I hereby forbid you to cross the street." If such a prohibition is successful, the facts in the world come to match the content of the speech act and Bernadette has from that moment on the duty not to cross the street. In this case, the relation between the performance of the speech act and the facts in the world is constitutive by nature; the performance of the speech act constitutes the duty. We might therefore speak of the "constitutive down direction of fit."

Notice that this down direction of fit relates a speech act to a duty, not to the compliance with the duty. Efficacy is here the coming about of the duty which the speech act aimed to create. The duty itself can also be efficacious in the sense that it is complied with, but that would be an example of the causal down direction of fit.[12]

The third kind of down direction of fit concerns the effects of "constraints." Constraints will be discussed more extensively in Sect. 3.3, but here we will use one kind of constraint as example: the conceptual rule (rule of meaning) that makes that the bachelors are unmarried man. Given this rule, if somebody is a bachelor, he must be unmarried. This "must" depends on the conceptual rule that defines the relation between being a bachelor and being married. Given this rule, it cannot be otherwise than that a person who happens to be a bachelor is also unmarried.[13] The facts in

[12] Seemingly, the causal down direction of fit can also exist between duties and behavior, and not merely between speech acts and behavior. However, that would mean that non-material entities such as duties can exert causal influences, which do not sit well with our ideas about the nature of causation. It is therefore more coherent (with our views of causality) to say that the causal down direction of fit can exist between the belief that one has a duty, as realized by a brain state, and behavior.

[13] It may be disputed whether the fact that bachelor is unmarried depends on a conceptual rule and whether this conceptual rule does not depend itself on some ontological constraint (ontological nominalism or ontological realism). For the present purposes, this does not matter, however.

the world adapt themselves to the constraint, and that is what is meant by the down direction of fit of constraints.

In Sect. 6.2, we will see that the constitutive down direction of fit is a special case of the down direction of fit of constraints and more in particular of dynamic rules.

3.2 Possible Worlds

We are all familiar with the distinction between what the facts actually are and what the facts might have been. The sun is shining, but it might just as well have been raining. In Western Europe, there is peace, but there might have been a war. In the common law, judge-made law plays a paramount role, but its role might have been subordinate to statute-based law.

Logicians use possible worlds' terminology to deal with this distinction between what the facts are and what they might have been. They say, for instance, that in the actual world the sun is shining, but that in some other possible world it is raining. Intuitively, a possible world is a set of facts which makes some descriptive sentences true and others false. The set of facts that defines a possible world is complete in the sense that it determines for every non-modal descriptive sentence[14] whether it is true or false. The actual world is one of the many worlds that are possible, and in the actual world the sun is shining. However, in some other possible world, it is raining. That is another way of saying that although actually the sun is shining, it might have been raining.

Some things are necessarily the case. For example, five is necessarily bigger than three. If something is necessarily the case, there is no possible world in which it is not the case. In all possible worlds, five is bigger than three. And in all possible worlds, if Janet is either in Berlin or in London, and she is not in London, then she is in Berlin. Being necessary may just as well be circumscribed as being the case in all possible worlds. What is necessary is the case in all possible worlds, while what is impossible is not the case in any possible world. What is contingent is the case in some, but not in all possible worlds.

What makes that a world is possible, and how can we distinguish possible worlds from impossible ones? To answer these questions, we need to apply the idea of constraints to possible worlds. Constraints on possible worlds are limitations on which facts can go together and which facts exclude each other. Possible worlds satisfy these constraints, while impossible worlds violate one or more of them.[15]

[14] For ease of exposition, we will ignore here exceptional descriptive sentences, such as the sentences "The king of France is bald" and "This sentence is false." The clause "non-modal" was added to take into account that modal sentences which express necessity may be interpreted as dealing with more than one possible world.

[15] The metaphysics of possible worlds is the central topic of an anthology edited by Loux (1979). To the present authors' knowledge, however, the idea that possible worlds are relativized to sets of constraints is not treated in that anthology, nor in more recent overviews of the discussions about possible worlds, such as Menzel (2015).

Examples of such constraints are for instance physical laws. A physically possible world is a world that satisfies all physical laws, including the law that metals expand when heated. Since all physically possible worlds satisfy this constraint, in all these worlds pieces of metal expand when heated. The same thing, stated in terms of facts that go together, is that in all physically possible worlds the facts that M is a piece of metal and that M is heated go together with the fact that M expands. So, if we only look at physically possible worlds, it is necessarily the case that a piece of metal will expand if it is heated. It is also the case that a piece of metal would have expanded if, counterfactually, it would have been heated.

3.3 Constraints

Necessity and possibility are not absolute phenomena. Something is always necessary or possible relative to some set of constraints.[16] If all constraints, including those of logic, are left out of consideration, everything is possible, even what is logically impossible. Not all constraints are physical or logical. There are also conceptual constraints, such as the constraint that bachelors are unmarried males, that a rectangle has straight corners, and—perhaps more controversial—that gold is a metal. Some constraints seem to defy any category, such as the constraints—if they are actual constraints—that all colored objects have a surface, that an item cannot simultaneously be at two different places, that every event has a cause, and that causality does not operate backward in time.

We can try to imagine a world that is not constrained in any way. In that world, all facts are independent of each other, as are the truth values of propositions that purport to describe these facts. The truth of one proposition has no connection at all to the truth of any other proposition, and the relations between the truth values of all propositions would be like the relations between the truth values of atomic propositions in propositional logic. Although it may be hard to imagine, the proposition “It is now five o’clock and it is raining” might be true, while at the same time the proposition “It is raining” would be false. A world in which this is the case is logically impossible, but it is still a possible world if the logical constraints are left out of consideration.

A logically possible world is a possible world in which logical constraints determine (not necessarily exclusively), which combinations of facts always hold, and which other combinations of facts never occur. Exactly which combinations of facts are necessary or impossible depends on the precise nature of the logical constraints. One such a constraint is that a fact and the “opposite” of this fact cannot go together. For instance, if the fact that it is raining obtains in a logically possible world, then the fact that it is not raining does not obtain in that world. Or—to say the same thing in term of truth values of propositions—a logically possible world does not make

[16] The theory about constraints as exposed here has some remarkable similarities to the theory of modalities defended by Frändberg (typescript).

both the propositions “It is raining” and “It Is not raining” true. More in general, a logically possible world does not allow that a proposition and its negation are both true. Nor does it allow that a proposition and its negation are both false. These are both constraints on logically possible worlds.[17]

Constraints may seem mysterious entities, and the question is justified in which manner they exist. An attempt to explain necessity and possibility in terms of constraints seems a bit like the explanation of the sleep-inducing nature of opium by pointing to the *vis dormitiva*, the sleep-inducing power, of opium in Molière’s play *The Imaginary Invalid*. That constraints somehow exist must be concluded from the fact that some things are necessary and others impossible. If we know that circles are necessarily round, we know something not only about actual circles, but also about the characteristics something would have if it were a circle, that is knowledge about possible circles. Knowledge about necessity is knowledge about hypothetical situations. This knowledge must be a priori (no dependent on sensory perception) and must be based on reasoning. The only feasible explanation that not only actual circles, but also possible circles are round is that the world is constrained in a manner that disallows non-round circles, and that this constraint also applies to the world that contains the possible circles about which we know that they must be round.

At first sight, it does not make much sense to search further for the nature of constraints and their mode of existence, but a little bit more can still be said. Take again the roundness of circles. This is often considered to be a conceptual truth. Unless one is a conceptual realist who assumes that concepts exist “out there,” to be discovered by intelligent beings, one can assume that concepts are being created by human beings. In particular with regard to artificial concepts, such as “computer,” it is plausible that they are human creations and could, at the time of their creation, be arbitrarily defined. It is still possible to modify the concept of a computer, to make it, for instance, include or exclude smartphones.[18] What a computer is, is a matter of convention, and the convention might have been slightly different from what it actually is. However, given the convention as it actually is, computers have some characteristics essentially and necessarily—for instance, having one or more processors—and other characteristics—for instance, their color—contingently and only possibly. The necessary characteristics of computers are based on a convention that functions as a constraint on what a computer can and cannot be.[19] Apparently at least some constraints are man-made, and rules belong to this category of man-made constraints.

[17] These constraints on logically possible worlds are typically represented in the semantics of logical theories by characteristics of the valuation function that assigns truth values to propositions. See, for instance, Navarro and Rodríguez (2014, 16).

[18] As a matter of fact, the concept of a “planet” has recently been redefined, taking away the status of a planet from the former planet Pluto. See https://en.wikipedia.org/wiki/IAU_definition_of_planet (last visited on December 24, 2015).

[19] This relation between conventions and the necessity based on them is explored a bit more in Hage (2013).

3.4 Rules as Soft Constraints

Why are rules a kind of constraints? Because they behave in many ways as other constraints. In the world in which a rule exists, the rule imposes itself on the facts of that world with the down direction of fit that other constraints also have. So, if some possible world contains the rule that thieves are punishable, then in this world thieves are punishable. Moreover, the rule also supports conditional and counterfactual judgments: If John had been a thief, he would have been punishable.

In a world that contains the rule that thieves are punishable, it is not merely a contingent matter of fact that thieves are punishable, but a necessary one, because being a thief makes one punishable. In this connection, something remarkable is the case. Rules allow for exceptions in the sense that sometimes the consequences of a rule do no hold, even though the conditions of the rule are satisfied. For instance, John, who is a thief, is also a minor and therefore the rule about the punishability of thieves cannot be applied to John. This possibility of exceptions seems hardly compatible with the necessary connection between being a thief and being punishable. And yet, the necessity and the exceptions have the same ground, which is that they are based on a constraint. Otherwise than descriptive sentences (see Sect. 4.6), constraints can have exceptions.[20] However, constraints also cover hypothetical and counterfactual situations, and that explains why judgments based on a constraint can express necessary relations such as the relation that thieves are necessarily punishable. Strangely, necessity and exceptions go hand in hand.

Rules have a lot in common with more traditional constraints such as the logical and physical ones, but there are also major differences. One such a difference is that rules only apply locally: The laws of one country are, for example, different from the laws of another country. The necessity of rule-based judgments seems therefore to be merely local necessity. This is different for logical and physical laws, which seem to have a universal scope of application.[21]

The scope of rules is not only limited in space, but also in time. Many rules can be created or derogated, and in that sense they differ from the more traditional constraints which somehow seem outside the scope of human manipulation. When the rule that thieves are punishable is introduced, suddenly all thieves become punishable. And when the rule is repealed again, the punishability of thieves disappears with the rule.

As a consequence of these differences, there can be some logically and physically possible worlds in which a particular rule exists, and other possible worlds in which the same rule does not exist. In a sense, it might be said that logical and physical constraints create necessities that are themselves necessary, while rules create

[20] Not only rules can have exceptions. There can also be exceptions to logical constraints (some descriptive sentences are not true or false) and to physical constraints (some physical laws are not applicable in extreme circumstances).

[21] This difference should not be overestimated, however. The geometrical law that the three corners of a triangle add up to 180° only holds for relatively small triangles and (which may be the same issue) for triangles in a flat plane. See also the discussion of the scope of physical laws in Toulmin (1953, 69 and 78).

contingent necessities. For this reason, rules will be categorized as “soft constraints,” as opposed to the hard constraints that do not depend for their existence on human decision making or social practices.[22]

4 Kinds of Facts

If the notion of a fact is taken broadly as that aspect of reality which makes a true descriptive sentence true, there are many different kinds of facts: facts that exist “objectively,” facts that depend on recognition, facts that are the results of rules or the use of reason, facts that are independent of all other facts, facts that “supervene” on other facts, “neutral” facts and facts involving evaluation, “inert” facts and facts that motivate or guide behavior. For a proper understanding of norms, the distinction between “inert” facts and facts that guide behavior may be the most important distinction between kinds of facts, but this distinction cannot be seen separate from many other kinds of distinctions between kinds of facts. Therefore, the present section and its subsections are devoted to a number of distinctions between kinds of facts. Their main purpose is to open up conceptual space for “deontic facts,” facts that involve that something should, or ought to be the case, of that somebody should or ought (not) do something. These deontic facts are crucial to understand norms.

4.1 Objective Facts

Some facts seem to be objective. They include the facts that Mount Everest is a mountain, that it is higher than 8000 m, that there are lions and other kinds of animals, and that there are *N* suns, with *N* being some as yet unknown natural number. The objectivity of these facts lies in their being mind-independent, that is independent of anyone’s beliefs, linguistic practices, conceptual schemes, and so on.

The idea that there are objective facts stems from a distinction that we make within our beliefs. To some beliefs, we ascribe a counterpart that somehow exists independently of what we humans think about them. This counterpart would consist of objective facts, and the facts mentioned above typically belong to this category. They are distinguished from other facts, which we take to depend on the human mind in some way. For example, many people take it that value judgments do not reflect an independently existing world of values, but rest on the way we humans evaluate things. No doubt evaluation is typically also based on objective characteristics of things, for instance the sharpness of the picture offered by the computer monitor, but these objective characteristics must be combined with a man-made standard to lead to the evaluation that this monitor is a good one. This dependence on a man-made

[22] This theme is elaborated in Hage (2015).

standard makes that the value judgment is not objective and that the fact expressed by it is not objective either. Objective facts are different, however. They are taken to obtain independently, and our knowledge, if it is objective, reflects these objective facts as they really are.

It may be argued that there are no objective facts in the sense of "objective" that is presently at stake. The reason is that every fact is the fact that …, where the dots are to be completed by some descriptive sentence. Facts depend on language, and since language is not mind-independent, facts are not mind-independent either, not even the "objective" ones. Many people would object against this conclusion because there must be something "out there" that precedes human categorization. That would be the "real" facts, and we humans try to develop concepts that fit this pre-linguistic substrate as well as possible. Whether such a pre-linguistic substrate really exists is an open question, but it certainly does not consist of the conceptualized reality in which we humans live. The assumption of a pre-linguistic substrate is the result of theorizing, not a precondition of it.

4.2 Brute Social Facts

When we are satisfied with a very coarse categorization, social facts may be described as facts which exist because the members of some group collectively recognize or accept them as existing. There are two variants on collective recognition. In the case of what we will call "brute social facts" the facts themselves are recognized by the members of some social group, while in the case of "rule-based facts" the facts are the result of some rule.[23] The facts based on rules that exist because of social recognition are perhaps better known as "institutional facts."

Brute social facts are the result of collective recognition. Important aspects of collective recognition are that sufficiently many and/or sufficiently important members of a social group believe the fact to be present, believe that the sufficiently many and/or sufficiently important other members also believe the fact to be present, and believe that these mutual beliefs constitute the believed fact.

Suppose that about 20 persons together make a foot trip to the top of a mountain. They believe that they are *as a group* walking to the mountain top, they believe that the others also believe that they are walking as a group to the top, and they all believe—minimally in the sense of not denying it when asked—that their mutual beliefs about acting together make that they are walking as a group to the mountain top, rather than as a set of individuals. In this case, the people in the group make a foot trip to the mountain top as a group. This is a brute social fact.

The above example deals with collective recognition of acting together, but collective recognition does not always deal with collective action. Suppose that one member of the group, say Henriette, utters strong opinions about which path to take

[23] The facts based on rules are perhaps better known as "institutional facts" (MacCormick and Weinberger 1986, 10).

to the top of the mountain and that most of the group members tend to act on these opinions. After having several times chosen a particular path because Henriette proposed to take it, most group members recognize the leading role of Henriette. They believe that Henriette has become the group leader that most other group members hold the same belief and that Henriette is the leader of the group because she is recognized as such by most group members. In this example, the brute social fact concerns the possession by Henriette of the status of leader of the group.

In the two above examples, recognition took the form of believing. However, sometimes mere believing does not suffice. If the leadership of Henriette in the group of mountain climbers has been sufficiently established, the group members may collectively recognize an order from Henriette as a reason for acting. Suppose that Henriette ordered Susan to walk on the back of the group to see to it that nobody stays behind. Then, Susan is considered to be obligated to walk on the back on the basis of collective recognition. This involves she is liable to be criticized by group members, including herself, if she does not walk on the back. Susan is obligated to walk on the back as a result of collective recognition, but Henriette's competence to create such a duty for Susan is also based on collective recognition. The group members collectively recognize an order from Henriette as creating a duty for the person who was ordered. They do this by collectively recognizing the duties that ensue from the orders. In combination, this amounts to the recognition of a power to create duties by giving orders. In its turn, the power actually exists by being recognized. This is an example of a fact—the existence of a power—that exists as the result of collective recognition, where the recognition does not consist in a belief, but in a complex set of dispositions to act.

4.3 Social Rules

If in our example about the mountain climbers the group members normally recognize the duties imposed by the leader of the group, whoever that may be, it may be said that the group has the rule that the group leader can impose duties. This rule exists through being recognized and therefore as a matter of social fact. The difference between having this rule and the recognition of the power of Henriette is the abstraction from the actual person having the power. When a power is not anymore ascribed to a particular person, but to a role—in this case the role of group leader—the acceptance of an ordinary social fact has become the acceptance of a social rule.

It is tempting to follow Hart (2012, 57) in assuming that the existence of a social rule involves the existence of a critical reflective attitude with regard to behavior covered by the rule. This characterization of social rules is quite adequate for rules that prescribe behavior, but less so for other kinds of rules such as power-conferring rules. A broader, and therefore more adequate, characterization of a social rule is that *a social rule exists within a group if sufficiently many (sufficiently important) members of the group recognize the consequences of the rule when the rule is applicable*. For a mandatory rule, this means that sufficiently many group members assume the presence of a duty or obligation if the rule attaches this duty or obligation to an actual fact situation. If the duty or obligation applies to a specific group member, this recognition typically involves that the group member is motivated to comply with the rule. For a power-conferring rule, this means that sufficiently many group members recognize the power of a person to whom the rule conferred the power. This recognition typically consists in the recognition of the effects of the exercise of the power. In our example, this was illustrated by the group members recognizing the duty that Henriette imposed on Susan.

4.4 Rule-Based Facts

If the group of mountain climbers has the rule that its leader has the power to create duties for group members, the power of Henriette to impose the duty on Susan to walk on the back is an example of the application of this rule. Henriette's power exists because of the rule and does normally not require separate recognition of Susan's duty by the other group members. Because the group has this rule about the powers of its leader, Susan has the duty to walk behind as soon as Henriette has imposed that duty on her.

The fact that Susan has this duty exemplifies a rule-based fact. Rule-based facts are those facts which exist because they were attached by a rule to some other fact, including the occurrence of some event.[24] Law provides telling examples of rule-based facts. Suppose that the parents of Joan own the Blackacre Ranch. When they die, Joan inherits the Blackacre Ranch and becomes owner of the ranch, at the moment that her parents die, even though it may still take some time before people, including Joan herself, receive the information that this is the case and before people are in a position to recognize that Joan has become the owner.

When Joan becomes the owner of the ranch, she also becomes competent to mortgage the ranch and to transfer the ownership of the ranch to somebody else. Most likely, the ownership of the ranch also brings for Joan the duty to pay real estate taxes. All these facts obtain solely because of the application of rules to the existing facts. Rules can attach consequences to facts and to events, and these consequences are new facts, rule-based facts.

One kind of rule-based fact is the existence of another rule. Again, this phenomenon is particularly important in law where most rules exist because they were explicitly created. That the rule-creating events actually lead to new rules is because other rules attach the consequence that a rule exists to these events. A properly created rule immediately exists, even if its consequences do not receive any recognition yet. However, if the recognition of the rule consequences never occurs, or—in other words—if the rule is completely inefficacious, the rule stops existing. If the term "valid" is used for the existence of rules, this means that rules that belong to a legal

[24] In Sect. 4.6, these rule-based facts will be called "immediate rule-based facts," and they will be distinguished from "mediated rule-based facts."

system are valid if the system as a whole is efficacious; efficacy of the individual rule is in first instance not required. However, if an individual rule is or becomes inefficacious for a longer stretch of time (*desuetudo*), this may take away the rule's validity (Kelsen 1960, 215–219).

Also outside the law, rule-based facts play an important role. It becomes easier to recognize this when one sees that standards at the hand of which value judgments are given are also a kind of rules.[25] Suppose that a group uses the standard that a soccer match is good if the play is aggressive but not foul. Then, if some match has aggressive but not foul play, this match is good. This situation is not very different from that of the group that recognizes that Susan has the duty to walk on behind because the group leader said so. The fact that Susan has this duty is just as "real" as the fact that the soccer match is good. Obviously, the existence of evaluative facts depends on a presupposed standard, but this holds for all rule-based facts.

Still another example of rule-based facts is the facts expressed by the theorems of some branch of mathematics, systems of formal logic included. The theorems express facts, they are true propositions, but they derive their truth from the axioms and the semantic rules (the "valuation function") of the formal system to which they belong.

4.5 Creation and Derogation

Because the existence of rules is often rule-based facts, there is a risk that rules are confused with the events by means of which they were created. Legislation is then, for instance, seen as a collection of rules, rather than as a means to create rules.

A similar misunderstanding underlies the view that legal norms are a kind of commands.[26] Such a characterization of norms would be wrong because a command is a speech act and therefore a kind of event, while a norm is a rule and therefore not an event. The temptation to see norms as a kind of commands may be explained from the shared normativity of commands and norms; it almost seems as if the norms command to act in a particular way. However, a proper understanding of the mode of existence of norms should focus on norms being a kind of rules, rather than on the specific kind of rules that norms are.

Basically, the same kind of mistake is made if the existence of norms is somehow connected to a legislator.[27] Many norms have been created intentionally by some

[25] An important difference between these standards and, for instance, legal rules is that legal rules also generate exclusionary reasons (Raz 1975; Schauer 1991; Hage 1997), while evaluative standards typically do not. This difference has no fundamental consequences for the role of evaluative standards as underlying rule-based facts, however.

[26] Famously, Austin defined laws in his first lecture in *The Province of Justice Determined* (Austin 1954, 24) as commands which oblige persons generally to acts or forbearances of a class.

[27] Von Wright makes this mistake for a particular category of norms, the laws of the state. He calls such norms "prescriptions" and defines prescriptions as having their source in the will of a norm-authority (Von Wright 1963, 7). A similar mistake seems to be made by Alchourrón and Bulygin

authority, but being a norm, not even being a legal norm, is not the same as being created by a state authority or any other kind of authority.

The speech acts by means of which some norms are created should not be identified with the norms created by them or with any other kind of norm. *Mutatis mutandis* this also holds for speech acts by means of which norms are derogated or repealed. By passing a bill, a legislator can derogate existing norms, but that does not make the bill into a norm. The idea that there are derogating norms therefore rests on a mistake.[28]

4.6 Factual and Descriptive Counterparts of Rules

The rule that thieves are punishable makes it impossible that thieves are not punishable.[29] Or, to state the same thing affirmatively, the rule necessitates that thieves are punishable. The (existence of) the rule that car drivers must drive on the right makes that car drivers have the duty to drive on the right. And the rule that cars count as vehicles for the Traffic Law makes that cars are (count as) vehicles.

If some rule—or, more in general, a constraint—exists, this means that some general descriptive sentence will be true. This sentence has more or less the same formulation as the rule, but it is not the rule formulation but a sentence that aims to provide information about the facts. Since these facts obtain because of the constraint, this sentence will be true.

Such sentences are open generalizations. An open generalization is a generalization over potentially infinitely many items. Examples would be that pedestrians wear shoes (a false open generalization) and that atoms have a nucleus (a true open generalization). Examples of closed generalizations are that all desks in this classroom are brown and that all instances of the Olympic Games lasted less than four months. Open generalizations can have counter-instances and still be true. For example, the open generalization that birds can fly is true, notwithstanding the existence of ostriches. A counter-instance to a closed generalization falsifies this generalization.

A rule of thumb to distinguish open from closed generalizations is that a closed generalization requires the use of the word "all," or some equivalent, while this word can be left away in case of open generalizations. For example, the sentence "Desks in this classroom are brown" expresses an open generalization which is almost certainly false, even if all desks in the classroom happen to be brown. ("Happen" is another word that indicates a closed generalization.) The open generalization requires for its truth a law-like connection (a constraint) between being a desk in the classroom and

(1981), when they recognize an "expressive conception of norms," according to which norms are essentially commands.

[28] This mistake was made by Kelsen when he allowed the possibility of derogating norms (Kelsen 1960, 57, 1979, 1).

[29] Remember that this necessity is compatible with exceptions to rules. See Sect. 3.4.

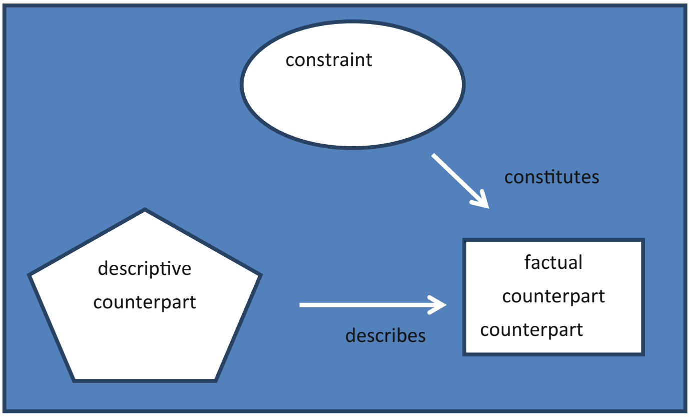


Fig. 2 Constraints, factual, and descriptive counterparts

being brown. Interestingly, it is precisely this law-like connection which makes that open generalizations can have counter-instances and be still true.[30]

The open generalizations that describe the effects of rules typically have the same formulation as the rule the effects of which they describe, and they are true because that rule exists. They describe facts that will be called the "factual counterpart" of the rule, and they may themselves be called the "descriptive counterparts" of rules. These descriptive counterparts of rules describe facts that are based on the existence of rules, but mediated by the rule-based facts that are immediately based on the rules.

Where rules impose themselves on the world by way of their down direction of fit, but are not true or false, the descriptive counterparts of rules are descriptive sentences, which are true or false, usually depending on the existence of the rules of which they are the counterpart. In schema (Fig. 2).

4.7 Norm-Propositions

The descriptive counterparts of norms have some resemblance with what are in the literature on deontic logic sometimes called "norm-propositions." Von Wright used the term "norm-proposition" for a proposition stating that a particular norm exists. Apart from the observation that Von Wright apparently saw norms as a kind

[30] See also Sect. 3.4. Because of the way open generalizations are often represented in formal logic, they have also become known under the misnomer "defeasible conditionals." Generalizations are not conditional sentences, even though they tend to be represented in predicate logic by means of conditionals. Moreover, open generalizations are true or false and not defeasible—but conclusions based on them may be defeasible (Hage 2005, 14). However, the truth conditions of open generalizations differ from those of closed generalizations, because the former are not necessarily falsified by counterexamples, while the latter are.

of entities that can exist, that is as a kind of logical individuals (see Sect. 6.1), these "norm-propositions" are not at all similar to descriptive counterparts of rules. Where norm-propositions in the sense of Von Wright talk about norms, descriptive counterparts talk about the subjects of norms (rules). For example, the descriptive counterparts of the norm "No vehicles in the park" are about vehicles, not about a norm.

Alchourrón and Bulygin (1981) and later Navarro and Rodríguez (2014) seem to follow Von Wright in calling statements about (the existence of) norms "norm-propositions," but they add that norm-propositions are statements about what is mandatory, prohibited or permitted relative to some set of norms. Since these statements are statements about prescribed, prohibited, or permitted states of affairs or actions, they are not statements about norms, so this identification seems to be based on a confusion. However, they are correct in pointing out that there are descriptive sentences, made true by existing norms (better: rules), stating that particular kinds of actions have a particular deontic status. Such statements describe if they deal with individual acts or with acts to be performed by a specific agent, rule-based deontic facts. Examples are the descriptive sentences "This killing was permitted" (based on the license to kill for secret agents) and "John must clear away the snow from the pavement before his house" (based on the rule prescribing house owners to clean away the snow before their houses). If the truth of these descriptive sentences is based on legal rules, these sentences would also express what Kelsen called "Rechtssätze" (Kelsen 1960, 57).

If such statements describe general prescriptions, prohibitions, or permissions, they are descriptive counterparts of rules. Examples would be "House-owners must clear away the snow from the pavement before his house" and "Secret agents are licensed to kill in the performance of her majesty's secret service." Kelsen called these general descriptive sentences "Rechtssätze" too (Kelsen 1960, 85), but also "legal rules" (Kelsen 1945, 45).

4.8 "Entailed" Norms

The recognition of descriptive counterparts of rules is important, because they are "ordinary" descriptive sentences to which deductive logic is applicable, whereas the applicability of deductive logic to rules, and to norms as a species of rules, is dubious since rules are from a logical point of view individuals (see Sect. 6.1). Many inferences which seem to have rules as their conclusions may well be interpreted as arguments with descriptive counterparts of rules as their conclusions. For example, the argument "Volkswagens count as cars. Cars owners must pay road tax. Therefore owners of Volkswagens must pay road tax" is dubious as an argument in which a rule is derived from two other rules. As an argument in which two descriptive counterparts of rules are used to derive another open generalization, it is valid.[31]

The possibility to derive open generalizations from other open generalizations also explains the phenomenon of "deontic inheritance" (Hage 2001), "entailed norms" (Navarro and Rodríguez 2015), or "normative consequences" (Araszkiewicz and Pleszka 2015). An example would be that a prohibition for vehicles in the part would entail a prohibition for Volkswagens in the park. It is highly dubitable whether these entailed "norms" are rules that can be traced back to some official legal source. However, it is obvious that if vehicles are prohibited in the park, then—normally speaking—Volkswagens will be prohibited as special case of this general prohibition. The one deontic fact—vehicles being prohibited—encompasses the other—Volkswagens being prohibited—and there is no objection against deriving the proposition that expresses the latter fact from the proposition expressing the former fact. However, interpreting such a derivation as the derivation of one norm from, among others, another norm would be misguided (Hansen 2013).

5 Deontic Facts

5.1 Deontic Facts and Motivation

Norms are normative because they lead to duties or obligations.[32] Duties and obligations are entities that exist in time but not in space. They can be created and destroyed, and they can have all kinds of characteristics such as being a nuisance or being suitable to deal with societal problems.

The existence of a duty or an obligation is a deontic fact. It is a fact about an immaterial "thing," a duty, or an obligation, and it is deontic (normative) in the sense that it guides behavior. Suppose that Susan and Thera have concluded a labor contract. After that event, Susan has an obligation toward Thera to pay her a monthly salary, while Thera has an obligation toward Susan to work the afternoons of all weekdays. The sentences describing the existence of these obligations are true just like any other descriptive sentences. Facts that involve the existence of a duty or an obligation, and also some other kinds of facts, including the existence of permissions, may be called "deontic facts," after the convention that has arisen in logic to call logics that deal with duties, and obligations and with everything else that ought to be done "deontic logic."

[31] If the word "rule" is also used to denote open generalizations, there is no problem in deriving rules *in this sense* from other rules, *also in the sense of open generalizations*, and facts. It is this kind of reasoning about "rules" that seems to be at stake when MacCormick (1978, 100–108) writes about second-order justification of rules.

[32] The difference between duties and obligations as it is made here will be discussed in Sect. 5.2.

Sometimes the duties and obligations themselves, or their contents, are called "norms." As explained in Sect. 1, we adopted a different terminology here.

Even though the existence of a duty or an obligation is a fact, it is also deontic, normative, or behavior guiding, whatever you may want to call it. Although this is not the place to go into details with regard to the nature of normativity (however, see Sect. 2), it may nevertheless be useful to say a little about why duties (and obligations) are normative.[33] The normativity of duties lies in the connection, however remote that may sometimes be, between the existence of a duty and the motivation of persons to act in a particular way.[34] Typically, the acting person is the holder of the duty, and the behavior at issue is compliance with the duty. Agents tend to have a disposition to comply with their duties.[35] Although it is possible for an agent to recognize that he has a duty without being motivated at all, it is not possible that agents typically would not be motivated when they recognized to have duties. The reason is that if duty holders would not normally be motivated to comply with their duties, the very notion of a duty would not make sense.

The existence of a duty is not only based on the behavior or the disposition thereto of the duty holder; the behavior of agents in the environment of the duty holder is relevant too. This behavior consists of praise—in case the duty was complied with—or blame—in case the duty was violated, or—to use Hart's phrase—of a critical reflective attitude.

Sometimes the connection between a duty and the motivation to act is indirect. That is for instance the case with duties based on rules which exist themselves as a matter of rule-based fact, such as legal duties. Then, the disposition to comply with duties is the disposition of the addressees of the normative system as a whole to comply with duties based on the normative system. It is not the case anymore that every duty based on the system must lead to a disposition for compliance. Moreover, the efficacy of the system as a whole—because that is what we are talking about—-must consist in recognition of the consequences of the system's rules, and since the rules are not always mandatory, the required efficacy is not always compliance with duties. It can, for instance, also be recognition of the power to make rules.

The connection between a deontic fact such as the existence of a duty and behavior may be quite complicated, but first, normativity cannot exist without such a connection, and second, there is nothing more to normativity than this connection. There is, for example, no such a thing as "binding force" apart from the disposition to motivate. That means that there is, for instance, no need to postulate the existence of a "norm"

[33] The following paragraphs only discuss duties, but *mutatis mutandis* the argument also applies to obligations.

[34] Sartor (2005, 454) seems to express the same idea when he characterizes obligations in terms of the intention to act on them.

[35] Human agents often critically evaluate the "duties" that they have according to a particular normative system, such as positive morality or positive law, and sometimes this evaluation leads to the conclusion that they should not comply with some "duty." However, such a refusal to comply with a "duty" which is not up to standard is often motivated by saying that the "duty" turned out not to be a "real" duty after all. In that case, the link between real duties and the motivation to act upon them remains intact. Obviously, much more can and needs to be said on this issue, but this is not the place to do so. Interested readers are referred to Hage (2013).

next to the rule-based facts that persons have certain duties and the constitutive rules on which these facts are based.[36]

5.2 *Duties and Obligations*

In normative discussions, words like “ought,” “should,” “must,” “duty,” and “obligation” are often used interchangeably. Although the meanings of these words in natural language are not fixed and overlapping, there exist different categories of deontic facts. The differences between these categories are important, although the words used to denote them are not. In the following, we will consider some distinctions and adopt particular words to denote the newly delineated categories. Let us start with some examples:

- Everybody has the duty not to steal, but normally there is no obligation to that effect.[37]
- *A* and *B* are under obligations toward each other, because they entered into a sales contract, but these obligations are not duties.
- From the fact that *P* is under an obligation or a duty to do something, it follows, *pro tanto*—that is, if only this reason is taken into account—that *P* ought to do it, but not the other way round.

The first distinction to be made is between duties and obligations. The existence of both a duty and of an obligation is a reason why somebody ought to do something, but neither the duty nor the obligation coincides with the fact that this person ought to do it. *A* duty is often connected to a role or status. It is, for instance, the duty of house owners to pay real estate tax and the duty of a mayor to maintain the public order in a municipality. All human beings[38] are under a duty not to kill other human beings. However, as our example of Bernadette and Adrian (Sect. 2.2) illustrated, it is possible to have a duty as the result of a command, and such a duty is not connected to a particular role or status.

Whereas duties are often connected to a particular status or role, an obligation is the outcome of an event and depends on that event having occurred. Typical examples of such obligation generating events are causing damage, making a promise, or contracting. Moreover, whereas a duty is not a duty with regard to

[36] Such a postulation seems to be made in the account that Navarro and Rodríguez give of the relation between norms and normative propositions (Navarro and Rodríguez 2014, 78).

The constitutive nature of the rules on which deontic facts are based is discussed in Sect. 6.5 of the present contribution.

[37] The term “obligation” derives the technical meaning that is proposed here from the civil law tradition, according to which an obligation is a particular kind of bond between a debtor and a creditor (for the historical roots of this word use, see Zimmermann 1996, 1). In the English literature, the difference between duties and obligations is not drawn sharply, possibly under the influence of the common law.

[38] Being a human being might be the most abstract status to which duties are assigned.

somebody in particular, obligations are always "directed," obligations toward somebody else.[39] This directedness of obligations still holds if this "somebody else" is (as yet) unknown, as when, for instance, a car was unlawfully damaged but the owner of the car is still unknown. Duties are not directed in this way.[40]

5.3 Being Obligated and Owing to Do Something

The term "obligated" will be used here as a term of art to denote the common denominator of duties and obligations: A person who is under a duty to do *B* is obligated to do *B*; a person who is under an obligation to do *B* is also obligated to do *B*. Being obligated is not directed. If *A* has contracted with *B* to pay him €100, then *A* has an obligation toward *B* to pay him €100, and *A* is also obligated to pay *B* €100, but *A* is not obligated *toward B* to pay *B* €100.

By now, we have encountered three normative concepts, "duty," "obligation," and "being obligated." They all differ from the normative concept that is often used as a catchall for all kinds of normativity, the concept of "ought." In connection with duties, obligations, and being obligated, the more relevant notion is ought-to-do. The word "ought" as defined here stands for the outcome of the interplay of one or more reasons for acting, a kind of aggregate of these reasons. Examples are the legal ought, as the aggregate of legal reasons for action, and the moral ought as the result of the aggregate of moral reasons.

An ought itself is not a reason for acting, but merely the outcome of one or more reasons. So, where the fact that *X* is under a duty to pay real estate tax is a reason why *X* ought to pay real estate tax, the fact that *X* ought to pay the tax is not a reason for paying it, although it *presupposes* the existence of such a reason (the duty, for example). An ought is comparable to being obligated in the sense that it abstracts from the precise reasons for acting, but nevertheless indicates (through presupposition) that such reasons exist. Where being obligated is tied to precisely one such a reason, owing to do something is based on a set of reasons, even though this set may contain one reason only. Being obligated can therefore be seen as a *pro tanto* ought.[41]

The difference between, for instance, an obligation and the ought based on it becomes clear if one considers what happens in case it is impossible to perform one's obligation. For instance, if Antony contracts with Giovanni to transfer his car to Giovanni, and if he also contracts with Guido to transfer his car to him, then

[39] For a logical discussion of these "directed obligations," see Herrestad and Krogh 1995.

[40] An example of a duty without a person toward whom the duty exists is the duty to stop for a traffic light, even if nobody is approaching. However, even if a duty mentions persons, e.g., the duty not to kill prisoners of war, this is not a duty toward these persons. Other persons can also address the duty holder about compliance with the duty. This is different for obligations, where typically only the right holders can demand compliance.

[41] The notion "prima facie ought" is more fashionable, but is strictly speaking an epistemic notion: If *A* prima facie ought to do *X*, then *for all we know A* ought to do *X*.

Antony both has an obligation toward Giovanni and toward Guido. It is impossible for Antony to comply with both obligations, and therefore[42] it is not the case that Antony both ought to transfer the car to Giovanni and to Guido. The law has a simple solution for such cases. Both obligations have an equal status (*paritas creditorum*). If Antony complies with his obligation to Giovanni, he must default on his obligation toward Guido, and—because the obligation still exists—Antony must compensate the damage of Guido. The question which obligation supersedes the other has a clear answer: neither one of the obligations *as such* supersedes. However, in determining what Antony ought to do, the reasons for acting have to be balanced. If the above account of the legal situation is correct, the outcome will be that Antony is legally permitted to deliver the car to any one of his creditors[43] and that he will have to financially compensate the other creditor.[44]

5.4 Permissions

Traditionally, permissions have been treated as the opposite of prohibitions. For example, if *P* is forbidden to do *A*, this means (is the same as) that *P* is not permitted to do *A*. However, it has turned out that the relation between on the one hand permissions and on the other hand the other deontic notions, such as "ought," "obligated," "duty," and "obligation," is not straightforward (Hansson 2013). Characteristic in this connection is the distinction, popularized by Von Wright (1963, 85–87), between weak and strong permissions. A weak permission would be nothing else than the absence of a prohibition, while a strong permission would involve a prohibition to interfere with an agent's freedom in a certain respect.

We will have a brief look at several possible interpretations of permission, and start with the possibility that act tokens, acts that have already been performed, were

[42] Although the principle "ought implies can" is in the eyes of the author not a logical constraint, there is from the moral and the legal point of view much to be said for it. In the law of obligations, for instance, impossibility is the main reason for assuming *force majeure*. That is why the principle is applied in the present argument.

[43] Whether Antony is also permitted not to deliver the car to anyone of his creditors depends on the legal system. In the common law, where "specific performance" is exception rather than the rule, Antony would be permitted to financially compensate both creditors rather than delivering the car to any one of them. In the civil law tradition, Antony would still be obligated to deliver the car to the creditor he does not compensate financially (Smits 2014, 194–202). This example illustrates in the first place that the relation between the existence of an obligation and what a debtor legally ought to do depends on the law, not on logic alone, and in the second place that it is useful to study the law from a comparative perspective to see the respective roles of law and logic. Where legal solutions differ, they cannot be a matter of logic.

[44] This account may not be correct for every legal system. In some systems, obligations to transfer a good do have a priority, with the older obligation superseding the more recent one. In those systems, the debtor legally ought to transfer the object to the oldest creditor. Also, this example illustrates that the relation between legal obligations and what an agent legally ought to do are in the first place governed by law.

permitted. In this brief discussion, the agents performing actions will mostly be left out of consideration and the talk will be about action types or act tokens that are, or are not, permitted. The evaluation of act tokens may be undertaken from several points of view, such as the legal and the moral point of view. Here, we confine ourselves to the legal point of view.

What does it mean that a particular act *token* was legally permitted? Basically, it means that this act, as performed by this agent, does not belong to a legally prohibited action type. Either there was no legal norm prohibiting this type of action, or in the concrete case there was an exception to the norm. Suppose that Ellen takes a break by making a walk on the lawn. Taking a bread was allowed, but walking on the lawn was not. Therefore, what Ellen did was not permitted, because her act can be subsumed under at least one prohibited action type. However, if Ellen would have received a special permission to walk on the lawn, her act was permitted, because the granted permission makes an exception to the general prohibition to walk on the lawn.

An action *type* is permitted if there is no norm which directly or indirectly prohibits that type of action. We will return to this distinction between direct and indirect prohibition soon, but first it is necessary to say something about default deontic status. Most of us live in a society where everything that has not been prohibited is permitted. Being permitted is the default deontic status of all action types, and it requires a prohibitive norm to change that status for a particular kind of action. Things might have been different, however. Logically, it would have been possible that everything that has not been permitted is prohibited. The reader should keep this in mind and be prepared to turn the following account of prohibited and forbidden action types around for the theoretical case that a society would have prohibition as its default deontic status.

An action type is directly prohibited if and only if there is a norm that explicitly prohibits that type of action. So, if the norm exists that prohibits lying, the action type lying is directly prohibited. If this norm does not exist, it might still be possible that there is a norm that explicitly forbids cheating. Then, cheating is directly prohibited. Suppose now, for the sake of argument, that lying involves cheating.[45] Then, barring exceptions, every instance of lying is also an instance of cheating. In that case, the direct prohibition of cheating makes that lying, the action type, is indirectly forbidden by the norm that prohibits cheating. Using this distinction between directly and indirectly forbidden action types, we can say that an action type is permitted if this type is neither directly nor indirectly prohibited.

Thus far we discussed permissions in the sense of absence of prohibition. An act token was permitted if it could not be classified as belonging to a prohibited type. An action type is permitted if there is no norm that directly forbids this type of action and if the performance of an act of this type does not involve the performance of a prohibited action. However, as was already pointed out by Von Wight, some

[45] The precise nature of this involves-relation which may hold between action types is crucially important in this connection. *A* first approximation would be that action type *A1* involves action type *A2* if necessarily every token of *A1* is also a token of *A2* (Hage 2001). It should be noted in this connection that the approximation presupposes that one act token can belong to more than one action type, and that the constraints that determine what counts as necessary remain unspecified.

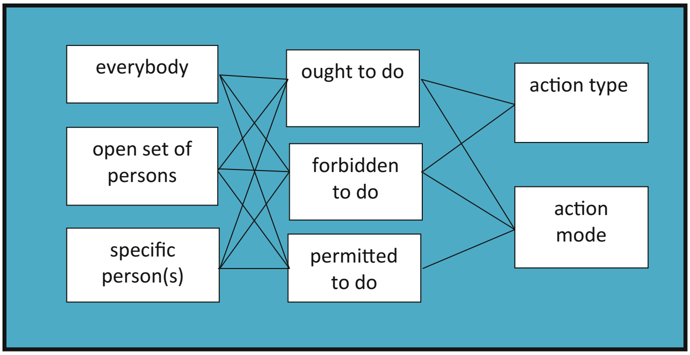


Fig. 3 Anatomy of ought-to-do

action types are explicitly permitted by a permissive norm. Such a permissive norm is typically—but not logically necessarily—connected to a freedom right. For example, the right to vote includes a permission to vote, and the freedom of expression includes a permission to utter one's opinions.[46] The permission that is included in some rights should not be confused with other elements that are also included in these rights. The idea, for instance, that strong permissions include Hohfeldian immunities—the legislator would, for instance, not have the power to forbid a citizen to vote—seems to be based on this mistake. In a right, a permission may be combined with an immunity, but this combination does not mean that the permission somehow includes the immunity. It is the right that includes both the permission and the immunity.

It is possible that an act token can be classified as belonging to two types, one type being explicitly permitted and the other type being forbidden. This would, for instance, be the case if it is forbidden to set people against each other, while it is permitted to express one's political opinions. Then, there is a norm conflict, which should be treated in the same way as other norm conflicts.

5.5 The Anatomy of Ought-to-Do

To understand the nature of deontic facts and of norms, it is useful to distinguish the elements of deontic facts. We will enumerate these elements for states of affairs of the ought-to-do type, but most of the discussion can mutatis mutandis be applied to duties and obligations as well (Fig. 3).[47]

[46] The inclusion of permissions—and also of competences—in rights is somewhat analogous to the involvement of one action type by another action type. An adequate theory about the nature of rights should include an elaboration of this includes-relation, but this is not the place to address this topic.

[47] The main difference is that whereas an ought and a duty contain three (or four) elements, an obligation is directed toward a creditor and therefore contains four (or five) elements, the extra element denoting the creditor.

An ought-to-do state of affairs involves that somebody is either permitted, required, or prohibited to do something, or to do something in a particular way, or at a particular time or place. An ought-to-do state of affairs consists of three or four elements, the deontic modality, the addressee, the act specification, and—occasionally—the specification of the act modality.

Take the following examples:

a. It is forbidden to murder.
b. Car drivers ought to carry a driver's license.
c. Leon is allowed to eat asparagus with his fingers.

In example a, everybody is an addressee; the modality is a prohibition, and the object of the deontic state of affairs is the performance of an action type (to murder).

In example b, the addressees are the members of the open class[48] of car drivers, the modality is an ought, and the object of this ought is the performance of an action type (carrying a driver's license).

In example c, the addressee is a single agent, the modality is a permission, and the object is an action mode (using one's fingers to eat asparagus).

It may be tempting to treat example b as expressing a conditional sentence: If somebody is a car driver, then he ought to carry a driver's license. This temptation is even strengthened if one considers how the deontic fact described in b can be used to argue why a particular car driver, say Lenny, ought to carry a driver's license. The following modus ponens style argument seems to do the job well: All car drivers ought to carry a driver's license; Lenny is a car driver; therefore, Lenny ought to carry a driver's license. Still, it seems a better idea to follow Von Wright (1963, 82) by distinguishing conditions under which a deontic fact obtains and the agents for which this deontic fact holds.

Notice, by the way, that this distinction presupposes that there are no conditional deontic facts, but only conditions for the existence of a deontic fact. If car drivers should place a warning sign before their cars if the car has broken down while it is dark, the deontic fact is that car drivers should put a warning sign before their cars, and this fact is present if the cars break down while it is dark. Of course, the sentence in which this relation between darkness and the duty to place a warning sign is expressed is itself conditional. However, that does not make the deontic fact conditional.

Only deontic facts where the addressee is a single agent (or a closed group of agents) concern the existence of a duty or obligation, and only these can be immediate rule-based facts. If it is a fact that car drivers ought to carry a driver's license, this is a fact only because every individual car driver, actual or merely hypothetical, ought to carry a driver's license. These individual oughts are most likely based on the rule (norm) that car drivers ought to carry a driver's license. A similar argument

[48] That the class is "open" means that the denoting expression refers to everybody who may happen to be a car driver and not merely to the fixed set of actual car drivers. The "openness" of the class makes that the deontic fact also deals with hypothetical car drivers, as in "If Thera would have been a car driver, she would have to carry a driver's license." See also the discussion of open generalizations in Sect. 4.6.

shows how it can be a fact that it is (for everybody) forbidden to murder. From this perspective, it can be seen that the sentences a and b are ambiguous: They may be read as rule formulations, but also as descriptions of deontic facts. In the latter case, they are the descriptive counterparts of the rules (see Sect. 4.6).

6 Of Norms and Other Rules

Given the facts that there is no fixed terminology concerning norms and that norms are closely related to rules and to normativity, it seems worthwhile to explore the idea that norms are rules—a kind of constraints—that lead to deontic facts. They lead in the first place to the existence of duties and obligations and—derived from these duties and obligations—to facts of the obligated—and ought type. We will explore that idea in this section, and to that purpose we start with distinguishing three kinds of rules, dynamic rules, fact-to-fact rules, and counts-as rules. Then, we consider the relevance of these kinds of rules for the constitution of deontic facts. This section is closed with some remarks on competence-conferring rules and other rules that confer status.

6.1 Rules as Individuals in the Logical Sense

Rule formulations such as "Thieves are liable to be punished" and "The Mayor of a municipality is competent to issue emergency regulations for that municipality" have, as far as their formulation is concerned, much in common with general statements. Moreover, rules can be used in rule-applying arguments which look like arguments of the modus ponens type. Nevertheless, there are important differences between rules and statements. Otherwise than statements, rules exist in time: They can be created and repealed (derogated; abolished). They can become outdated and can stop being used (*desuetudo*). Moreover, in contrast to statements which have the up direction of fit, they have the down direction of fit (of constraints). It is possible to predicate something of rules, such as in the sentence "This rule has been studied by legal historians for dozens of years." Rules can also be part of a relation, as can be stated in the sentence "The rule that thieves are liable to be punished exists longer than the rule that gives Mayors the competence to create emergency regulations."

Because of these latter reasons, there is much to be said for treating rules as a kind of things, rather than as statements describing what is the case. In the terminology of logic, such "things" are called individuals. If rules are from a logical point of view individuals, it is easy to see why rules as such cannot be parts of deductive logical derivations.[49] Deriving something from a rule would be comparable to deriving

[49] This does not exclude that rule-applying arguments are studied in logic. There are several ways to do so. One is to drop the demand that all elements of an argument are propositions. Second is

something from a chair. Norms do not figure in deductive arguments, but the reason is not that they are deontic or like imperatives (*pace* Jörgensen 1937/8), or that they have the down direction of fit of constraints,[50] but that they are from the logical perspective individuals.

6.2 Dynamic Rules

All rules connect facts to each other. These facts may be simultaneous, or they may succeed each other in time. The latter is the case with dynamic rules: They create new facts, or modify or take away existing facts as the consequence of the occurrence of an event.

Examples of events to which rules attach consequences are that John promised Richard to give him €100 and that Eloise was appointed as chair of the French Parliament. John's promise has the consequence that from the moment of the promise on John has the moral obligation to pay Richard €100. The appointment of Eloise as chair has as its legal consequence that from the starting point of the chair's new term on, Eloise will be the chair of the French Parliament. Other examples in which a dynamic rule attaches legal consequences to an event are that a Bill was passed, with as consequence the existence of new rules, or that Lionel committed theft, with as legal consequence that Lionel is liable to be punished.

Like all rules, dynamic rules have an element of generality. They apply to events of a particular kind and attach to these events facts of a particular kind. Dynamic rules may be conditional, in which case their consequence is only attached to the event under certain conditions. An example is the rule that if it is dark, the occurrence of a car accident obligates the drivers to place a sign on the road before to the cars.

If a juridical act or some other constitutive speech act is performed and a dynamic rule attaches consequences to this act, this is both a case of the constitutive down direction of fit and of the down direction of fit of constraints (see Sect. 3.1). The former focuses on the speech act by means of which the consequences were constituted; the latter focuses on the rule that attaches consequences to the performance of the act. The constitutive force of speech acts rests on the effects brought about by dynamic rules, and therefore the constitutive down direction of fit of constitutive acts is a special case of the more general down direction of fit of constraints and more in particular dynamic rules.

to allow entities without truth values as propositions. And third one is to use statements about the existence of rules as premises in rule-applying arguments. All of these options have consequences for the systems of logic that can be used to study rule-applying arguments that reach farther than accommodation for the defeasibility of rule-applying arguments.

[50] That the down direction of fit of constraints does not preclude them from being parts of arguments becomes clear from the example that if the world is constrained in such a way that Volkswagens are vehicles and that vehicles are not allowed in the park, the world is also constrained in the sense that Volkswagens are not allowed in the park. However, from the fact that the former two constraints exist as rules, it cannot be derived that the latter constrains also exist as a rule. See also Sect. 4.8.

6.3 Fact-to-Fact Rules

Where dynamic rules govern the succession of facts in time, static rules govern the coexistence of facts. One kind of static rules is fact-to-fact rules, rules which make that one kind of fact tend to go together with some other kind of fact, where the latter fact depends (supervenes) on the former. The relation between the kinds of facts is timeless, in the negative sense that the one kind of fact is not the occurrence of an event after which the second kind of fact comes into existence.[51]

A logical example of a fact-to-fact rule is that a conjunction is true if both conjuncts are true.

A moral example is the rule that spouses should be faithful to one another.

Legal examples of fact-to-fact rules are the rules that

1. The owner of a good is allowed to use this good.
2. The mayor of a municipality has the competence to issue emergency regulations for that municipality.
3. House owners must keep the pavement before their houses clean.
4. The king of Belgium is the commander in chief of the Belgian army.[52]

Characteristically, all the legal example rules attach consequences to the possession of a certain legal status. Important legal examples of fact-to-fact rules are rules that impose legal duties (example 3), rules that confer competences on people with a particular status (example 2), and rules that attach a specific status to the presence of some other status (example 4).

6.4 Counts-as Rules

The second kind of static rules that will be discussed here is *counts-as rules*. They have the structure: Individuals of type 1 count as individuals of type 2. These "individuals" may be human beings, as in the rule that the parents of a minor count as the minor's legal representatives. Often, however, the "individuals" that count as another kind of individual are events. For instance, under particular circumstances, causing a car accident counts as committing a tort, or offering money to another person counts as attempting to bribe an official.

Usually, counts-as rules are conditional, meaning that individuals of type 1 only count as individuals of type 2 if certain conditions are satisfied. An example from Dutch law (art. 3:84 of the Civil Code) would be the rule that the delivery of a good counts as the transfer of that good if the person who made the delivery was competent to transfer and if there was a valid title for the transfer.

[51] Notice that, this timeless relation between the conditions and the consequences of a fact-to-fact rule is compatible with the existence in time of the rule. Only as long as the rule exists, the condition facts and the conclusion facts go together in a timeless fashion.

[52] This last rule may also be interpreted as a counts-as rule.

Counts-as rules cannot create deontic facts by themselves. However, they often make that something counts as something to which a norm attaches deontic facts. Causing a car accident may count as a tort to which a rule of tort law attaches the obligation to pay damages. Being a person against whom serious objections exist counts as being a criminal suspect, a fact that gives police officers permission to arrest you.

6.5 Norms

To focus our discussion, we have defined norms as rules that constitute deontic facts. Let us focus our discussion even more, by assuming that there are only two kinds of basic deontic facts that matter for norms, that is duties, which are not directed toward a corresponding right holder, and obligations, which are directed in that way.

This stipulation contains two clauses that need justification. The first clause is that we confine ourselves here to *basic* deontic facts. That excludes facts of the types that somebody is obligated to do something or that somebody ought to do something. These latter facts are not basic, because they supervene on the existence of duties and obligations (see Sect. 5.2).

The second clause is that we confine ourselves to deontic facts that matter for norms. This has to do with the second-person perspective of norms (see Sect. 2.3): Norms justify claims for compliance. Mere requirements of practical rationality, such as the requirement that everybody who is thirsty should drink something, do not justify such claims. These requirements of practical rationality are not duties, let alone obligations, and the constraints underlying them, such as the constraint that somebody who is thirsty should drink, are typically not called norms. The practical relevance of this second point is mainly semantic, however.

We assume therefore that norms are rules that constitute duties or obligations. Examples of such norms are the norms that:

A. Nobody should steal (everybody has the duty not to steal).
B. Car drivers must (have the duty to) drive on the right-hand side of the road (perform acts in a particular way).
C. Paul must (has an obligation toward Patty to) compensate the damage of Patty.

As was to be expected, these norms have formulations that are identical to their descriptive counterparts. From the formulations, it is not possible to detect whether we are dealing with a norm or a descriptive sentence. The two major differences between norms and their descriptive counterparts are that norms have the down direction of fit of constraints, while their descriptive counterparts have the up direction of fit, and that norms are from the logical point of view individuals, and not part of language, while their descriptive counterparts are sentences and therefore part of language.

The account that was given above of norms emphasized the constitutive nature of norms: They create, rather than are, duties and obligations. This may look somewhat strange at first sight, since the notion of a norm is more often associated with guidance of behavior than with the constitution of facts. Still this finding should not surprise if the idea is accepted that there can be facts that guide behavior. Rules that constitute such behavior guiding facts—that is, norms—are both constitutive and behavior guiding (regulative).

6.6 Competence-, Power-, and Other Status-Conferring Rules

The emphasis that is often placed on norms may draw our attention away from the rules that do not primarily aim at guiding our behavior. Still these other rules are crucially important for the functioning of more complex normative systems. To illustrate this, we will briefly pay attention to some rules that are not norms.

Dynamic rules attach new facts as legal consequences to the occurrence of some event. By performing some act that triggers the operation of a dynamic rule, an agent can bring about these legal consequences. For example, by committing a crime, a person can make himself liable to be punished, and by moving to a different municipality a citizen can change the amount of municipality tax he has to pay. The existence of a dynamic rule has a side effect that persons can do things which they could not have done without the presence of these rules. In that sense, they confer powers upon agents who through their acts are able to trigger the operation of dynamic rules and in that way bring about legal consequences.

The two examples given above of powers resulting from the existence of dynamic rules are not the most characteristic ones for law. The powers that can be exercised by performing a juridical act are much more characteristic. In this connection, a juridical act may be defined as an act that is aimed at bringing about legal consequences, to which these legal consequences are typically attached by a dynamic rule for the reason that this was the aim of the act.[53] Typical examples of juridical acts are contracting, making a last will or an association, legislating, pronouncing a judicial verdict, and granting a license.

An agent who performs a juridical act and thereby creates legal consequences must be competent to create these legal consequences by means of this kind of juridical act. This competence is assigned by a legal rule, although not necessarily an explicitly created one. For example, in order to be able to transfer the ownership of a piece of land, the transferor should be competent to alienate the land. This competence is

[53] The formulation "aimed at" has been chosen instead of the more natural sounding "performed with the intention to" to include acts that are performed without a conscious intention, such as acts performed by implemented computer programs. A public officer who signs a license without even reading it also performs a juridical act, and it should not be precluded by definition that a computer program that buys and sells securities thereby performs juridical acts. For this reason, the aim of an act should not be identified with the intention with which the act was performed. Aims are ascribed to acts, and (ascribed) intentions are merely factors that play a role in ascribing aims. Thanks go to Hester van der Kaaij for pointing out to me how important this innocuous-seeming difference between intention and aim is.

a legal status that is typically attached (by a fact-to-fact rule) to the ownership of the land. Being competent is a necessary condition for the successful performance of a juridical act and therefore also for the existence of a legal power that must be exercised through the performance of a juridical act. However, as we have seen from the two examples of the previous paragraph, legal powers, that is powers that exist because of the existence of legal rules, do not always require juridical acts for the exercise.

It is sometimes possible that somebody who lacks the competence for a particular juridical act nevertheless can succeed in bringing about the legal consequences of this act. For example, a public officer may succeed in providing somebody with a valid license, even though the officer lacked the competence to do so. Another example is that a non-owner may succeed in making somebody else the owner of a good. Both examples illustrate that legal consequences that could not be brought about through a valid juridical act may nevertheless occur for reasons of legal certainty. These examples illustrate that an agent can have the power to bring about legal consequences while lacking the competence to do so by means of a valid juridical act.

Rules that confer an agent the competence to perform a particular kind of juridical acts can both be dynamic and fact-to-fact rules. An example of the former is the dynamic rule that governs contracts and that by and large involves that the legal consequences which the contract partners intended to bring about will actually hold. If one contract partner wanted to make the other party competent to perform juridical acts in his name, that is to act as a legal representative, the effect of the contract is that this latter party has become competent to perform juridical acts in the name of the former contract partner. An example in which a competence is provided by a fact-to-fact rule is that the owner of a good is competent to alienate this good. The rule attaches the competence to the status of ownership.

Having a particular competence is a status assigned by a legal rule. There are many other examples of status assigned by legal rules. All legal counts-as rules assign a status to entities or events, such as the status of being a vehicle in the sense of the Road Act, being the president of Germany, being the commander in chief of the Belgian army, being a suspect in the sense of penal law, being the owner of a good, being wedded.[54] Very often norms attach deontic consequences to the presence of such a legal status.

7 Summary

In this contribution, an attempt was made to clarify the notion of a norm by elaborating the idea that norms are rules that lead to deontic consequences. The elaboration focused both on the nature of rules and on the nature of deontic facts.

[54] In a sense, even having a duty or an obligation can be seen as having a particular status, but this stretches the idea of legal status to its limits.

Rules, it was argued, are a kind of constraints on possible worlds. They determine which kinds of facts necessarily go together or cannot go together. Three kinds of rules were distinguished: dynamic rules which attach consequences to the occurrence of events; fact-to-fact rules which attach one fact to the presence of some other fact; and counts-as rules, which make that some things (often events) also count as something else. It was pointed out that the existence of a rule makes that some facts obtain: the factual counterparts of the rules. In this sense, all rules are constitutive. The descriptive sentences that express these facts, the descriptive counterparts of rules, are open generalizations, and they have often the same formulations as the rules from which they derive their truth.

By distinguishing between objective, brute social, and rule-based facts, an attempt was made to overcome resistance against the idea that facts might be normative and that there might be deontic facts. That something is mind-dependent does not exclude that it is a fact. Deontic facts are mind-dependent, because they are facts that tend—often in an indirect way—to induce a motivation to comply in agents to which they apply. Deontic facts are most often the result of the application of fact-to-fact rules (duties) or dynamic rules (obligations). A distinction was made between two kinds of basic deontic facts—the existence of duties and of obligations—and two kinds of supervening deontic facts: being obligated and owing to do something.

References

Alchourrón, C.E., and E. Bulygin. 1981. The expressive conception of norms. In *New studies in deontic logic*, ed. R. Hilpinen. Dordrecht: Reidel.

Alvarez, M. 2010. *Kinds of reasons. An essay in the philosophy of action*. Oxford: Oxford University Press.

Anscombe, G.E.M. 1976. *Intention*, 2nd ed. Oxford: Basil Blackwell.

Araszkiewicz, M., and K. Pleszka. 2015. The concept of normative consequence and legislative discourse. In *Logic in the theory and practice of lawmaking*, ed. M. Araszkiewicz, and K. Pleszka, 253–297. Cham: Springer.

Austin, J. 1954. *The province of jurisprudence determined*, ed. H.L.A. Hart. London: Weidenfeld and Nicholson. (1st ed. 1832).

Bertea, S. 2009. *The normative claim of law*. Oxford: Hart.

Broome, J. 2013. *Rationality through reasoning*. Chicester: Wiley Blackwell.

Castañeda, H.-N. 1972. On the semantics of ought-to-do. In *Semantics of natural language*, 2nd ed, ed. D. Davidson, and G. Harman, 675–694. Dordrecht: D. Reidel Publishing Company.

Darwall, S. 2006. *The second-person standpoint. Morality, respect and accountability*. Cambridge: Harvard University Press.

Frändberg, Å. Typescript. *The legal order. Studies in the foundations of juridical thinking.* (Typescript of a book in preparation).

Hage, J.C. 1997. *Reasoning with Rules*. Dordrecht: Kluwer.

Hage, J.C. 2001. Contrary to duty obligations. A study in legal ontology. In *Legal knowledge and information systems. JURIX 2001: The fourteenth annual conference*, eds. B. Verheij, A.R. Lodder, R.P. Loui and A.J. Muntjewerff, 89–102. Amsterdam: IOS Press.

Hage, J.C. 2005. *Studies in legal logic*. Dordrecht: Springer.

Hage, J.C. 2013. The deontic furniture of the world. In *The many faces of normativity*, ed. J. Stelmach, B. Brożek, and M. Hohol, 73–114. Kraków: Copernicus Press.

Hage, J.C. 2015. Separating rules from normativity. In *Problems of normativity, rules and rule-following*, ed. M. Araszkiewicz, P. Banaś, T. Gizbert-Studnicki, and K. Pleszka, 13–30. Cham: Springer.

Hansen, J. 2013. Imperative logic and its problems. In *Handbook of deontic logic and normative systems*, ed. D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, 137–191. London: College Publications.

Hansson, S.O. 2013. The varieties of permission. In *Handbook of deontic logic and normative systems*, ed. D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, 195–240. London: College Publications.

Hart, H.L.A. 2012. *The concept of law*, 3rd ed. Oxford: Oxford University Press. (1st ed. 1961).

Hartmann, N. 1962. *Ethik*, 4th ed. Berlin: De Gruyter.

Herrestad, H., and C. Krogh. 1995. Obligations directed from bearers to counterparties. In *Proceedings of the 5th international conference on artificial intelligence and law (ICAIL'95)*, 210–218. New York: ACM.

Hilpinen, R., and P. McNamara. 2013. Deontic logic: A historical survey and introduction. In *Handbook of deontic logic and normative systems*, ed. D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, 3–136. London: College Publications.

Jörgensen, J. 1937/8. Imperatives and logic. *Erkenntnis* 7: 288–296.

Kelsen, H. 1945. *General theory of law and state*. Cambridge, Mass.: Harvard University Press.

Kelsen, H. 1960. *Reine Rechtslehre*, 2nd ed. Wien: Franz Deuticke.

Kelsen, H. 1979. *Allgemeine Theorie der Normen*, eds. K. Ringhofer and R. Walter. Wien: Manzsche Verlags- und Universitatsbuchhandlung.

Loux, M.J. (ed.). 1979. *The possible and the actual. readings in the metaphysics of modality*. Ithaca, N.Y.: Cornell University Press.

MacCormick, N. 1978. *Legal reasoning and legal theory*. Oxford: Oxford University Press.

MacCormick, N., and O. Weinberger. 1986. *An institutional theory of law*. Dordrecht: Reidel.

Menzel, C. 2015. Possible worlds. In *The Stanford encyclopedia of philosophy,* ed. Edward N. Zalta. http://plato.stanford.edu/archives/sum2015/entries/possible-worlds/.

Navarro, P.E., and J.L. Rodríguez. 2014. *Deontic logic and legal systems*. Cambridge: Cambridge University Press.

Navarro, P.E., and J.L. Rodríguez. 2015. Entailed norms and the systematization of law. In *Logic in the theory and practice of lawmaking*, ed. M. Araszkiewicz, and K. Pleszka, 97–114. Cham: Springer.

Raz, J. 1975. *Practical reason and norms*. London: Hutchinson.

Sartor, G. 2005. *Legal reasoning, a cognitive approach to the law*. Dordrecht: Springer.

Schauer, F. 1991. *Playing by the rules*. Oxford: Clarendon Press.

Scheler, M. 1954. *Der Formalismus in der Ethik und die materiale Wertethik: Neuer Versuch der Grundlegung eines ethischen Personalismus*, 4th ed. Bern: A. Frankcke.

Searle, J. 1979. *Expression and meaning. Studies in the theory of speech acts*. Cambridge: Cambridge University Press.

Smits, J.M. 2014. *Contract law. A comparative introduction*. Cheltenham: Edward Elgar.

Toulmin, S. 1953. *The philosophy of science. An introduction*. New York, N.Y.: Harper Row.

Ullmann-Margalit, E. 1977. *The Emergence of Norms*. Oxford: Clarendon Press.

von Wright, G.H. 1963. *Norm and action. A logical enquiry*. London: Routledge and Kegan Paul.

Williams, B. 1981. Internal and external reasons. In Id., *Moral Luck*, 101–113. Cambridge: Cambridge University Press.

Zimmerman, R. 1996. *The law of obligations. Roman Foundations of the Civilian Tradition*. Oxford: Oxford University Press.

Values

Carla Bagnoli

1 Euthyphro Dilemma and Other Questions About Value

There are some fundamental problems concerning the meta-ethical status of values, their normative scope, and implications. These issues are importantly related, but it is useful to consider them separately. First, there is the *ontological* question: Is there value in the world? Are values part of the fabric of the universe, or else artifacts, projections of the mind, products of social conventions? In the platonic dialogue, Euthyphro asks whether the gods love the good because it is good or else the good is good because the gods love it (Plato 1991 6e–9e, 9d–11d). This is a fundamental dilemma about the nature and the ontological status of values, which divides realists and anti-realists. If what is good is good because gods love it, then the good is relative to them and dependent on their attitude. Vice versa, if gods love the good because it is good, it must be a real feature of the world, independent of what the gods happen to like or dislike. In the former case, values are subjective and their aspiration to objectivity seems problematic. In the latter case, the ontological status of values seems firmer and grounds the aspiration of value judgments to objectivity.

However, if values enjoy such an independent ontological status, it becomes unclear how they relate to other factual elements of the fabric of the world, but also how they matter to evaluators. This is a question that applies also in the case of the Platonic gods, but it becomes particularly interesting in the case of standard evaluators. How do objective values affect the life of such evaluators, presumably limited in their epistemic and practical rationality? This is the core issue of meta-ethics that aims to vindicate both the aspiration to objectivity and the practical relevance of

C. Bagnoli (✉)
Dipartimento di Studi Linguistici e Culturali, Università di Modena e Reggio Emilia, Modena, Italy
e-mail: carla.bagnoli@unimore.it; carla.bagnoli@gmail.com

C. Bagnoli
Department of Philosophy, Classics, History of Art and Ideas, University of Oslo, Oslo, Norway

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_6

value judgments. The tension between these two aspirations to objectivity and to practicality is so central that it has been dubbed "*the* moral problem" (Smith 1994). However, as we shall see, this apparent tension can be solved by understanding objectivity in a way that does not privilege ontology. After all, an objective value is also expected to exercise authority and bear some relation to our practical lives. Thus, the key problem is how to understand value objectivity.

Some of the problems generated by the search for objectivity are related to a well-established dichotomy between *matters of fact* and *matters of value*. Is this dichotomy to be taken for granted? The axiological status of subjects, agents, and objects crucially depends on what happens in the natural world. Differently than other forms of dependence, the dependence of values on facts is a conceptual truth. That is, it appears to be true a priori that the axiological status on an object cannot vary unless something in its natural makeup varies (Moore 2010; Hare 1963). In other words, two states or affairs or objects that are identical in their natural makeup must have identical axiological status, hence be evaluated similarly. How to understand and account for this sort of co-variance between factual and evaluative aspects is a problem that any theory of value should address. Reductive naturalists insist that the co-variance is implicated by a relation of identity; moral facts vary according to natural facts just because they are reducible to, or identical with, them. By contrast, anti-naturalists hold that moral facts or moral properties supervene upon natural facts and properties, but they are not reducible to them. An intermediate view holds that moral and natural concepts are not identical, but they share the same natural ontology. This view has the advantage of preserving the independence of the evaluative conceptual apparatus without committing to a special ontology. By contrast, an extreme position requires that ethical concepts be eliminated because they are not supported by any ontology.

A second, related, question concerns the *epistemology of value*. Can there be knowledge of right and wrong, good and evil? If there is, how is it accessible to us? Cognitivism and non-cognitivists offer opposite answers. If we can acquire knowledge of values, is it by education or by exercising and developing natural dispositions? This question admits of large varieties of positions. We can distinguish two main approaches, the rationalist and the sentimentalist. On the first approach, value is accessed by exercising rationality. The two main varieties of this approach are the rationalist intuitionism and constructivism, Rational intuitionism claims that we know right from wrong because reason recognizes it (Moore 2004, 2010; Ross 1930 ; Audi 2001; Dancy 2000, 2004; Skorupski 2010 ; Parfit 2011). By contrast, constructivism holds that practical knowledge is provided by engaging in practical reasoning (O'Neill 1989; Korsgaard 1996; Scanlon 1998, 2014; Bagnoli 2013; Wong 2008). Instead, on the sentimentalist approach, sensibility is the source of value and of moral distinctions (Blackburn 1998; Nichols 2004; Prinz 2007; Schroeder 2010, 2011). On this view, reasoning is driven by emotions and we have knowledge of right and wrong because we exercise virtues, which are a natural endowment of our sensibility (Aristotle 1984;Baier 1985, 1994; Foot 2000; Geach 1956; Hurka 2000; MacIntyre 2007, 2008; Slote 1992, 2010).

A third, related issue concerns the *semantic and logical* status of value judgments and the logical grammar of evaluative concepts. Are value judgments assertions about properties or states of affairs, or are they akin to prescriptions or expressions of normative states? Do evaluative concepts represent properties or facts of the world, or are they expressive of mental states? If evaluative judgments are not assertions about states of affairs and evaluative concepts do not represent properties and facts of the world, how are they anchored to reality? If they are not anchored to reality, how can they be responsive to facts and subjected to rational criteria of justification and revision? Realism is the view that ethical judgments are assertions about properties or facts of the world, and thus behave exactly like other sorts of factual assertions. Anti-realism denies all the above. Emotivism holds that evaluations are merely expressions of mental states, hence cannot be true or false (Ayer 1936; Stevenson 1937, 1963, 1979). The major problem of emotivism is that it seems to fail to account for central linguistic phenomena such as embedding moral and other evaluative expressions within conditionals (Geach 1960). The major problem of realism is that it bases its claims on truth of a special ontology, which is problematically related to natural ontology and may seem to require a special faculty of knowledge (Blackburn 1993; Mackie 1977, 44ff; Wong 1986).

These positions identify the two extremes of a large spectrum of more nuanced and hybrid positions. Universal prescriptivism is a similar view, but it claims that evaluative judgments have logical and semantic properties that generate practical reasoning (Hare 1952, 1963, 1981). Normative expressivism is the view that judgments of value are expressive of emotions of normative states, which are themselves governed by logical norms, in ways that seem to avoid Geach's problem about embedded contexts (Gibbard and Macintyre 1995). Quasi-realism holds that evaluative judgments only resemble assertions but they are not. They exhibit some sort of objectivity, but they have to earn their right to truth (Blackburn 1984). Error theory holds that value judgments are like assertions, but they are all false because they do not represent any real property or fact of the world. Both quasi-realism and error theory involve a systematic error on the side of the evaluator; hence, they are committed to a projectivist account of moral phenomenology, which is supposed to explain how it happens that people are trapped in this illusion, and how value judgments can guide them nonetheless. By contrast, cognitivist irrealism holds that evaluative judgments have cognitive import, but they do not represent a special sector of reality (Skorupski 1993, 2010). Kantian constructivism can be plausibly defended as a form of cognitive irrealism (Bagnoli 2013).

Finally, a large cluster of questions concerns the *normative authority* of values. Do values produce normative reasons? Do such reasons apply universally, across all evaluators, in all relevant circumstances? Does the adoption of values commit the rational agent to act for the pursuit of value? At least some values entail normative reasons. For instance, if Spartacus values equality, then he has a reason to support policies that promote equality. Such reasons may not be overriding; that is, they can be defeated by stronger reasons. Brutus values friendship, and he has a reason to love and protect his friend Caesar. Yet Brutus decides to participate in Caesar's assassination, not because he loved Caesar less, but because he loved Rome more. In this case, the value of civic friendship trumps the value of personal friendship.

Does civic friendship always trump personal friendship? Do patriotic values generate normative reasons that apply to all of us?

Arguably, moral values, such as the value of humanity, produce reasons that trump all other kinds of normative reasons. Especially for such values, the question arises about the scope of their applicability. According to some ethical theories, moral values are universally authoritative and give everyone normative reasons. Rationalist theories and Kantian theories hold that moral obligations are requirements of reason, hence applicable to all rational agents (Kant 1997; Audi 2005). They claim that the moral values are universally compelling exactly because they are required by reason (Parfit 2006).

However, a further question is whether the requirements of reason are universally compelling (Broome 2013; Kolodny 2005). Some are skeptical that reason may be universally binding not because human agents are defective but because rationality produces merely conditional requirements. Relativist theories hold that moral values are rooted in particular traditions and that their authority is local (MacIntyre 1988; Wong 2006). Consequently, values generate reasons that are binding and authoritative only for the members of specific communities, and even though they claim universal authority, such a claim is misplaced. The traditional objection against relativism is that of incoherence (Lyons 1976), but Harman (1977) attempts to avoid the problem by distinguishing between normative and meta-ethical relativism. In contrast to all above, naturalist accounts typically hold that the question of normativity is bogus, and explain the apparent authority of normative claims with psychological and sociological processes of internalization and enforcement. Typically, these accounts propose a genealogy of moral values meant to undermine the special authority of moral normativity (Mackie 1977; Joyce 2001). Not all genealogies are debunking, however. Building on studies in evolutionary biology, Gibbard (1990) and Nozick (2001) argue that ethical norms and values are selected because they favor coordination to mutual benefit. While Gibbard holds that the approach supports an expressivist account of normativity and normative language, Nozick favors a more complex theory, akin to structural realism, according to which there are invariant ethical structures identifiable under admissible transformations. A plausible philosophical explanation should focus on such structures, even though there can be competing explanations of the same phenomena. Nozick's explanation relates the raise of normativity to a normativity module favored by evolution, and it is capable of defending the possibility of moral progress, which he identifies through "a multistage process whereby cooperation between distinct groups gets established" (Nozick 2001, 263). A similar conclusion about the relation between progress and agreement on a core set of practical norms is drawn by David J. Velleman, even though he emphasizes the relativist implications of the claim (Velleman 2009, 2013).

These debates further contribute to larger and deeper issues about the *objectivity of value*. On a realist standard, the objectivity of values is an issue determined by their ontological status (Harman and Thomson 1996). However, this position makes it impossible to explain how values have an impact on our life and action. If values are intrinsically motivating properties, then they are queer properties. This is the objection that J. L. Mackie moves against objectivist theories of values. It is similar to Hume's objection that moral distinctions do not originate in reason because reason alone cannot move us to action (Hume 1739, 456–457; Schroeder 2009, 2011). Hume and Mackie assume that the authority of values and their significance to action can be understood in terms of their motivational force. Mackie concludes that the motivating power of ethical judgments shows the inadequacy of all objectivist theories of value. Hume, instead, takes this to be an argument against rationalism. Both assume that objectivity is granted by ontology. However, many philosophers suggest that the ontological conception of objectivity is useless or even dangerous in ethics. Some hold that ontological objectivity is inapplicable because there is no hope to converge on an independent reality (Williams 1985). Others argue, instead, that this is a misconception of the objectivity of value, which has to be rejected. Sensibility theorists, such as John McDowell and David Wiggins, argue that values bear an interesting resemblance with secondary qualities such as colors, insofar as the exercise of evaluative judgments is cognitive and yet concerns properties that are neither part of the causal structure of the world nor reducible to them. Our sensibilities seem to be inevitably implicated in evaluations, and this should indicate that the objectivity of evaluative standards cannot be totally independent of our distinctive sensibility. This is one way in which sensibility theories attempt to reconcile the aspiration to objectivity with the action-guiding aspiration of evaluative judgments. The outstanding question is whether they can thereby vindicate the authority or compellingness of some categories of ethical values.

Focusing on such categories, Roderick Firth and Richard Brandt propose to conceive of objectivity and normativity in terms of idealized conditions of rationality, which importantly include dispositions to respond to value (Firth 1951; Brandt 1996). By contrast, others explore a conception of objectivity centered on the principles of practical reason (Baier 1985; Toulmin 1950; von Wright). In particular, Kantians propose a practical conception of objectivity, which is based on shared reasons constructed via reasoning, under idealized conditions (Rawls 1980a, b; O'Neill 1989). It is often assumed that the practical conception of objectivity is weak and more modest than the ontological conception. In fact, the practical conception is more ambitious and demanding insofar as it includes other important dimension of values, such as their authority (Bagnoli 2013). For G. E. Moore, the objectivity and autonomy of values is warranted by the reality of irreducible normative relations (Moore 2004, 2010). On the Kantian view, instead, the autonomy of ethics is granted by the supremacy of pure practical reason. For Kant, practical reason has the unique power of producing its proper objects and should be recognized as sovereign. John Rawls has revived the Kantian conception of ethical objectivity in order to overcome an impasse in political theory, which he attributed to an inadequate conception of the standard of objectivity (Rawls 1980a, b). Neither Rawls nor Kantian philosophers are oblivious of the fact that the authority of value is often based on relations of powers. However, they think that progress can be made in building a dialogue governed by mutual respect and recognition. They also think that reasoning together is an activity constructive and productive, that is, generative of reasons that we could all share. To this extent, these philosophers believe in the reconciliatory and cooperative powers of reason.

In contrast to rationalists, relativists account for the apparent authority of moral values by invoking mechanisms of social enforcement, such as blame, and institutionalized devices to compel and coerce into compliance (Williams 1985; Blackburn 1998). On this view, the special status of moral values depends on the specific manner of their social enforcement, rather than on ontological and epistemological features of values. For Nozick, coercive enforcement is legitimate only insofar as it concerns the basic ethical norms of respect and cooperative virtues. He considers it a sign of moral progress that personal values are not a matter of social enforcement (Nozick 2001, 264–265). Gibbard distinguishes between repressive and coercive enforcement, showing that repression is always morally objectionable, while coercive enforcement may be necessary to the stability of society. Furthermore, he argues that the facts of value disagreement may be so divisive that it might be prudent to endorse principles of accommodation (Stevenson 2009).

The issue of the legitimacy of enforcement acquires center stage in liberal theories. Famously, John Rawls advocates a principle of political legitimacy according to which the exercise of power is legitimate only insofar as it respects all citizens as free and equal, hence capable to reasonably disagree about matters of values (Rawls 1993, 137). Reasonable disagreement is to be protected, rather than erased or undermined; this is a consequence of an argument that treats persons as "self-originating sources of valid claims" (Rawls 1980a, b, 582). Ontological issues appear to have crucial normative implications for law as a coercive system of norms. Should law protect values? Which values should law protect? Can values be enforced, and should they? And, most importantly, which values should be enforced? A crucial example is the case of human rights: If they are based on the value of humanity, do they bind even those who do not endorse this value? Which authority is in charge of the enforcement of human rights, in the face of a disagreement about the basis of their justification?

Both Rawls and Nozick agree that coercion and the exercise of power stand in need of rational justification. For Rawls, political power is legitimate when used in ways that all citizens can reasonably be expected to endorse. Thus, the legitimacy of a particular set of basic laws is determined by the so-called *criterion of reciprocity*, according to which citizens must reasonably believe that all citizens can reasonably accept the enforcement of a particular set of basic laws. This is an epistemic principle that governs the mutual expectations of citizens. An analogous epistemic principle governs the normative expectations of groups that recognize mutual dependence and thus coordinate for mutual benefit (Nozick 2001, ch. 5; Gibbard 1990). In a just society, citizens are in the epistemic and moral position to endorse the fundamental political arrangements freely, that is, free from manipulation, repression, and lack of adequate information. As we will see in Sect. 5, this issue is particularly relevant in the case of a pluralistic citizenry.

2 Aims and *Desiderata* for a Theory of Value

While the basic questions about values admit of different answers, there is a significant convergence about the aims and desiderata of an adequate theory of value. Theories of values are expected to explain the nature of values, specify their varieties, and account for their place in our lives. Correspondingly, the most important criterion of adequacy for a theory of value is its capacity to offer good philosophical explanations of why and how we should value things, objects, and activities. Call this the *explanatory capacity* requirement. Further requirements apply insofar as the basic aim of the theory is to produce convincing explanations of phenomena concerning our evaluative practices. Arguably, a good explanation is coherent. *Coherence* is a logical requirement, which applies especially if the theory of value is systematic and normative, but it seems also important for those theories of value based on a rational justification of value. In view of debates about the nature and ontological status of values, there is a consensus about the criterion of *ontological parsimony*, according to which one should not introduce more entities than they are required to explain the phenomena. However, philosophers disagree about the sort of properties that perspicuous explanations should admit in order to make sense of valuing practices and discourse. Theories of value are also expected to account for whether and how we know values, or else explain why the quest for knowledge of this sort is misplaced. Concerning the relation between facts and values, an adequate value theory seems superior to others when it explains why some values or clusters of values are indispensable to explain a range of facts. In other words, a value theory places values in the best explanation of what we experience in the world. According to some, values do not figure in the best explanation of what we experience in our evaluative practices; hence, they are dispensable entities (Harman 1977, 2000; Harman and Thompson 1996). According to others, instead, values have a place in the best explanation of how things stand (Sturgeon 1986, 1998). This position is supported by various versions of the *open-question argument*, originally devised by G. E. Moore. Moore argued that for any definition of good in terms of other properties, it is an intelligible question to ask whether such a property is good. This means that the meaning of the concept of good is not exhausted by its analysis. The implication is that matters of values cannot be explained away in terms of matters of (natural) facts, and that any reduction of value in terms of other properties leaves the crucial question unanswered (Moore 2004, 44). This argument is directed to any attempt at reducing value to other properties, either metaphysical or natural. Moore also argued that there is a naturalistic fallacy involved in such a reduction, but many critics have pointed out that there is not (Frankena 1939). Naturalists claim that the only properties necessary belong to a naturalist ontology, while realists claim that we should admit also of moral and evaluative properties, which are not identical with or reducible to natural properties. Subtler disagreements concern the very definition of natural and non-natural properties, and the forms and status of naturalistic reduction.

Thomas Scanlon offers a reductivist analysis of the concept of good in terms of reasons, which does not seem vulnerable to the open-question argument. His "buck-passing" analysis avoids indicating which properties make and object good. The analysis is supposed to show that once the evaluator has identified the reasons she values something as good, there is no further work for the concept of value to do. This is the claim of redundancy of value (Scanlon 1998, 97). Furthermore, the analysis of value in terms of reasons demystifies evaluative concepts and dispels the aura of gravity and inexplicable compellingness that surrounds terms such as good (Scanlon 1998, 98).

Disagreements about the possibility of reducing value to other concepts are relevant to determine the perspicuity of explanations. According to a general criterion of *descriptive plausibility*, a theory should explain why values appear as they do. For realists, the criterion requires that we consider values as part of the fabric of the world and take evaluations at face value, as assertions about values. Instead, others hold that a good explanation of value should fit our ordinary understanding and practices of value, and propose a requirement of *congruence* with the subjective experience of valuing we have as agents and citizens. On this view, it would be implausible to offer an account of values that systematically discounts and undermines the subjective experience of evaluators. By contrast, error theorists adopt a weaker criterion that allows an adequate theory to explain away the appearances of our evaluative practices as due to some systematic illusion or error (Mackie 1977). Likewise, projectivism holds that evaluative judgments formally behave like assertions about what is the case, but they are in fact the result of subjective projections or patterns of objectification (Blackburn 1984). Through such patterns of objectifications, values have gained a relatively solid ontological status, even though they originate in our mind and social practices, rather than being features of the world. These weaker criteria of descriptive plausibility seem problematic, however. On the one hand, it is unclear whether their semantic apparatus suffices to warrant our claim to truth. On the other hand, their analysis of value judgments as involving a systematic error or projection seems ultimately self-defeating: It corrodes moral authority and demands such a radical revision of ordinary evaluative practices that they become unsustainable as ordinarily known. An adequate explanation must not be self-effacing; that is, it should be one that does not undermine or undercut the importance attributed to moral values in ordinary thinking. I will call this the requirement of *reflective stability*, and others call it *transparency* (Korsgaard 1996, ch. 1).

In addition to criteria that measure the explanatory capacity of a theory of value, there are others that are meant to identify its *normative capacity*, that is, the capacity to guide action, attitude, and belief. It is a criterion of adequacy of value theory that it accounts for the relation between *values and rational requirements*. This criterion is interpreted in different ways, according to the various theories of rational authority. Consequentialists reduce all issues of deontic relations to value. The contrary view reduces value to deontic or normative relations, for instance the relation of reasons. This seems to be the case of the so-called buck-passing account (Scanlon 1998, 97), which reduces claims about the value of an object to the reason-providing properties of the object. The concept of value can thus be analyzed in terms of reasons and the properties of objects that provide them for evaluators.

In relation to the normative capacity to guide action, some philosophers hold that a theory of value should also account for the fact that at least some values are motivating, other things being equal. As mentioned above, Humeans hold that the authority of value amounts to its motivational and conative force. Endorsing the Humean perspective, some further argue that adopting a value commits to being motivated and bringing it into the world. For some, such a commitment is intrinsic to the very idea of adopting a value, so that when we endorse a value judgment we are thereby motivated to act accordingly, absent any incapacitation, interference, or impediment. On this view, the relation between value and motivation is thought to be internal and conceptual. For others, instead, there is no such a relation. When values motivate, it is because of an intervening external factor, such as a desire or an interest, whose independent force motivates action in conformity to value judgment. If Spartacus promotes equality, it is because he wants to promote equality, not merely because he thinks that equality is valuable. This desire attaches to and accompanies the evaluation and works as a trigger of action, a force external to the evaluator's practical reasoning and judgment. By contrast, Kantian philosophers such as Thomas Nagel defend the view that practical principles themselves generate desires (Nagel 1979). For instance, the desire to promote equality is generated directly by the adoption of the principle that prescribes equality. Furthermore, Kantians do not think that the normative capacity is exhausted by the capacity to motivate action. On the contrary, they hold that the key notion that articulates the normative capacity of value is that of rational requirement. Their view is that action motivated by the very idea of duty is of special value. By contrast, actions that conform to duty, but are motivated by self-interest or inclinations, do not have moral value. Such views link theories of value to moral philosophical psychology; hence, the debates about standards of objectivity are placed at the juncture among different disciplines. A further methodological issue is whether such disciplines are thoroughly empirical or may accept a priori arguments, concerning for instance the conditions of possibility of valuable action or the form of a moral agency (Anscombe 1957, 1958; Murdoch 2013; Thompson 2008).

3 Some Substantive Questions About Value

In addition to meta-ethical questions about the reality of values, the role of evaluative practices, and the logical grammar of evaluative judgments, there are substantive questions about value. What things have value, and how so? There are many things that seem good and a variety of ways in which they are good. Pleasure, knowledge, beauty, and personal relations are natural candidates and often thought of as indispensable ingredients for a flourishing life. Some theories take this variety seriously and hold that there are many different values and many ways in which things are good. These are pluralist theories, and they vary according to which goods they include and how they think such goods are related, if they are related at all. By contrast, monistic theories recognize only one kind of value. For instance, hedonism takes pleasure to be the only value, and utilitarianism takes utility to be the only value.

Value is a generic term that is attached to several varieties of items, which can be grouped into three basic categories of values. Values can be attached to (i) actions, policies, and institutions; (ii) objects, plans, projects, lives, prospects, events, states of affairs, outcomes, consequences, and effects; and (iii) to character, character traits, dispositions, intentions, and actions understood as bearing the marks of agency. The first category is generally governed by concepts such as right and wrong. The second category is understood as the sort of goodness that identifies something as worthy of choice. A fair taxation is good because it brings about good effects, promotes equity, and favors the least advantaged. The third category is often qualified as moral, in that character and intentions are traditionally objects of moral evaluation and assessment. There are important disagreements also regarding how we value things belonging to these categories. For instance, ethical theories take items in the third category as special sorts of values, but they differ in accounting for what makes them especially valuable.

When focusing on the act of valuing, it is useful to distinguish three kinds of value judgments. First, evaluators value an object because of the specific features it has as an object of a certain kind. A diamond is good because of its particular luminosity, purity, and saturation. Second, we may value something as a good thing; for instance, when we say that peace is a good thing to live or die for, or to bring about. Third, we may regard some things to be good absolutely, in that the world is better because of their existence; beauty, love, and friendship may be proposed as instances of this kind (Kolodny 2003).

When we say that some things are good, one may intelligibly ask "Good for what?" This question does not seem always pertinent, because it is often the case that the answer is implicitly contained in the description of the object that is said to be good. For instance, it may be superfluous to ask what a knife is good for, insofar it is obvious what the function of a knife is and also what properties make it suitable to fulfill its function. In other cases, instead, how to identify the relevant function is more problematic. What is the function of a human being? Is there any one function that fully defines what humans are? What is the function of government? What functions do moral virtues help us realize? Inspired by Aristotle, many seem to think that the disanalogy is only apparent and that all evaluative judgments admit of a similar analysis: They are relative to a function, even though reference to such function is often implicit. In support of this view, some have argued that the term good is a predicate modifier, rather than a predicate (Aristotle 1984; Geach 1960).

The disagreement about the function of the concept of good is particularly relevant in ethics as it bears normative and deontic implications. When the question "for what?" applies pertinently, then the value at stake is instrumental. For instance, practicing piano everyday is good for learning to play piano and drinking water is good for health, money is good for buying things we need. The deontic implication of these propositions is that money is good only to the extent that and insofar as it is a means to get some other goods, e.g., an object, a title, or a status. Many objects and activities are instrumentally good, but are all goods instrumental values? Some philosophers hold that goodness is a property, which can be predicated also *simpliciter*. In this case, the value at stake is intrinsic. A typical example of intrinsic value is the value of persons, even though this claim is not uncontested. Utilitarians and Aristotelians identify happiness as the overarching value, but they define it in different ways. For pluralists, happiness is a complex condition to which many different goods contribute. Some of these goods are objects, activities, or personal relations.

The difference between instrumental and non-instrumental values can be explained in terms of perspectives. What is good for *F* is good for her own perspective. Being good *simpliciter* means being good from the point of view of the universe, hence from no specific perspective at all, i.e., good "from nowhere," that is (Nagel 1986). Agglomerative theories, instead, hold that goodness *simpliciter* amounts to the sum of all perspectival goods. For instance, utilitarianism is committed to this view. Egoism typically holds that what is good *simpliciter* is defined by what is good for the evaluator. A distinct but related question is whether the bearer of value is always a state of affairs. Some are inclined to think that if good concerns states of affairs it is always attributive, and then goodness *simpliciter* can be analyzed in terms of what it is good for. By contrast, others hold that it is a mistake to conceive of the constitutive aspect of valuing in terms of bringing about a state of affairs.

On some theories, the distinction between instrumental and final goods collapses on, or is equated to, another distinction between intrinsic and extrinsic goodness. Instrumental goods are also held to be extrinsic goods, and final goods intrinsic goods. However, the distinction between intrinsic and extrinsic goodness concerns the sources of values, while the distinction between final and instrumental goods concerns the way we attach value to items. To bear in mind this distinction allows us to appreciate that there are things valued as extrinsically good, and yet as final ends. For instance, painting or horseback riding might be final ends because of the interest we take in them. Separating these distinctions allows us to distinguish two ways to treat this case. Some realists hold that the goodness of final ends is intrinsic, absolute, and independent of interests and desires (Moore 2010). By contrast, Kantians allow for extrinsically valuable ends that are valuable because people take an interest in them (Korsgaard 1983). Such ends are things we want for there own sake, but whose justification resides in something else, e.g., in the fact that we want them. They are not unconditionally good, since their being good is relative to us, but they are nonetheless rationally justified on the basis of their desirability.

4 Theories of Value

Traditionally, the theory of value represents one of the two main branches of ethical theory, along with the theory of right. While the theory of right tells us what we ought to do, the theory of value specifies to what end we should act, what states of affairs are good or bad, hence the distinction between virtue and vice. We can distinguish five main theories. *Hedonism* is the view that only pleasure is intrinsically good, and only pain intrinsically bad. This view admits of many different formulations, which vary in the definition of pleasure. According to some, pleasure is a sensation which is felt as pleasant. According to others, pleasure is a sensation that people want to have because of its subjective quality or sensation. The latter formulation seems to include a richer view of value that the term pleasure does not capture. A key issue for hedonism is measurement, since in order to justify determinate evaluative judgments, one must clarify how subjective sensations can be compared and ranked. For Jeremy Bentham and Henry Sidgwick, pleasure and pain are symmetrical, but others hold that pain is a greater evil than the absence of pleasure, and attribute greater ethical significance to intense pain.

While the view that pleasure is good and pain is bad is very plausible, the hedonist claim that pleasure is the only good is highly controversial. A crucial argument against this view is produced by Robert Nozick, and it involves the thought experiment of the pleasure machine (Nozick 1974, 43). This is a fictional case in which we consider whether to be plugged into a machine that gives us pleasure by interacting with our brain. If hedonism were right, then we all have decisive reason to plug in. By contrast, Nozick argues that there are three reasons not to plug in, all referring to how we care about knowledge. First, what matters to us as agents is not simply to have the experiences associated with actions; rather, agents want to be the persons who do such actions. Second, agents care about being persons, rather than a mass floating in a tank capable of feeling pleasure. Third, agents care about experiencing actual reality rather than a simulation. The pleasure machine simulates the experience of pleasure without the subject experiencing any actual reality. It matters to us that we feel pleasure through experiences, without losing grip of reality.

Desire theories argue that pleasure is too restrictive a notion, insofar as one might desire goods and experiences that are not themselves pleasurable, as Nozick's pleasure machine thought experiment shows (Nozick 1974, 43). In Nozick's case, knowledge of how things stand and acquaintance with reality is a more desirable experience than the comforting and pleasurable sensations grounded on false belief. This shows that there are things that are important and matter to us but not because they are pleasurable. As for hedonism, there are different formulations of desire theories. One is that good is a state of affairs where the agent obtains the good he desires. Others commit to more normative claims about what constitutes the well-being for a person, which may differ from what one actually desires in particular circumstances. It is controversial whether this kind of theory avoids Nozick's objection, insofar as it seems committed to say that whenever people desire to be plugged into the pleasure machine, this is good for them. It seems that to avoid Nozick's objection, one needs to offer stronger grounds.

Perfectionism identifies a plurality of goods and typically regards knowledge as the highest value. This theory has been defended in various forms, from Plato and Aristotle, to Aquinas, Hegel, Nietzsche, and G. E. Moore in the analytic tradition. Some perfectionist accounts emphasize the plurality of goods as a ineliminable feature of the theory; others recognize the plurality of goods, but they also insist on the importance of giving a unitary account in the explanation of goods, and also, and more importantly, to organize the plurality of concrete goods under a more general and abstract characterization or formal structure (Hurka 1993, chs. 8–10). Perfectionism about the good is typically combined with an account of the virtues that make such goods achievable and realizable.

Since the goods are many, and virtue is what makes us fit for the good, the question arises as to how to harmonize the virtues. Perfectionists in the early twentieth century, such as G. E. Moore and W. D. Ross, do not offer a unified theory of the virtues, but rather insist on the distinction between things that are good and dispositions of the mind and of characters that qualify as virtues. However, these perfectionists were sensitive to the explanatory question raised above and tried to account for a common ground that explains both why some things are good and why some character traits and dispositions of the mind are virtuous (Moore 2004, 204, 208–211; Ross 1930, 134–160). Instead, more recent perfectionists argue against the view that a unified explanatory account is needed (Adams 2006, 31). More ambitious theories aim to offer a unified theory of value and virtue. The first prominent example of this case is Aristotle. On the Aristotelian account, the most accomplished human life consists of a complex hierarchy of material goods, excellences of character, and excellences of the mind, which all work together toward the realization of the human being as a rational and political animal. Among all activities that are distinctive and typical of human beings, there are cooperative activities among friends.

5 Value Disagreement

People disagree about what to value, why, and how. Some love knowledge, and others search for pleasure. Some think knowledge is valuable because it helps us survive and others claim it is good in itself. Further disagreements concern the normative implications of valuing objects such as knowledge, pleasure, or people. What kinds of normative commitments and attitudes do valuing require? If one values knowledge, one is committed to promoting knowledge. If one values pleasure, one has a normative reason to engage in activities that are of one's satisfaction. If one values persons, does one have the same sort of normative reasons as in the previous cases? It seems that valuing persons requires different normative attitudes than those associated with realizing states or affairs or bringing about consequences. This is apparent in the case of conflicts of values where the life of others is at stake and there is no policy that relieves the agent from guilt.

Some argue that value disagreements are more widespread and more pervasive in ethics than they are in science (Harman 1977; Williams 1985). They remark that even when there is an agreement about basic facts of the matter, a value disagreement may persist. For instance, conservatives and liberals may handle the very same data about stem cell research and yet disagree about the moral and political permissibility of it. How to understand such disagreement and what conclusions to derive from the apparent facts with which we do disagree is an open question. However, the brute fact of value disagreement has supported two meta-ethical positions: nihilism about the existence of value and skepticism about the possibility of moral knowledge.

These two positions are often conjoined, but they support different claims about the nature of value. Nihilism is the view that there are no values and it does entail that there cannot be knowledge of values, where this is understood to be knowledge of peculiar objects. However, moral knowledge may be construed differently, as a kind of practical knowledge that originates in a special relation one entertains to oneself as a practical subject. In this distinctive sense, moral knowledge does not require any moral ontology, and thus, it is not in contrast to nihilism. Furthermore, one may hold that values exist, but lie beyond our reach. This would be a case in which skepticism about moral knowledge presupposes realism about value and makes the condition of possibility of knowledge so distant from our standards that they are never met, and thus, knowledge is never obtained. This is a logically consistent position, but of little practical import. Normally, however, nihilists tend to be skeptical also about the possibility of moral knowledge.

Whether widespread and persistent disagreement about value demonstrates nihilism and supports skepticism is a controversial and complex matter, which we can begin to address by considering the sort of disagreement that would challenge moral truths and the existence of values. A first consideration is that disagreements about values can be identified and discerned against the background of shared practices. For instance, liberals and libertarians strongly disagree about the sort of cases in which it is legitimate to use coercive instruments to redistribute wealth such as taxation, but this disagreement is intelligible only against the background of shared practices informed by the principle of political legitimacy, which requires one to justify the deployment of coercive enforcement. Second, interesting disagreements often concern specific claims and are intelligible against the background of general principles. For instance, different traditions have delivered very peculiar catalogs of the virtues, the excellences of character. Philosophers as diverse as Aristotle, Hume, and Kant disagree about the significance of wit, or the morality of magnificence, but they would broadly agree that inflicting unbearable suffering for fun is morally objectionable or that unqualified pain is a disvalue (MacIntyre 1988, 2007, 179–181; Scanlon 1995). This seems to suggest that there are some very basic and general norms that are invariant, while more specific moral norms vary across societies. Naturalists explain this fact on the hypothesis that the core ethical norms favor coordination for mutual benefit and different societies might find different equilibria (Gibbard 1990; Nozick 2001, ch. 5).

It is noteworthy that there is an asymmetry about how the facts of value disagreement and agreement are used in the debate about the ontological and epistemological status of value. Persistent value disagreement in the face of factual agreement is thought to be evidence for skepticism and nihilism; but the presence of large areas of agreement and concordance is not typically used to show that, at least on such areas moral knowledge can be obtained (Nagel 1979; Parfit 1984, 452–453). Third, some philosophers insist on vagueness as a possible epistemic source of disagreements. On this analysis, our value disagreements depend on the fact that we do not know what is the correct position on value matters, even though there is one correct answer in each case (Brink 1984; Boyd 1988; Shafer-Landau 1994). There are interesting cases of value disagreement that rest on factual ignorance, irrational ignorance, self-deception, or implicit bias, and thus can be fruitfully explained otherwise, without endorsing nihilism about values. For instance, two colleagues might disagree about the opportunity to increment diversity in a department, not because they disagree about the value of diversity or the proper means to implement it, but because they do not have access to the same facts about the presence of minorities in positions of powers or holding offices. Furthermore, another colleague might disagree about the same issue, not because she does not have access to the relevant facts of the matter, but because she self-deceptively resists the evidence that the department is not as diverse as she thinks it should be, discounting the facts that would support a different belief about the state of the department and its recruitment policy (Shafer-Landau 1994). These are interesting analyses because they uncover some complexities in the case of value disagreements, whose roots are not always identifiable as either factual or evaluative. Evaluators often disagree not only on facts per se, but also on their relevance and importance, on the nature and strength of evidence provided, and on the question who bears the burden of persuasion.

There are different perspectives about the significance of genuine value disagreements. The presence of irresolvable value conflicts is generally taken to undermine the aspiration to objectivity of ethics. For instance, Bernard Williams argues that the pervasiveness and untreatable character of conflicts of values shows that the aspiration to ethical objectivity is not a matter of logic and cannot be understood in terms of convergence on an independent moral reality (Williams 1985). There is a pressure toward resolving disagreements, but this is to be understood as a psychological and sociological need, functional to peacefully living together. Conflicts do not reveal any pathology of practical thought, but exhibit the richness and varieties of values. Thomas Nagel shares this diagnosis and further concludes that what we call ethics is not a homogeneous field, but a complex and heterogeneous cluster of claims, which does not admit of a systematic treatment. Nonetheless, there is a standing request for objective standards of correctness for evaluations, which in some domains may allow for weighting reasons (Nagel 1979, 180). A more optimistic view emerges as we distinguish levels and layers of disagreements. According to Scanlon (1995), this approach allows us to recognize broad areas of agreement. This is not to deny that there are radical disagreements in values, but they are not as large as the skeptics believe. When we clarify how we differ in values, we find large areas of consensus and we can further consider whether aiming to erase all sorts of value disagreement is morally appropriate.

How to decide on the latter issue, how to treat radical differences? Following Rawls, Scanlon holds that this is a matter for political philosophy, rather than for epistemology or ontology. The latter position has been argued extensively in political philosophy, especially as an argument for political liberalism. According to Rawls, reasonable citizens accept the burdens of judgment, and this is why they are inclined to tolerate disagreements, when this does not undermine or violate human rights. Reasonable citizens recognize that it is difficult, if not impossible, to settle disagreements of values, which are the root of political, religious, and moral disagreements. They do not have to deny that there ultimately are definitive answers about truthful values, but they have to admit that there are such serious epistemic limitations that tolerance of disagreement is the most reasonable option open to them. To coercively enforce or impose by repression, a value that is not shared would be a violation of freedom. In dealing with untreatable disagreements, Rawls addresses the problem of legitimacy within a liberal society, and his argumentation is based on the claim that the citizens of a liberal society would share not only the political institutions of a constitutional regime, but also the public traditions of their interpretation, as well as historic texts and documents that are common knowledge (Rawls 1993, 13–14). That is, they share a public political culture. In addition, they conceive of justification as a thoroughly public affair, that is, as a normative practice addressed to other citizens within the framework of a liberal society (Rawls 1993, 101). Apart from its relevance in debates about political legitimacy, this view is representative of a distinctive position in the epistemology of value disagreement. Rawls points out that the practice of normative justification governed by mutual respect and recognition has the potential of developing a free and reasoned agreement in judgment. Rawls' main concern in dealing with value disagreement is to legitimately ground the stability of pluralistic societies. However, the possibility of developing a basis of agreement in judgment is one of the foci of the debate about value pluralism. Agreement on some core values, such as respect of the dignity of others, is something that many identify as a sign of moral progress. But protecting value disagreement is also largely considered a sign of moral progress. Is it possible to defend these positions solely on political grounds, without taking sides about the nature of value?

6 Pluralism and Incommensurability

The fact of value disagreements requires a plausible explanation, and the theories of value offer competing ones. Monist theories of value reduce such disagreements to ignorance, epistemic vagueness, and bad reasoning, hence implying that there cannot be genuine disagreement about value. Pluralist theories, instead, take seriously the facts of value disagreement and try to account for plurality. In fact, how to account for plurality or the lack thereof is a challenge for both monism and pluralism. By discounting value disagreements, monist theories commit themselves to offer a plausible explanation of the sources of spurious disagreement, while pluralist theories commit themselves to explain how we can choose and behave rationally in contexts marked by pluralism. Neither task proves to be easy to accomplish. The strategies differ in relation to what we take value pluralism to be. Clearly, no theory can deny that subjects recognize different values and goods. However, pluralist theories take seriously the phenomenon of value disagreement, this plurality as irreducible, and account for such irreducibility on the basis of the claim that plural values are incommensurable. By contrast, monist theories hold that differences among values are superficial and can be explained away. This dispute over the sources and nature of value has significant normative implications, especially regarding the case of conflicts of values. To account for such implications, it is useful to distinguish different interpretations of value incommensurability.

In its stricter formulation, incommensurability is the claim that there is no common unit of measurement of value (Wiggins 1987; Stocker 1989, 175ff.; Chang 1997; Finnis 1980, 113ff; Wong 1989, 1992). Incommensurability comes in two varieties. *Weak incommensurability* holds that there is no single cardinal scale by which every value can be measured. *Strong incommensurability* holds that between any two particular values, there is no single unit by which they can be measured (See Wiggins 1987, 259; cf. Richardson 1995, 104–105). Weak incommensurability is often combined with the view that some values, such as human life and rights, have a special axiological and normative status (Anderson 1993, chs. 7–9; Lukes 1997). Supporters of this view remark that it is congruent with common valuing practices about the sacredness of life and treat some goods as radically different from commodities. Among such practices, there are procedures for establishing legal and moral constraints on the legitimacy of treating human life, personal relations, relations to non-human animals, and to the environment, along with marketable goods. For instance, this is the foundation of legal and moral arguments against slavery, prostitution, exploitation, and the trafficking of human organs.

This view carries two important implications. First, the claim that values do not differ in kind and are measurable by a single unit of value facilitates the decisions of public policies, but it undermines common evaluative practices and sensibilities, and thus, it requires a process of revision and adjustment. Second, the recognition of the irreducible varieties of values seems to require the development of an appropriate sensibility, receptiveness, and emotions (Stocker 1998). As Elizabeth Anderson remarks, different values demand different evaluative attitudes and responses from evaluators (Anderson 1993). To ensure that the recognition of the plurality of values be effective from a normative point of view, evaluators should be endowed with psychological resources apt to appropriately respond to this axiological variety (Bagnoli 2011; Deigh 2008; Slote 2010). This axiological and psychological complexity certainly complicates matters, and presumably it severely constrains the legitimacy of value transactions, but it also ensures that our lives are rich and nuanced. Vice versa, the claim that all values are ultimately the same and measured by a single unit of value simplifies our economic transactions, but it also impoverishes and flattens our emotional lives (Nussbaum 1990, 116–120; Anderson 1993).

An interesting section of this debate deals with a more practical dimension of incommensurability as it invests the life of any entity entitled to assess and capable of evaluating. For instance, Isaiah Berlin defines incommensurability as a claim about abstract values that cannot be jointly realized in the world (1969, 49–50, 53–54). On his reading, incommensurability amounts to *incompatibility*, in a given context. Insofar as it originates in concrete situations because of contingent features of the context of choice, it does not raise any logical issue about the incoherence of value system. Ronald Dworkin talks of incommensurability of specific instances of values of which one cannot say that they are always as good as, or better than, instances of other values. On this interpretation, incommensurability generates a phenomenon called *trumping, which happens when a certain sort of considerations overrides any other sort of consideration* (Dworkin 1977, xi). Trumping assumes incommensurability because it is not merely the case that some considerations are stronger than others. On the contrary, trumping occurs when a category of considerations overrides other considerations of a totally different sort. In particular, Dworkin holds that rights trump other considerations because they have a special normative force, which insulate them from trade-offs (Dworkin 1977). For instance, advocating the trumping power of the right to free expression, John Stuart Mill writes that one cannot be silenced by the majority as much as the majority cannot be silenced by one (Mill 1988, 20). A similar view, called *discontinuity* or *threshold lexical superiority*, is designed to capture cases in which some threshold amount of one value trumps any amount of the other value (Griffin 1986, 85).

These formulations of incommensurability serve well deontological ethical theories, where the categories of duty and right are supposed to be separated from the category of utility. For some philosophers, incommensurability is a *constitutiv*e feature of values such as respect for persons or friendship (Raz 1986, 345–357).

Kant distinguishes between value and price (Kant 1997, 4.432). Things have a price and thus are inter-substitutable and mutually fungible. These features warrant the commensurability of commodities, which is the basis of market relations. By contrast, dignity does not admit of measurement, and this blocks any form of compensation in kind when different persons are at stake. The opposite view is typically associated with monist utilitarianism, such as Jeremy Bentham's theory, which takes pleasure to be the only value, in terms of which all other values could be measured. More contemporary versions of utilitarianism take informed preference as the basis of such assessments, but they all adopt commensurability (Hare 1981; Harsany 1974). While commensurability is typically associated with consequentialism in its utilitarian variants, it is arguable that consequentialism may recognize at least weak incommensurability. For instance, some philosophers argue that while it is true that there is no rate of substitution among values and thus no general maximizing principle that can guide action, there is a general duty to bring about the best consequences (Finnis 1980, 113; Stocker 1998) . The content of duty is not uniform across contexts because the definition of what counts as the "best consequence" is relative to incommensurable values.

Strictly speaking, comparisons and rankings do not require commensurability. Therefore, the philosophically significant problem is comparability. Unfortunately, the claim of strict incommensurability (i.e., the lack of a cardinal unit by which values can be measured) is not always neatly distinguished from various failures of "incomparability" (i.e., lack of ranking relatively to a covering value). Considerations about how to rank values often merge with considerations about the limited cognitive and practical capacities of evaluators, which result in imprecision and indeterminacy. This idea turns on the claim that comparability is a matter of precise cardinal comparison. Derek Parfit holds that there are cases of imprecise cardinal comparisons (Parfit 1984, 431ff.). When incomparables are really indeterminately comparable, that is, it is neither true nor false that they stand in a positive value relation or "rough comparability" (Griffin 1986, 80–81, 96) or "vagueness" in comparison (Broome 1997). A general view of value that might explain the last three phenomena is that values are not determinate quantities but are metaphysically indeterminate (Chang 2002, 143–145).

Most philosophers in this debate hold that there are three positive value relations of comparison: "better than," "worse than," and "equally good." By contrast, Ruth Chang suggests that there is a fourth basic value relation, called "parity," which indicates a relation between two objects different than equality. Whether parity is really a distinct fourth value relation, not reducible to equality, is a debatable claim, but there are two important considerations in its support. First, parity seems indistinguishable from equality only on the presumption that values should be modeled on the relations among real numbers, which is a contestable claim (Chang 2002). Second, there are comparisons that do not seem to fit the standard tripartition outlined above. For instance, suppose an expert consultant is required to assess a policy regulating promotion in a research laboratory and concludes that: "*x* is a better policy than *z* because it is meritocratic." The consultant does not thereby imply that *x* is a better policy than *z*, absolutely. Her comparison is constrained by the context, which is limited to evaluating research and, therefore, assumes some values that are relevant in research assessment exercises. Were she required to compare the two policies in relation to another context, e.g., a public geriatric hospital, the consultant's judgment would likely be very different because it would be informed by other values, e.g., the values of health care. It would be grotesque to consider meritocracy as a dominant value in regulating patients' admission. The example shows that comparisons are made according to a range of considerations governed by substantive values, such as merit or fairness. Formally, then, comparability is a three-place relation: For any value *x* and *y*, *x* is comparable with *y* with respect to *V*, "a covering consideration" (Chang 1997). Also conversely, incomparability is a three-place relation: *x* is incomparable with *y* with respect to *V*, where *V* is a covering consideration. On this interpretation, then, incomparability does not amount to non-comparability, which holds when the formal conditions required for comparing values are not met.

Chang introduces an important complication in the relation among values, but does not challenge the basic assumption that deliberation and rational decision require comparing or commensurate options. A profound question is whether commensurability is a morally appropriate requisite for any transactions about value (Scanlon 1991). A radical view is that at least when some important values are at stake, comparability is a misplaced expectation and inappropriate method (Raz 1986, 322; Lukes 1997, 185–186). According to Steven Lukes, this method encourages an impoverished and sometimes even corrupted conception of value. To ask for comparison is a moral mistake, which aptly attracts blame. Conversely, refusal to compare shows that the evaluator correctly comprehends and understands the practice of value (Lukes 1997, 185–186; Raz 1986, 345–357). This radical view makes clear that the shared assumption about commensurability as a requirement for rational choice commits to commodification, that is, the claim that all sorts of values are like commodities.

The advantage of this approach is that practical reasoning about what to do becomes a form of calculation. An important example of the role of cardinality is the discussion of well-being, which spreads from ethics to welfarist economy. For instance, John Broome (1991) argues that how people's states of well-being (i.e., how well off they are) should be aggregated in order to determine the value of an overall distribution of well-being. According to "the interpersonal addition theorem," for a single group of people, one distribution of well-being is better than another if, and only if, the weighted total of the well-being of its members is greater. The same reasoning applies when we compare the values of distributions of well-being for different groups of the same size and then groups of different sizes; and, similarly, when we compare the states of well-being of a single person at different times (i.e., states of "temporal well-being"), in order to determine her overall, lifetime level of well-being. Broome concludes that the lifetime well-being is the sum of the values of all of one's states of temporal well-being, and he proposes that the value of a distribution of well-being for any population is the sum of the amounts by which the lifetime well-being of each person exceeds the neutral level for adding a life (Sidgwick 1907). Broome's demonstration requires that well-being can be measured cardinally. A problem typical of this approach is that it seems counterintuitive and unfair. For instance, it generates what Parfit names the repugnant conclusion, that is, that given a population of any size in which everyone enjoys an extremely high level of well-being, it would be better to have a much larger population of people, all of whose lives are barely worth living. Further qualifications about what it takes to have a life worth living may lessen the problem of unfairness, but do not solve it.

Are there compelling reasons to accept the claim about commensurability? Elizabeth Anderson argues that there are not. Commensurability is dispensable because it is not really useful. It does not make any sense to call objects good when they bear no relation with agents (Anderson 1997, 91; Slote 1989). Anderson defends a pragmatist theory according to which judgments of value are constructions of practical reason that guide practical reasoning. This is to say that judgments of value make sense within a practical domain, when we consider what to do. She thinks this conception commits to the view that we value things only insofar as they relate to us in some significant way, and also that we can justify or value judgments only pointing to practical functions (Anderson 1997, 91). A further implication of the pragmatist conception is that rational deliberation is never only about means, but always also about the ends. That is, values are always implicated when reasoning about what to do. This view has the merit of showing that discussions about the nature of value are strongly connected to rational choice and to issues such as the integrity and practical identity of agents.

By contrast, there are attempts to show that pluralism is an illusory phenomenon, which can be reduced to monism, without any loss of descriptive plausibility. Attempts of this kind typically distinguish between the subjective experience of values and the real ontological stance of value, e.g., a projectivist story about how values become part of the fabric of the world (Mackie 1977; Blackburn 1984, 1985). However, it is arguable that this reduction generates loss of descriptive plausibility. It is preferable a theory of value whose full-fledged epistemological and ontological story does not routinely and systematically discount subjective experience as illusory, but it is congruent with it.

On the basis of this argument, pluralists argue that monism is false to facts, while pluralism exhibits a high explanatory capacity. We have seen how it can explain the language of rights and the special value commonly attributed to personal relations. Furthermore, the pluralist claim helps us explain several predicaments of practical rationality. This is an important reason for placing incommensurability at the center of debates about the powers and limitations of practical reason. For instance, value incommensurability may be as one of the possible sources of *akrasia.* We generally take this case to be such that the agents do not conform to duty even when they know what they have to do. However, one other account of the phenomenon is that their reasoning does not fully determine what to do, but prescribes different and incompatible lines of action. The break in the line from the reason for action to action is not at the level of motivation, but it is situated earlier, in the contrasting values that inform the starting point of practical reasoning. The claim that the plurality of values blocks inter-substitutability helps explain why in the presence of different valuable ends, agents may respond with *akratic* behavior (Wiggins 1987, 239; Stocker 1989, 230ff.). It is debatable whether such phenomena are merely subjective illusions due to the cognitive and practical limitations of human psychology or instead depend on the ontological features of the value domain. The capacity to explain the phenomenology of valuing is certainly an asset of value pluralism. It may be objected that this is a consequence of its incapacity to guide choice.

7 Values and Rational Choice

The main significance of incomparability is that it threatens the possibility of rational choice. If two alternatives for choice are incomparable with respect to the values that matter in the choice between them, then, it is widely believed, there can be no rationally justified choice between them. Some admit of "existential plumping" for one alternative over the other (Chang 1997, 11; Broome 2000, 33–34). Incommensurability covers a wide range of phenomena where in case of conflicts, there is no principled way to rank the values at stake and no independent value that may be invoked as the umpire (Williams 1981). This view has three normative implications. First, from a normative point of view, it implies that value conflicts generate moral dilemmas where obligations clash, or practical dilemmas where reasons for action clash, and there is no resolution based on reason. In such dilemmatic contexts, decision always involves a loss in value. Second, this loss justifies and is congruent with the emotional experience of regret or guilt. Third, the phenomenology of choice is marked by attitudes and emotions that count as moral residue, even when there are attempts to trade off values, and within practices where compromise and compensations are legitimate.

Because of these profound implications for choice, some hold that the incommensurability of values undermines ethical theory as a theoretical and practical enterprise because it makes it impossible to provide a coherent and efficacious method of practical reasoning to help the agent determine what to do (Hare 1981). Value pluralism adds levels of complexity to monism, but it seems to undermine the rational basis for choice. A monistic account of value facilitates rational choice, but it oversimplifies the experience of valuing and it correspondingly impoverishes our evaluative life, seeming false to our experience as evaluators. The dispute between monists and pluralists concerning rational choice is articulated around the two meta-theoretical desiderata: normative determinacy, which is the alleged prerogative of monism, and descriptive plausibility and congruence, which are central for pluralism.

However, there are two orders of considerations for subverting either of these conclusions. The first order of considerations concerns the possibility of ranking incommensurable values so as to justify rational choice. Pluralists do not deny the possibility of ranking values, but argue that such orderings are not complete and admit of partial, dominant, and vague orderings (Sen and Williams 1982, 17). As mentioned in the previous section, most debates are based on the assumption that the lack of a cardinal scale for values prevents the comparison of instances of values. This assumption is false because the lack of a cardinal scale of measure does not entail incomparability (Bagnoli 2000, 2006; Chang 1997).

The second order of considerations for denying that pluralism undermines rational choice concerns the form of practical reasoning applicable in pluralist contexts. Pluralists have devised several strategies to rationally justify choice in pluralistic contexts, and all deploy rational deliberation. When incommensurability is defended as a feature of abstract values, the rational evaluator is required to deliberate further so as to make such values more specific. This deliberative strategy is called *specification.* Its effects are analogous to the effects of commensuration, but it does not require value commensurability and is advocated in contexts marked by strong incommensurability (Richardson 1995). Second, the method of *practical induction* suitably exploits the concrete practical experience of values. This method requires that the value dimension be considered across time and allows for increasing coherence among values over time (Millgram 1997, 151–184; Millgram 2002). The evaluator learns to assess his options over time. Deliberation makes options commensurable (Millgram 1997, 157–158). Coherence is thus an achievement of deliberation, rather than a formal property of values abstractly characterized. Attention to the historical dimension of value motivates a third approach to value, which attempts to order values by situating them in the context of one's entire life. While this strategy of *life-contextualization* does not lead to any principled view of reasoning, it grounds rational action on a broad and thick conception of agential integrity. If values fit together in the context of an entire life, this is a life with integrity (Taylor 1997, 179–180).

Practical integrity is also advocated as a moral standard. For Christine M. Korsgaard decisions should be respectful of the identities that we are *willing to reflectively endorse* as practical, that is, as those identities under which we attribute ourselves values. Such practical identities may clash, on some particular occasions, and the only guide we have from practical reasoning is that we must act on the basis of reasons that everybody can share. This requirement, akin to Kant's universalization, blocks the reasons that are immoral. This is to say that moral obligations rule out contrary considerations. However, this rational guide does not prevent the possibility of severe practical conflicts among special obligations that are rooted on our identities (Korsgaard 1996, ch. 4; Scanlon 2014). Finally, one may regard rational judgment as the locus of assessment of the normative relations among values. This strategy requires that we abandon the view that commensurability and incommensurability are hypotheses on the nature of value, and deal with them as *constitutive acts of evaluations*. Incommensurability is not a property of value that constrains reasoning, but the result of a judgment of comparative assessments, which articulates and organizes one's reasoned choice. Conversely, commensurability is the output of a successful rational deliberation, rather than its condition of possibility. Incommensurability is not an ontological feature of the value domain but a practical problem, which can be solved by engaging in deliberation. This approach to value requires a more complex view of deliberation.

As for normative determinacy, there are two related issues at stake. First, it is questionable that normative determinacy is a dominant requisite or a *desideratum* of any value theory. Second, it is questionable that commensurability and a fortiori comparability are sufficient to grant completeness in value ranking and normative determinacy of reasons for action. In other words, it is questionable that to determine what an agent ought to do we have to admit commensurability, and it is also questionable that commensurability suffices to determine what an agent ought to do. The case revolves around the relevance of so-called symmetrical dilemmas. Suppose Abe ought to financially support the synagogue and ought to financially support the museum, but he cannot afford donating to both institutions. If the deontic operator "ought" is agglomerative, then Abe ought to support either institutions, and he cannot support both, hence, he faces a dilemma. One may argue that Abe has only a disjunctive obligation: He chooses rationally if he chooses to support one of the two institutions (Herman 1993, 159–173). It would be irrational for him not to support either, but it is rational to support either one and it does not matter which one. But how does Abe decide which institution to support? The disjunctive obligation strategy leaves Abe with no decision procedure. More precisely, this strategy does not resolve the moral dilemmas, even though it indicates that the deliberative impasse can be overcome. Suppose Abe resolves to toss a coin (MacIntyre 1990). There are considerations of fairness that may guide this decision, but it is hardly the case that Abe's decision can be called fair. Perhaps one can say that it has a fair effect, but Abe's decision does not rest on any ground, nor is it chosen out of concern for fairness; hence, he cannot be judged as fair. In the case that the symmetrical choice concerns moral options, e.g., Abe has to choose between donating one instead of another, tossing a coin as a way to solve the conflict seems especially problematic, even when all deliberative routes have been explored. Some philosophers object that tossing a coin in such cases is an irresponsible act of self-indulgence (Rosalind 1999; Railton 1996, 153; Blackburn 1996, 129, 131). Others argue that the decision by randomization is arbitrary (Bagnoli 2006, 2013). These considerations support the view that symmetrical dilemmas are not trivial because of their symmetrical features. In fact, at least some symmetrical dilemmas are morally relevant hard choices. This shows that the monistic claim about commensurability does not warrant the sort of normative determinacy that many assume.

These cases are generally discounted as spurious or irrelevant on the assumption that, when there is no failure of commensurability, choice between symmetrical requirements is indifferent and can be determined by randomization. The appeal to randomization allows the agent to overcome a deliberative impasse, but it does not really resolve the moral dilemma. This is because randomization fails to provide the agent with a genuine decisive reason for action since reasoning does not fully determine nor explain our actions. Acting in such context is not irrational, but it does not count as a principled decision. This sort of arbitrariness may not be immoral because it may not result in unfairness or bias. However, arbitrary decisions of this kind do not fully express agency and authorship of action. Lack of authorship is a failure of agential authority over one's own action, but it is not a sign of irrationality or immorality, nor does it show a failure of value commensurability.

8 Persons and Values

Persons stand in a special relation to values because they are both bearers and sources of value. This claim carries important normative and deontic implications. On a widespread view, persons have value insofar as they are persons, and they are sources of value insofar as they are persons. Correspondingly, there are evaluative practices and attitudes directed to persons as valuable, and evaluative activities by which persons assign values to other objects. Furthermore, insofar as persons are values, they are also sources of valid claims on others (Rawls 1980a, b, 452). In its turn, this implicates that there are constraints on how to relate to persons, how we should treat them, how we should express our feelings toward them, and so on. Persons carry a special relation to values insofar as they are both loci of value or value bearers, and also sources of values. How to explain this complex sort of relation is a matter of dispute. One focus of this dispute concerns the features that make persons distinctive sources of values. The other one concerns the sense in which persons are sources of values, if they produce, create, or recognize. This is an important principle, but it is not uncontested.

The main normative implication of the claim that persons are independent sources of valid claims, reasons, and values is that they are separate individuals. The separateness of persons is the grounding reason for prohibiting trade-offs. More generally, interpersonal comparisons violate the separateness of persons (Nozick 1974, 33). However, disproportion in number is also, and very generally, taken to be a ground for resolving conflicts. Both claims are deeply rooted in our commonsensical approach to value. The crucial case is a conflict between action that affects one person and action that affects many. In this sort of deliberation, it seems that numbers matter. There is a large agreement between normative theory and ordinary moral judgment that it is preferable to save the many, for instance. If numbers are relevant, though, it is because there are quantitative comparisons across persons. Arguably, there are separate reasons to save each person, since the moral value of each person is the same, and these reasons somehow add up and result in the obligation to save the many. If persons have equal significance, then the presence of each additional person should make a difference (Kamm 1989, 240–241).

Why are persons unique bearers of value? Arguably, this is because of some (metaphysical or natural) features. Kant's view is that persons have a distinct value or dignity insofar as they have autonomy, which is a metaphysical property of the will and entitles them to exact respect from others (Kant 1997, 4.412; Kant 1996, 6.211–6.213, 227). A contemporary (and partial) rendering of this view is that persons are values because they are self-reflective, hence capable of rational and critical assessment (Nozick 1981; Frankfurt 1988; Korsgaard 1996). In this reading, reflexivity does not necessarily set humans apart from other animals (and other possible rational beings). For utilitarians, humans are morally significant insofar as they are sentient beings. To be a person does not add anything in terms of value and it is not a distinct category of value, and the personal identity of persons is irrelevant to the metaphysics of value (Parfit 1984). Persons have no special and distinct value insofar as they are persons, but only and to the extent that they are receptacles of utility. By contrast, for deontologist or a constructivist, the concept "person" indicates a normative status, which is warranted through practices of recognition, and it carries normative and deontic implications (e.g., moral claims and responsibilities). Among such implications there is respect for boundaries across persons. On this reading, then, the claim that persons have equal standing does not mean that they are commensurable items of equal value. For instance, the Kantian view is that equality of status entails that persons ought *not* to be treated as mere equivalents. Since persons have no equivalents, any exchange in value placed upon persons is morally prohibited (Kant 1997, 4.432).

9 Values and Emotions

Values stand in a complex relation with the most receptive aspect of practical rationality, which includes emotions, attitudes, and desires. These relations are complicated by the fact that there is no philosophical agreement about the concepts involved. Insofar as emotions are broadly understood as containing some conative states, along with desires, they have been often used in order to clarify the nature of evaluation and its motivational impact. For instance, emotivism holds that an evaluative judgment is not an assertion about a state of affairs or a property of an object, but expressive of emotions (Ayer 1936). As it appears, this schema of analysis assumes that the concept of emotion is clear enough to be able to explain the complex act of evaluation. However, there is a profound disagreement about what an emotion is, and whether it necessarily includes a conative state. The variety of emotions and the growing literature on the complex relation they bear with perception and reasoning strongly suggest that this analysis is doomed to failure.

However, there is a superficial agreement that at least some so-called moral and deontic emotions and values are interestingly connected. For sentimentalist theories, emotions and feelings are the source of moral judgment and also, with some corrective that pushes toward the general point of view, the cement of social life. This view is also supported by some evolutionary theorists, insisting that natural selection favors cooperative values and emotions of mutual support, recognition, care, and love may have a decisive, even though instrumental, role. According to the perceptualist theory of value, emotions are like perceptual judgments, which allow us to detect value in the world. This view was first proposed by Scheler (1954), who argued that emotions are perceptions of "tertiary qualities," which depend upon facts about social relations, pleasure and pain, and natural psychological facts. A similar view is currently articulated by Mulligan and Tappolet (2000). Emotions such as love, compassion, care, and forgiveness shape widely shared notions of moral value. Some authors analyze this role in terms of "response dependence" and argue that emotions are responses that depend on the values and norms that lie at the core of the moral life (McDowell 1985; D'Arms and Jacobson 2000).

The relation between morality and the emotions is problematic (Bagnoli 2011). On the one hand, emotions stand in the way of moral conduct, insofar as they provide independent motivations that undermine or compete with moral motives. On the other hand, emotions seem to be necessary to have moral understanding. According to the ethics of virtue, emotions such as love and compassion are natural dispositions, which when properly habituated and educated develop into virtues of character (Baier 1985; Doris 2002; Slote 2010; Smith et al. 1989). For others, deontic moral emotions such as guilt and remorse attend the violation of duties, explain normative behavior, and signal the capacity to be bound by norms (Hare 1981, ch. 2). Emotions such as shame and pride, instead, seem to be crucial modes of valuing the self and expose individuals to the gaze of others, hence showing how the value of identity is profoundly influenced by social criteria of membership. Furthermore, the presence of emotions has been used to recover conflicts of values, and tensions between values of membership and individualistic values. For instance, genuine moral emotions and affections drive Huckleberry Finn against his judgment to abide by the law and turn Jim, the escaped slave, into the authorities (Bennett 1974; McIntyre 1990). Emotions such as love and compassion here convey attention to values that are sanctioned by socially enforced moral standards, revealing a more authentic moral understanding and attachment. On the realist theory about value, Huckleberry apprehends the moral values of human fellowship and freedom via emotional acquaintance. According to Mulligan (1998), emotions justify axiological judgments and beliefs, even though they are not direct perceptions of value. This approach seems well placed to explain how emotions further value and help moral life. However, there are significant ontological and epistemological objections against this view, which is criticized for overemphasizing the analogy with perceptual judgment. By contrast, appraisal theories of emotions hold that emotions contain an evaluative thought, but not necessarily a belief or a cognitive judgment about the case. A key point of relating values to emotions is to develop a theory of virtue, which takes emotions as natural dispositions that can be shaped by education and habituation into competences.

Recent debates have focused on the role of emotions in practical and epistemic reasoning. In traditional views, emotions are often regarded as disturbances and interferences. By contrast, empirical studies show that emotions positively contribute to reasoning and at various levels. Their basic function is to call attention to details of the situation that matter; hence, they work as criteria of salience, which help generate reasons for action and reasons for belief. Many have also started to notice and study the role of values in theoretical reasoning, and to identify epistemic values (Pritchard 2007; Haddock et al. 2009; Williams 2002; Zagzebski 2004).

10 Valuing

Valuing is a complex activity, which concerns large varieties of objects, including properties, events, states of affairs, activities, practices, attitudes, and persons. In short, it seems that anything can be the object of assessment. Does this indicate that anything can be regarded as valuable and that there are no boundaries to what can be treated as value? As anticipated in Sect. 4, non-cognitivists and cognitivists differ in answering this question. However, they might agree that the activity of valuing admits to some constraints, even though they might disagree about what they are and how stringent. Non-cognitivists such as Hare (1963, 1981) and Stevenson (1979) hold that the constraints are so meager that any factual contents can be combined with a positive assessment, so as to identify value. If valuing is mainly an emotional attribution, even logical consistency may not apply. In fact, the varieties of pluralism, then, would be akin to ambivalences and other peculiarities of psychological lives. Arguably, the regularities and patterns that we register in recording the kinds of values cherished in the course of human history might be best explained by (evolutionary) psychology, rather than by logic or ontology (Gibbard 1990; Nozick 2001).

By contrast, cognitivists such as G. E. M. Anscombe argue that one cannot rationally value a saucer of mud (1957, 70) without indicating any rationale for supporting this preference. The rationale would be a characterization of the object in terms of its desirability; absent such characterization, the agent's valuing rests on no grounds and it is criticizable as irrational. Arguing toward a similar conclusion, Derek Parfit presents the case of somebody who cares equally about what happens every day of the week but lacks any concern for future Tuesday, conforming to a principle named "future Tuesday indifference" (Parfit 1984, 254). Singling out present Tuesday as the focus of one's valuing is irrational because it is arbitrary, lacking any justifying reasons.

The philosophical analysis of this complex activity highlights both rational and emotional components, but it is an outstanding question whether such components can be separate and, more importantly, whether we gain a better understanding of the practices of valuing by decomposing values in factual and non-factual components (Murdoch 2013). Borrowing from Murdoch, Hilary Putnam (2002, 28–45) has argued against the fact/value dichotomy on the ground that it is based on a poor understanding of evaluative language. This is a misunderstanding rooted in the empiricist tradition, which adopts a very narrow view of facts, and a very simplistic account of moral psychology. On the opposite view, facts and values are inevitably entangled. While non-cognitivism views description as devoid of values, Murdoch further suggests that one chief mode of assessing the world and deliberating about what to do is redescribing it. Moral agents become objective by constant efforts of attention, by which they attempt to redescribe reality as accurately as they can.

Murdoch's account vindicates a crucial aspect of valuing as a practice and activity importantly historical. The temporality of moral agency and of valuing is something that both realist and anti-realist have found hard to appreciate and explain. Valuing is not an occasional activity of human beings, and it does not appear to be something that we engage and disengage from at will. In fact, the very activity of valuing, both in its individual and social dimensions, seems to be profoundly related to the fact that we are temporal beings, rooted in the past, hooked in the present. Our valuing attitudes and preferences are sensitive to temporal constraints. Philosophers have identified several cases of temporal bias, in which agents discount the value of their options according to how they are situated in time. Our reasons for valuing seem to be driven by concerns that are sensitive to time. The practice of valuing thus intersects another philosophical debate about prudence and rationality over time. Philosophers disagree about whether there are or there should be normative criteria for assessing the rationality of our evaluative activities across time (Nagel 1979; Parfit 1984; Bratman 2009; Korsgaard 2008; Hedden 2015 ; Sidgwick 1907).

Furthermore, and more radically, philosophers have identified the temporality of agency as the main rationale for entering valuing practices and activities. Our subjective experience of life as temporally bounded connects crucially with what makes life worthwhile. Bernard Williams discusses the Makropulos' case—the case of an immortal but boring life, to show how the meaning of life is constrained by finitude. This is partly because the objects and desires that make a life worth living are finite and exhaustible (Williams 1973). Along these lines, Samuel Scheffler argues that what we care and value most—e.g., love and labor, intimacy and achievement, and solidarity—"have the status of *values* for us because of their role in our finite and bounded lives" (Scheffler 2013, 100). Current debates show that the issue of temporality and value is still to be placed on a clearly intelligible framework.

A promising approach is informed by the conviction that valuing is a rational activity, hardly reducible to the expression of preferences more or less intense. It is a complex activity not only because it involves a complex network of emotional and cognitive capacities, but also because it admits of various intertwined modes. We value in different ways, through varieties of evaluative judgments (e.g., along dimensions such as desirability, or reasonableness), attitudes (e.g., love, admiration, and respect), practices (ranking, mutual respect and recognition), and institutions (e.g., the market). The plurality of the modalities of valuing should be investigated not only by differentiating categories of values (e.g., intrinsic, extrinsic, instrumental, categorical), but also by understanding different (institutional and individual) modes in which we attribute and confer value in our life. A pluralist approach does not only allow us to recognize plurality of values, but it also encourages us to construct new arguments for comparing values and assessing the ethical limitations of institutions such as the market, beyond the traditional methods of welfare economics and traditional theories of justice, such as the cost–benefit analysis. Such a pluralist approach appears promising especially in consideration of the challenges faced by governmental institutions, which are required to take action in the presence of divisive conflicts of values and under uncertainty. For instance, in the debate of global warming, governments are required to take action under uncertainty, while appreciating values as diverse as safety, financial interest, and moral obligations to future generations. While some emphasize the complexity of weighing lives through times (Broome 2004; Hedden 2015), others argue that weighing is not the appropriate way of approaching problems of rational choice, precisely because the contexts in which we choose are profoundly marked by value pluralism (Anderson 1993). This is one dramatic example of the account that the nature of value has a direct, practical impact in our lives, and affects not only the quality of the present life of our co-habitants, but also the future of life on the planet.

References

Adams, R.M. 2006. *A theory of virtue: Excellence in being for the good*. Oxford: Oxford University Press.

Anderson, E. 1993. *Value in ethics and economics*. Cambridge, Mass.: Harvard University Press.

Anderson, E. 1997. Practical reason and incommensurable goods. In *Incommensurability, incomparability and practical reason*, ed. R. Chang, 98–101. Cambridge, Mass.: Harvard University Press.

Anscombe, G.E.M. 1957. *Intention*. Oxford: Blackwell.

Anscombe, G.E.M. 1958. Modern moral philosophy. *Philosophy* 33: 1–19.

Aristotle. 1984. *The complete works of Aristotle*. Princeton, N.J.: Princeton University Press.

Audi, R. 2001. *The architecture of reason: The structure and substance of rationality*. Oxford: Oxford University Press.

Audi, R. 2005. *The good in the right: A theory of intuition and intrinsic value*. Princeton, N.J.: Princeton University Press.

Ayer, A.J. 1936. *Language, truth and logic*. London: V. Gollancz Ltd.

Bagnoli, C. 2000. Value in the guise of regret. *Philosophical Explorations* 3: 165–187.

Bagnoli, C. 2006. Breaking ties: The significance of choice in symmetrical moral dilemmas. *Dialectica* 60: 1–14.

Bagnoli, C. (ed.). 2011. *Morality and the emotions*. Oxford: Oxford University Press.

Bagnoli, C. (ed.). 2013. *Constructivism in ethics*. Cambridge: Cambridge University Press.

Baier, A. 1985. *Postures of the mind: Essays on mind and morals*. Minneapolis, Minn.: University of Minnesota Press.

Bennett, J. 1974. The Conscience of huckleberry finn. *Philosophy* 49 (188): 123–134.

Blackburn, S. 1984. *Spreading the world*. Oxford: Clarendon Press.

Blackburn, S. 1985. Errors and the phenomenology of value. In *Morality and objectivity*, ed. Ted Honderich. London: Routledge.

Blackburn, S. 1993. *Essays in quasi-realism*. New York, N.Y.: Oxford University Press.

Blackburn, S. 1996. Dilemmas: Dithering, plumping, and grief. In *Moral dilemmas and moral theory*, ed. H. Mason, 127–139. New York: Oxford University Press.

Blackburn, S. 1998. *Ruling passions*. Oxford: Oxford University Press.

Boyd, R. 1988. How to be a Moral Realist. In *Essays on moral realism*, ed. G. Sayre-McCord, pp 181–228. Cornell University Press.

Brandt, R.B. 1996. *Facts, values, and morality*. Cambridge University Press.

Brink, D.O. 1984. Moral realism and the sceptical arguments from disagreement and queerness. *Australasian Journal of Philosophy*, 62(2): 111–125.

Broome, J. 1991. *Weighing goods: Equality, uncertainty and time*. Oxford: Blackwell.

Broome, J. 1997. Is incommensurability vagueness? In *Incommensurability, incomparability and practical reason*, ed. R. Chang, 67–89. Cambridge, Mass.: Harvard University Press.

Broome, J. 2000. Incommensurable values. In *Well-being and morality: Essays in honour of James Griffin*, ed. R. Crisp, and B. Hooker, 21–38. Oxford: Clarendon Press.

Broome, J. 2004. *Weighing lives*. Oxford: Oxford University Press.

Broome, J. 2013. *Rationality through reasoning*. Oxford: Wiley-Blackwell.

Chang, R. 1997. Introduction. In *Incommensurability, incomparability, and practical reasoning*, ed. R. Chang. Cambridge, Mass.: Harvard University Press.

Chang, R. 2002. *Making comparisons count*. London-New York: Routledge.

D'Arms, J., and D. Jacobson. 2000. Sentiment and value. *Ethics* 110: 722–748.

Dancy, J. 2000. *Practical reality*. Oxford: Oxford University Press.

Dancy, J. 2004. *Ethics without principles*. Oxford: Oxford University Press.

Darwall, S., A. Gibbard, and P Railton (eds.). 1996. *Moral discourse and practice*. Oxford University Press USA.

Deigh, J. 2008. *Emotion, values, and the law*. New York, N.Y.: Oxford University Press.

Doris, J. 2002. *Lack of character: Personality and moral behavior*. Cambridge: Cambridge University Press.

Dworkin, R. 1977. Taking rights seriously. *Philosophical Quarterly* 27 (109): 379–380.

Finnis, J. 1980. *Natural law and natural rights*. Oxford: Oxford University Press.

Firth, R. 1951. Ethical absolutism and the ideal observer. *Philosophy and Phenomenological Research* 12 (3): 317–345.

Foot, P. 2000. *Natural goodness*. Oxford: Clarendon Press.

Frankena, W.K. 1939. The naturalistic fallacy. *Mind* 48: 464–477.

Geach, P. 1956. Good and evil. *Analysis* 17: 33–42.

Geach, P. 1960. Ascriptvism. *Philosophical Review* 69: 221–225.

Gibbard, A. 1990. *Wise choices, apt feelings: A theory of normative judgment*. Harvard University Press.

Gibbard, A., and A. Macintyre. 1995. The viability of moral theory. *Philosophy and Phenomenological Research* 55: 343–356.

Griffin, J.1986. *Well-being: Its meaning, measurement and moral importance*. Clarendon Press.

Haddock, A., A. Millar, and D. Pritchard (eds.). 2009. *Epistemic value*. Oxford: Oxford University Press.

Hare, R.M. 1952. *The language of morals*. Oxford: Clarendon Press.

Hare, R.M. 1963. *Freedom and reason*. Oxford: Clarendon Press.

Hare, R.M. 1981. *Moral thinking: Its levels, method, and point*. Oxford: Oxford University Press.

Harman, G. 1977. *The nature of morality: An introduction to ethics*. Oxford: Oxford University Press.

Harman, G. 2000. *Explaining value and other essays in moral philosophy*. Oxford University Press.

Harman, G., and J.J. Thomson. 1996. Moral relativism and moral objectivity. *Philosophy* 71: 622–624.

Hedden, B. 2015. *Reasons without persons: Rationality, identity, and time*. Oxford University Press UK.

Herman, B. 1993. *The practice of moral judgment*. Harvard University Press.

Hume, D. 1739. *A treatise of human nature*. Oxford University Press.

Hurka, T. 1993. *Perfectionism*. New York: Oxford University Press.

Rosalind, H. (1999). *On virtue ethics*. Oxford University Press.

Hurka, T. 2000. *Virtue, vice and value*. New York, N.Y.: Oxford University Press.

Joyce, R. 2001. *The myth of morality*. Cambridge University Press.

Kamm, F. 1989. Harming some to save others. *Philosophical Studies* 57: 227–260.

Kant, I. 1996. *Metaphysics of morals*, trans. Mary Gregor, 1st ed, 1797. Cambridge: Cambridge University Press.

Kant, I. 1997. *Groundwork for the orals*, trans. Mary Gregor, 1st ed, 1785. Cambridge: Cambridge University Press.

Kolodny, N. 2003. Love as valuing a relationship. *Philosophical Review* 112: 135–189.

Kolodny, N. 2005. Why be rational? *Mind* 114: 509–563.

Korsgaard, C. 1983. Two distinctions in goodness. *Philosophical Review* 92: 169–195.

Korsgaard, C. 1996. *The sources of normativity*, ed. O. O'Neill. Cambridge: Cambridge University.

Korsgaard, C. 2008. *The constitution of agency: Essays on practical reason and moral psychology*. Oxford: Oxford University Press.

Lukes, S. 1997. Comparing the incomparable: Trade offs and sacrifices. In *Incommensurability, incomparability and practical reason*, ed. R. Chang, 185–186. Cambridge, Mass.: Harvard University Press.

Lyons, D. 1976. Ethical relativism and the problem of incoherence. *Ethics* 86: 107–121.

MacIntyre, A. 1988. *Whose justice? Which rationality?* Notre Dame, Minn.: University of Notre Dame Press.

MacIntyre, A. 1990. Moral dilemmas. *Philosophy and Phenomenological Research* 50: 367–382.

MacIntyre, A. 2007. *After virtue: A study in moral theory*. Notre Dame, Minn.: University of Notre Dame Press.

MacIntyre, A. 2008. Value and context: The nature of moral and political knowledge. *Journal of Moral Philosophy* 5: 151–154.

Mackie, J.L. 1977. *Ethics: Inventing right and wrong*. New York, N.Y.: Penguin.

McDowell J. 1985. Values and secondary qualities. In *Morality and Objectivity*, ed. Ted Honderich. London: Routledge. pp. 110–129.

Mill, J.S. 1988. Utilitarianism. In *Collected works of John Stuart Mill*, 1st ed, 1861, vol. 29, ed. J.M. Robson, 371–577. Toronto: University of Toronto Press.

Millgram E. 1997. *Practical induction*. Harvard University Press.

Millgram E. 2002. Commensurability in perspective. *Topoi* 21 (1–2): 217–226.

Moore, G.E. 2004. *Principia Ethica*, 1st ed, 1903. Mineola, NY: Dover Publications.

Moore, G.E. 2010. *Philosophical studies*, 1st ed, 1921. London: Routledge.

Mulligan, K. 1998. From Appropriate emotions to values. *Monist* 81: 161–188.

Murdoch, I. 2013. *The Sovereignty of good*, 1st ed, 1971. London: Routledge.

Nagel, T. 1979. *The possibility of altruism*. Princeton: Princeton University Press.

Nagel, T. 1986. *The view from nowhere*. Oxford: Oxford University Press.

Nichols, S. 2004. *Sentimental rules: On the natural foundations of moral judgment*. Oxford: Oxford University Press.

Nozick, R. 1974. *Anarchy, State and Utopia*. Oxford: Blackwell.

Nozick, R. 1981. *Philosophical explanations*. Cambridge, Mass.: Harvard University Press.

Nozick R. 2001. *Invariances: The structure of the objective world*. Belknap Press of Harvard University Press

Nussbaum M.C. 1990. *Love's knowledge*. Oxford University Press.

O'Neill, O. 1989. *Constructing authorities: Reason, politics and interpretation in Kant's philosophy*. Cambridge: Cambridge University Press.

Parfit, D. 1984. *Reasons and persons*. Oxford: Oxford University Press.

Parfit, D. 2006. Normativity. *Oxford Studies in Metaethics* 1: 325–380.

Parfit, D. 2011. *On what matters*. Oxford: Oxford University Press.

Plato. 1991. *Euthyphro*, ed. Chris Emlyn-Jones. Bristol: Bristol: Bristol Classical Press.

Prinz, J.J. 2007. *The emotional construction of morals*. Oxford: Oxford University Press.

Pritchard, D. 2007. Recent work on epistemic value. *American Philosophical Quarterly* 44: 85–110.

Putnam, H. 2002. *The collapse of the fact/value dichotomy and other essays*. Cambridge, Mass.: Harvard University Press.

Railton 1996, The Diversity of moral dilemma. In *Moral Dilemmas and Moral Theory*, ed. Mason, H. E. (1996). Oxford University Press.

Rawls, J. 1980a. Construction and Objectivity. *Journal of Philosophy* 77 (9): 554–572.

Rawls, J. 1980b. Kantian constructivism in moral theory. *Journal of Philosophy* 77 (9): 515–572.

Rawls, J. 1993. *Political liberalism*. Columbia University Press.

Rawls, J. 1994. Political Liberalism. Philosophical Quarterly 44 (177):542-545.

Raz, J. 1986. *The morality of freedom*. Oxford: Clarendon Press.

Richardson H.S. 1995. Beyond good and right: Toward a constructive ethical pragmatism. *Philosophy and Public Affairs* 24 (2): 108–141.

Richardson, H.S. 1999. Institutionally divided moral responsibility. *Social Philosophy and Policy* 16 (2): 218.

Richardson, H.S. 2013. Moral reasoning. *The Stanford Encyclopedia of Philosophy*.

Ross, W. D. 1930. *The right and the good*. Clarendon Press.

Scanlon, T.M. 1991. The Moral basis of interpersonal comparisons. In *Interpersonal comparisons of well-being*, ed. Jon Elster and John E. Roemer, 17–44. Cambridge University Press.

Scanlon, T.M. 1995. Moral theory: Understanding and disagreement. *Philosophy and Phenomenological Research* 55: 343–356.

Scanlon, T.M. 1998. *What we owe to each other*. Cambridge, Mass.: Belknap Press of Harvard University Press.

Scanlon, T.M. 2014. *Being realistic about reasons*. Oxford: Oxford University Press.

Scheler, M. 1954. *The nature of sympathy*. Hamden, Conn.: Archon.

Scheffler, S. 2013. *Death and the afterlife*. Oup Usa.

Schroeder, M.A. 2009. *Slaves of the passions*. Oxford: Oxford University Press.

Schroeder, M.A. 2010. *Noncognitivism in ethics*. London: Routledge.

Schroeder, M.A. 2011. Moral sentimentalism. *Philosophical Review* 120: 452–455.

Sen, A., and B. Williams. 1982. Introduction: Utilitarianism and beyond. In *Utilitarianism and beyond*, ed. A. Sen, and B. Williams. Cambridge: Cambridge University Press.

Shafer-Landau, R. 1994. Ethical disagreement, ethical objectivism and moral indeterminacy. *Philosophy and Phenomenological Research* 54: 331–344.

Sidgwick, H. 1907. *The methods of ethics*. Indianapolis, Ind.: Hackett.

Skorupski, J. 1993. The definition of morality. *Royal Institute of Philosophy Supplement* 35: 121–144.

Skorupski, J. 2010. *The domain of reasons*. Oxford University Press.

Slote, M. 1989. *Beyond optimizing*. Cambridge, Mass: Harvard University Press.

Slote, M. 1992. *From morality to virtue*. Oxford: Oxford University Press.

Slote, M. 2010. *Moral sentimentalism*. Oxford: Oxford University Press.

Smith, M., D. Lewis, and M. Johnston. 1989. Dispositional theories of value. *Proceedings of the Aristotelian Society* 63: 89–174.

Smith, M. 1994. *The moral problem*. Oxford: Blackwell.

Stevenson, C.L. 1937. The emotive meaning of ethical terms. *Mind* 46: 14–31.

Stevenson, C.L. 1963. *Facts and values*. New Haven, Conn.: Yale University Press.

Stevenson, C.L. 1979. *Ethics and language*. Norwalk, Conn.: Ams Press.

Stevenson, C.L. 2009. The nature of ethical disagreement. In *Exploring philosophy: An introductory anthology*, ed. S.M. Cahn. Oxford: Oxford University Press.

Stocker, M. 1989. *Plural and conflicting values*. Oxford University Press.

Stocker, M. 1998. Emotions. How emotions reveal value and help cure the schizophrenia of modern ethical theories. In *How should one live?: essays on the virtues*, ed. Roger Crisp. Clarendon Press.

Sturgeon, N. 1998. Moral explanations. In *Ethical theory 1: The question of objectivity*, ed. James Rachels. Oxford University Press.

Sturgeon, N.L. 1986. Harman on moral explanations of natural facts. *Southern Journal of Philosophy* 24 (S1): 69–78.

Tappolet, C. 2000. *Emotions et Valeurs*. Paris: Presses Universitaires de France.

Taylor, C. 1997. Leading a Life. In *Incommensurability, incomparability, and practical reasoning*, ed. R., Chang. Cambridge, Mass.: Harvard University Press. chapter 9.

Taylor, C. 1976. Responsibility for self. In *The Identities of persons*, ed. Amelie oksenberg rorty, pp. 281–299. University of California Press.

Taylor, C. 1989. *Sources of the self: The making of the modern identity*. Harvard University Press.

Thompson, M. 2008. *Life and action: Elementary structures of practice and practical thought*. Cambridge, Mass.: Harvard University Press.

Toulmin, S.E. 1950. *An examination of the place of reason in ethics*. Cambridge University Press.

Velleman, J.D. 2009. *How we get along*. Cambridge: Cambridge University Press.

Velleman, J.D. 2013. *Foundations for moral relativism*. Cambridge: OpenBook Publishers.

Wiggins, D. 1987. *A sensible subjectivism?*. Blackwell.

Williams, B. 1973. The Makropulos case: Reflections on the tedium of immortality. In Id., *Problems of the Self.* Cambridge: Cambridge University Press.

Williams, B. 1981. *Moral luck*. Cambridge: Cambridge University Press.

Williams, B. 1985. *Ethics and the limits of philosophy*. Cambridge, MA: Harvard University Press.

Williams, B. 2002. *Truth and truthfulness: An essay in genealogy*. Princeton, N.J.: Princeton University Press.

Wong, D. 1986. On moral realism without foundations. *Southern Journal of Philosophy* 24: 95–113.

Wong, D. 1989. Three kinds of incommensurability. In *Relativism: Interpretation and confrontation*, ed. M. Krausz, 140–158. Notre Dame, Ind.: Notre Dame University Press.

Wong, D. 1992. Coping with moral conflict and ambiguity. *Ethics* 102: 763–784.

Wong, D. 2006. *Natural moralities: A defense of pluralistic relativism*. Oxford: Oxford University Press.

Wong, D. 2008. Constructing normative objectivity in ethics. *Social Philosophy and Policy* 25: 237–266.

Zagzebski, L. 2004. Epistemic value and the primacy of what we care about. *Philosophical Papers* 33: 353–377.

The Goals of Norms

Cristiano Castelfranchi

Premise: The Foundational and Intrinsic Relation between N and G

The relationship between norms (Ns) and goals is not simply important but foundational, since *Ns are artifacts for social coordination through agents' goal manipulation*.

This relation is also multifaced and multifunctional; these faces and functions—in order to understand what Ns are and how they work—should be explicitly analyzed and explained.

Obviously, our object is "norms" in the "normative" (prescriptive) meaning/sense, not in the "normality" (descriptive or statistic or standard sense). However, there is an important and bidirectional goal relation between N1 (in the normative sense) and N2 (in the normality sense):

(a) N2 creates and becomes a goal for the actors and even an N1 (a prescription, something "due"), in order to conform, to be like others. This conformity is either a need of the individual or a need (and request/pressure) of the group, or both.

(b) N1 creates an N2, a normal conduct in the community, if it is respected: N conformity is "normal."

Moreover:

(b1) N1 has the goal and function to be respected and thus to create an N2, a normal behavior (at the individual, internal level this helps it to also become an automatic response, just a habit);

(b2) If N1 does not become/create an N2, it is weakened (Bicchieri 2006; Conte and Castelfranchi 1995), perceived as less credible and less binding.

C. Castelfranchi (✉)
Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche (CNR), Rome, Italy
e-mail: cristiano.castelfranchi@istc.cnr.it

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_7

Coming back to Ns as behavior regulation devices, we have:

On the one side, Ns are *only* for autonomous *goal-directed* agents, whose behavior depends on their goals/choices, preferences, that is, on their free decisions, and are thus assumed to be "responsible" for their action, since they could behave differently.

On the other side, norms "have goals" in the sense that they are built and used "for" something. They are interactional and societal *tools*.

But they have goal and are finalistic in another sense, too: They guarantee certain social emerging *functions*. This is a different teleological notion and a different kind of "goal." It is crucial to disentangle these notions and to have a theory of their structural and functional relations.

A good theory (and study) of Ns should be explicit in distinguishing and factually exploring the differences and relations between:

1. The intention (intended effects) of N; what we expect and want to produce by the issuing and monitoring a given N;
2. The actual effects of N, and in particular:

   2a. *N efficacy*, that is, the actual/realized intended effect (corresponding to its goals or unsatisfactory); such *efficacy* or failure clearly is evaluated with respect to the "goal" of N: (1).

   2b. *Functional effects*, usually unintended (desirable or even undesirable) and not understood, but such that they have feedback and select and reproduce that behavior or entity, including a given N; the outcome (in part) is responsible for the maintenance or reproduction of N.[1]

   2c. *Side effects*, neither intended nor functional but systematic, and either negative (as with addictive drugs) or positive (relative to some other goals not in the plan of N).

In other words, Ns have to do with *two kinds of teleology*: (i) mentally represented (and perhaps even intended) or *psychological goals* that regulate our conduct and (ii) nonmental goals, merely emergent and self-organizing "functions" (social or biological) impinging on our individual and collective behaviors. Let us use the term *goal* only for the internal control system, the mentally represented objective, and the term *function* only for the external selecting finality of a feature or behavior.

Given this distinction between two kinds of finality impinging on norms, the question is: What are the dialectics between them? More precisely:

(a) Is the goal of N the goal (G) represented in the mind of "subject" S, and driving her behavior? Is it at least the G in the mind of the "issuer"?
(b) Should S understand and pursue the G of the N while obeying it?

As for (a), the answer is: only partially; the goal induced in the S's mind and adopted by her is only a subgoal of N, not its aim in the mind of the authority nor its social function.

[1] Consider, for example, Marx's claim that prisons also reproduce themselves by reproducing delinquency; which of course is not the mission or the aim of prisons! But in a sense is a "bad *function*" of them.

And as for (b), the answer is: not at all; on the contrary, the ideal "obedience" and subject are without a cooperative sharing of the aim of N and in any case are not motivated by such explicit collaboration.[2]

So we have to examine what a goal is; what a goal-directed behavior is; how goals regulate human behavior; how Ns are *made to* induce goals and regulating our behavior, but also what the "functions" are and what the relation is between the goals of the "subject," the goal imposed by N in the subject's mind, the issuer's goal, and the function of N.

1 Goals–Norms: A Multiple Relation

Six are the main structural relations between Ns and Gs:

(A) Ns are communicative artifacts for "social manipulation," designed to influence autonomous, goal-directed behavior *by inducing or blocking goals* in a given set of addresses. *Ns are goal-oriented tools for the goal-directed behavior of autonomous agents.* They want to enter our preferences and decisions or bypass them by building a habit, an automatic conform response.

(B) Ns presuppose and need the *postulation of goals in the minds of subjects*; they are grounded in an "intentional stance" and the attribution of mental states and in particular of autonomous behavior internally regulated by goals and beliefs and then by "choices/decisions."[3]

(C) Ns are aimed at governing our conduct through our goal mechanism: at "regulating" our behavior by "influencing" us, that is, by *inducing* goals in us, goals to be *adopted*, so that they become internal and possibly prevail on personal/private motives.

Either by inducing us to do something or by creating a conflict and inhibiting a "wrong" behavior on our part. Norms *reduce* the subject's choices—the possible goals to be pursued—and/or *add* new goals to be pursued.

They have to "give" us new goals (or new reasons for a goal) and to block other possible goals we may have. Finally, they have to generate the *intention* regulating our action.

(D) Ns are for goal adoption (*adhesion*). Since we are autonomous agents (governed and motivated by our own (internal) motives), the N-goal must be internalized and adopted for some goal we may have. But N has the goal (and function) and (cl)aim to be adopted for specific goals and reasons. The ideal-typical adhesion to an N is for an intrinsic motivation, for a "sense of duty," recognition of

[2] When driving a car, you "have to" use a safety belt even if you disagree about this prescription and use: It is not the case that you have to use one on condition that you agree that it is better for you or for the costs placed on the community.

[3] It is useless to create "norms/laws" for animals; better to use orders and threats, or physical barriers. In humans, mental barriers can work, and frequently even physical barriers primarily have a signaling, communication, function.

authority, because it is right/correct to respect Ns. Only subideally should one respect Ns, in order to avoid external or internal sanctions. Normative education also goes in this direction.

Only after the goal adoption, there is "true" (intentional/aware) "violation" (disobedience) not just a behavioral violation.

Eventually, norm-conforming behavior can be proceduralized, automatized, routinized; however, on the one side, there is the implicit knowledge that it is a norm; on the other side, it socially counts *"as if"* were a deliberate and intentional act (responsibility ascription).

(E) Ns work thanks to the distributed, collective, expectation about others' goal of conformity, that is, a collective and mutual "prescription" (*goals about the goals* of others), and monitoring and sanctioning.

(F) Norms are "tools for" something: *They have a goal, and they are aimed at producing a given result* (coordination of actions and interests, social order, power distribution, reference rules, and trust). They have a "goal" (the goal of the "legislator," of the normative authority, possibly to be also understood and interpreted by the judge and the monitoring guy but not necessarily understood or intended by the subject (see later)) and some function. The "function" of Ns is not necessarily understood and intended, and not necessarily a good one and corresponding to the norm's official end/mission.

We will try to illustrate these faces of the Ns–goals relation.

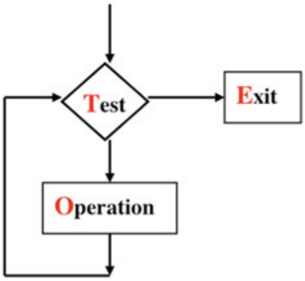
2 Teleologies of Mind: Goals, Functions, and Pseudogoals

2.1 What Are Goals?

In modern science, there are two well-defined teleological frames and notions:

- The one provided by *evolutionary approaches*, where it is standard (and correct) to speak in terms of functions, (adaptive) value, being *for* something, having a certain finality/end, providing some advantage, etc. In this context, *goal* (end, function, finality, etc.) means the "effect" (outcome) that has selected/reproduced and maintained a certain feature or behavior: initially just an accidental effect, an effect among many others, but later, thanks to the loop and positive feedback on its own "causes" (i.e., on the feature or behavior producing it) no longer a mere effect but the "function," the purpose of that feature, what makes it useful and justifies its reproduction.
- The one provided by *cybernetic control theory* and its postulated and representations cycle, in which the agent is able to adjust the world through goal-directed behavior, and to maintain a given "desired" state of the world (homeostasis), as illustrated in Fig. 1.

Fig. 1 TOTE cycle



Actually, there might be a third teleological/finalistic notion used in several sciences (from medicine to the social sciences): the notion of a "function" of X as a "role," a functional component, an "organ" of a global "system"; for example the "function" of the heart or of the kidneys in our body, or the function of families (or of education or of norms) in a society, or the function of a given office in an organization. However, this "functionalist" and "systemic" notion has never been well defined and has elicited a lot of problems and criticisms. My view is that this finalistic view is correct, but it is reducible to, and derived from, the previous two kinds of teleology. Either the "organs" are the result of an evolutionary selection—in that they contribute to the fitness and reproduction (maintenance) of that organism—or there is a "project," a "design," that is, a complex goal in someone else's mind, a goal which imposes particular subgoals on its parts, components, and tools, or both.

2.2 The Relations Between Psychological Goals and Behavior Functions

A serious problem for a (future) science of goals is that these two fundamental teleological notions/mechanisms have never been unified:

i. Neither conceptually, by looking for a common definition, a conceptual common kernel (e.g., in terms of circular causality, feedback): Do we have and is it possible to have a general, single notion of "goal" with two subkinds (functions vs. psychological goals)?
ii. Nor by solving the problem of the interaction between the two coexisting forms of finality.

This constitutes a serious obstacle and reveals a real ignorance gap in contemporary science.[4]

[4] As for issue (i), without the aforementioned conceptual unification, we cannot have a unitary theory of communication—or a theory of cooperation, of sociality, etc.—in animal and humans. What are today presented as unified theories are just a trick. In fact, these notions—which necessarily require a goal (e.g., "communication" requires not only a sign-"reader" but also a "sender": The information

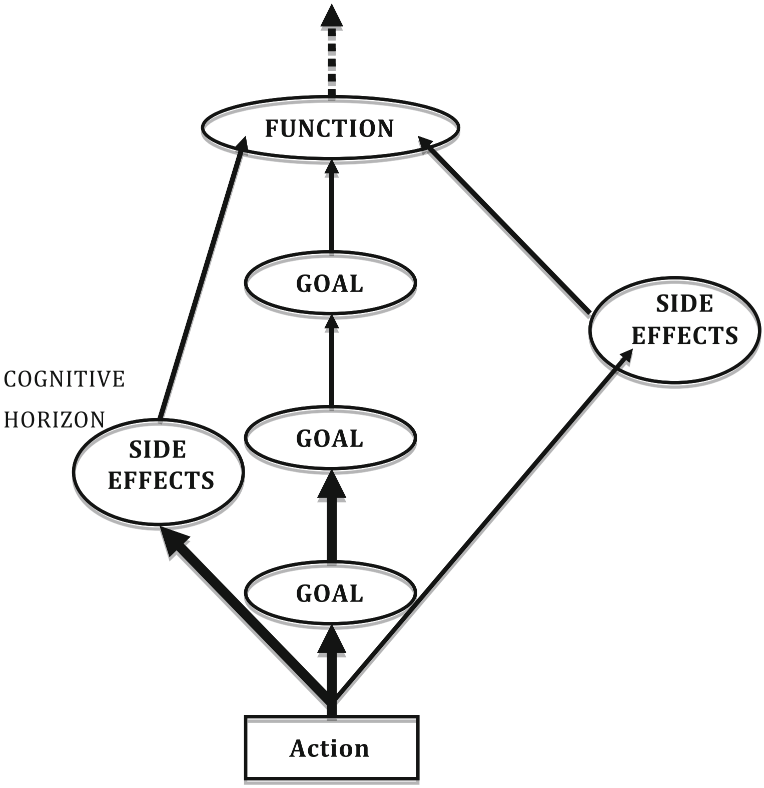


Fig. 2 Mental goals and possible functions

As for point (ii): *What is the relationship between the internally represented goals (motivations and concrete objectives) of an agent regulating its behaviors from the inside and the adaptive functions that have selected that agent and its behaviors?*

Usually, in purposive, goal-driven agents/systems, the "function" of their conduct, the adaptive result that has to be guaranteed, is *not* internally represented and pursued; it is neither understood nor foreseen (Fig. 2). Of course, not all foreseen outcomes or all side effects have a "function."

The internal motivations (and whatever solutions and instrumental goals they generate) are just subgoals of the "external" goals of the behavior, of its functions;

is deliberately "given" to the "receiver/addressee")—are defined in terms of adaptive functions when applied to simple animals (like insects), whereas in humans they are defined in intentional terms. Thus, there is no unified notion (or theory) of "communication," in that we do not know the common kernel between a "functional" device and an "intentional" device. A remarkable attempt to deal with these problems is Ruth Millikan's work.

they are just "cognitive mediators" of the (biological or social) functions that would be nonrepresentable and mentally noncomputable.[5]

2.3 Goals Versus Pseudogoals

It is also very important to disentangle true goals from *pseudogoals* (Castelfranchi 2012), that is, goals that only seem to be there and to regulate the system and its behavior. However, they are not in fact there as goal mechanisms: They are not represented in and "governing" the system. They are just functional ways in which the system has been "designed" (by evolution, by learning, by the designer); they are the system's goal-oriented way of working, its operational rules. For example, a real thermostatic system (thermostat, thermometer, room, radiator, boiler, etc.) has been designed in order to reduce naphtha consumption, heat loss, etc., as much as possible. These are (pseudo) goals of the system, which also works to guarantee them, but they are not true cybernetic goals like the thermostat's set point. They are not represented, evaluated, and "pursued" by the system action cycle.

Analogously, our minds have been shaped (by natural selection, or culture and learning) in order to have certain working principles and to guarantee certain functions, which are not explicitly represented and intended. It seems (from our behavior) that we have certain goals, but they are not real goals, only pseudogoals. This is the case, in our view, with some well-known (and badly misunderstood) finalistic notions, like utility maximization, cognitive coherence, and even pleasure. No doubt, we often choose between different possible goals so as to maximize our expected utility, giving precedence/preference to the greater expected value, that is obvious and adaptive. However, this does not mean that we have *the* goal (the single and monarchic goal) of maximizing our utility, regardless of specific contents and goods. On the contrary, we are moved and motivated by specific, qualitative terminal goals we want to achieve (esteem, sex, power, love, etc.), but the *mechanism* that has to manage them has been designed and works so that it maximizes expected utility.

In the same vein, we maintain coherence among our beliefs, and need to avoid and eliminate contradictions. That is why we can reject certain information and do not believe all the data we get (sometimes even what we directly perceive; "we do not believe our eyes," literally); the new data must be "plausible," credible, integrable, within the context of our preexisting knowledge; otherwise, we have to revise our previous beliefs on the basis of new (credible) data. This coherence maintenance is frequently completely automatic and routine. We have mechanisms for checking and

[5] For example, only very recently have we discovered why we have to eat, the real functions/effects of our food in our organisms (proteins, carbohydrates, vitamins, etc.), and very few people eat in view of such effects. We eat for hunger or pleasure or out of habit. Analogously, we do not usually engage in courtship and sex in view of reproduction: We are driven by other internal motives. We can even cut off the "adaptive" connection between our motives and their original functions, as by deciding to have sex without inseminating or without establishing/maintaining any friendly/affective relation or support.

adjusting coherence. We do not usually have any real intention about the coherence of what we believe. Thus, knowledge coherence is a pseudogoal of ours, not a real meta-goal guiding meta-actions.

2.4 Subjective Kinds of Goals

- Neither *goals* nor *motives* mean “desires.” Desires are just one kind of goal. Desires are endogenous (and usually pleasant), and with Ns, we just have to cut off some possible course of action by *making some desire of the S practically impossible or nonconvenient*. It is ignored that *“duties” are not “desires”;* they are *goals from a different source*, with a different origin: They come from the outside (*exogenous*),[6] and they are imported, “adopted”; they are “prescriptions” and “imperatives” from another agent.
- Not all goals have to be “actively pursued” or just “pursued”; some of them (like having a sunny day) are not within our power: Whether they can be realized is not up to us but depends on other “agents” or external forces, so we cannot really “pursue” them. Other goals are such that their realization is only partly up to us: We have something to do with it, but then the final result depends on others or on luck, an example being winning a lottery or being acquitted of a crime we did not commit. As was previously pointed out, a goal is not a goal only if/when actively pursued.

  Thus, we may have *actively* pursued goals (goals pursued through our active actions), but also merely *passive* goals, and the latter can be of two very different kinds:

  – Goals we can only wait for, hoping that they will be attained, for they do not depend on us at all: We cannot do anything (else).
  – Goals whose realization depends on us and on our “doing nothing,” that is, on our abstaining from possible interference. We would have the power to block that event/result, and we decide to do nothing, in order to let it happen (inaction, “passive action”). This case also involves our “responsibility,” since the result is due to our decision and (in)action.

- An important distinction in motivational theory is that between avoidance goals and approach goals. This is how Wikipedia summarizes the difference: “Not all goals are directed towards *approaching* a desirable outcome (e.g., demonstrating competence). Goals can also be directed towards *avoiding* an undesirable outcome” (for scholarly discussion, see, for instance, Elliot 2006). More than that, avoidance and approach represent two mental frames, two different psychological

[6] However, see the later discussion on the internalization of “authority” and on internal moral imperatives.

dispositions and mind-sets (see Higgins's 1997 avoidance and approach "regulatory *focus*").[7]

In avoidance goals, success is to pass from Q to not Q (to end the negative state) or to prevent passage from not Q to Q, while in approach goals, success is to pass from not P to P or from P to P (i.e., maintaining P as the desired status quo).[8]

- Not all our goals are "felt," in part because not all of them are represented and defined in a sensory-motor format.[9] The two most important kinds of felt goals are *desires* and *needs*.

  In the "felt need" for a given object O, we perceive a current, unpleasant, or disturbing bodily or affective stimulus S (for instance, we perceive dryness in our throat when we "*feel* the need for water") that we cognitively ascribe to the loss of O. Similarly, in felt desires, we just "imagine" and anticipate the pleasant sensations/emotions that we will/would have if/when attaining our cherished object.
- Intentions are those goals that *actually drive our voluntary actions or are ready/prepared to drive them*. They are not another "primitive" (like in the BDI model inspired by Bratman's theory: see, e.g., Rao and Georgeff 1995), a mental object different from goals. They are just a kind of goal: the final stage of successful goal processing, which also includes "desires" in the broad sense,[10] with very specific and relevant properties (see also Castelfranchi and Paglieri 2007; Castelfranchi et al. 2007).

In a nutshell, in our model, an *intention* is a goal that

(1) has been activated and processed;
(2) has been evaluated as not impossible, and not already realized or self-realizing (achieved by another agent), and whose achievement is therefore *up to us*: We have to act in order to achieve it[11];

[7] Notice how this terminology (e.g., "approach") is related to a semantics/connotation of "goal" that was criticized in Sect. 3, point A, as being too strongly inspired and constrained by behaviorism. Moreover, many motivational theories about avoidance and approach (such as Higgins's) remain essentially hedonistic.

[8] Another important difference is between gradable and all-or-nothing goals, or between achievement and maintenance goals. But these distinctions here are less relevant (Castelfranchi 2012).

[9] This means that we cannot say, for example, "I feel the intention of…"—for the simple reason that sensory-motor format of the represented anticipatory state is not specified in the notion of intention. Intention is a more "abstract" notion of goal, with an unspecified codification. Looking at a goal as an "intention," we abstract away from its possible sensory components.

[10] The creation of two distinct "primitives," basic independent notions/objects (desires vs. intentions), is in part due to the wrong choice of adopting "desires" (also in accordance with common sense) as the basic motivational category and source. We have already criticized this reductive move (Sect. 4) and introduced a more general and basic (and not fully common sense) teleonomic notion. This notion also favors a better unification of kinds of goals and a better theory of their structural and dynamic relationships.

[11] An intention is always an intention to "do something" (including inactions). We cannot really have intentions about the actions of other autonomous agents. When we say something like "I have the intention that John go to Naples," what we actually mean is "I have the intention *to bring it about that* John goes to Naples."

(3) has been chosen against other possible active and conflicting goals, and we have "decided" to pursue it;
(4) is consistent with other intentions of ours; a simple goal can be contradictory, inconsistent with other goals, but once chosen, it becomes an intention and has to be coherent with our other intentions (Castelfranchi and Paglieri 2007; Castelfranchi et al. 2007)[12];
(5) implies the agent's belief that she knows (or will/can know) how to achieve it, that she is able to perform the needed actions, and that there are or will be the needed conditions for the intention's realization; at least, the agent believes that she will be able and in a condition to "try";
(6) is being "chosen" implies a "commitment" with ourselves, a mortgage on our future decisions; intentions have priority over new possible competing goals and are more persistent than the latter (Bratman 1987);
(7) is "planned"; we allocate/reserve some resources (means, time, etc.) to it, and we have formulated or decided to formulate a plan consisting of the actions to be performed in order to achieve it. An intention is essentially a two-layer structure:

    (a) the "intention *that*," the *aim,* that is, the original processed goal (e.g., to be in Naples tomorrow) and
    (b) the "intention *to do*," the subgoals, the planned executive actions (to take the train, buy the tickets, go to the station, etc.). There is no "intention" without (more or less) specified actions to be performed, and there is no intention without a motivating outcome of such action(s);

(8) thus, an intention is the final product of a successful goal processing that leads to a goal-driven behavior.

After a decision to act, an intention is already there even if the concrete actions are not fully specified or are not yet being executed, because some condition for its execution is not currently present. Intentions can be found in two final and prefinal stages:

(a) *intention "in action,"* that is, guiding the executive "intentional" action;
(b) *intention "in agenda"* ("future-directed," those more central to the theories of Bratman, Searle, and others), that is, already planned and waiting for some lacking condition for their execution: time, money, skills, etc. For example, I may have the intention to go to Capri next Easter (the implementation of my "desire" of spending Easter in Capri), but now it is February 17, and I am not going to Capri or doing anything to that end; I have just decided to do so at the right moment; it is already in my "agenda" ("things that I have to do") and binds my resources and future decisions.[13]

[12] Decision-making serves precisely the function of selecting those goals that are feasible and coherent with one another, and allocating resources and planning one's actual behavior.

[13] I would also say that an "intention" is "conscious": We are aware of our intentions, and we "deliberate" about them; however, the problem of unconscious goal-driven behavior is open and quite complex (see Bargh et al. 2001).

Ns are aimed at producing (through a choice) *intentions* in the subject; on the basis of given beliefs: "Is it really an N? Is it valid and respected? Is it my case?… Are there other conflicting, more important Ns?… Am I able and in the condition?…"

This is the cognitive *N-processing* in minds (Castelfranchi 2013; Conte et al. 2010).

3 Features of the Goal of an N

3.1 Impersonal

The goal of "ordering"/"prohibiting," of issuing an imperative, cannot be personal/private; you have to play your role and to implement, or actuate a goal of the role; so your will cannot to be your will but an "institutional" normative will.

It is the goal "of the authority," that is "who" we consider "*entitled*" (meta-norm) to issue an N: God; the group or community, the impersonal anonymous "we" and "one," "nobody," etc.; the boss or leader; the institutionalized authority (the king, Parliament), and an intrinsic part of acknowledging and treating that impinging will as an N (and thus as a deontic obligation) is to recognize the goal as a nonpersonal and rightful one.

So an N and a simple prescription, imperative, or impositive request differ not by the "object" ("Do not smoke," "Close the door!") by for the required motivation, and so, in a sense, by their "content," since it is part of the content of the "linguistic" (or communicative) normative act (issuing or instancing N) to prescribe specific motives (goals) for adhering and for doing. We have to not just behaviorally conform, or adopt the goal and formulate the intention: We have to do that for specific "reasons" (recognition), for specific higher "motives": obedience, sense of duty, respect, and so on.

Sanction avoidance, the penalty, is not the *goal* that should motivate (drive) your adhesion and choice (the goal of the *sanction*—if any, apart from blame—is fairness, that you pay for your abuse, and that you and others be sent a message confirming N, its monitoring and equity: those who commit a wrong, who do harm (a public good like social order, or other people) have to pay.

"Sanctions" (penalties) are also for "learning" to pay "attention," to be aware of the existence of N and of its context and circumstances, and to not violate N out of unawareness. We have to become normative–attentive, and this is done by meta-Ns and imperatives ("Be responsible," "Be careful") and by shocks and learning.

3.2 Avoidance

As noted, Ns want to create a specific kind of goal, an *avoidance goal*. Or, better yet, they want us to perceive that imperative in an avoidance focus; it is a matter of "framing": We have to frame the situation from the perspective of coercion, danger, duties, possible harms, and possible sanctions ("Prevention focus": Higgins 1997). This is because in this frame the decision is more coercive, and the prescription more effective.

In our view, the prospect of punishment and of sanctions, condemnation, violation, or being "bad" is precisely meant to place us in this cognitive and emotional "frame"; its purpose is more educational and communicative than practical and "economic," in part because we cannot detect and punish all deviations, and we need a strong *internal* control and precaution and later a feeling of guilt, as well as internal persecution and punishment.

The conflict between two avoidance goals, or between an avoidance goal and an approach goal, makes us feel a sense of astriction: We chose between a threat, a worry, and something else (negative or positive).

In fact, if the N has to eventually create or come into conflict with some other non-normative goal, it has to win. That is why the Ng takes the (psycho)logical form of an "obligation." Obligations are more stringent, cogent, this for two reasons:

(A) They are shaped in terms of "ought," that is, of necessity; that is, there are no alternatives, this is the only way: this or nothing (violation). It is not simply *useful* and within your choice; it is "imperative" in both senses. Even the technical "ought" is conceived as a necessity: "if you want x you have to y (if you cannot, you will not succeed)." In the deontic, normative "ought," this necessity perspective, this lack of alternative is even stronger, since—given that you don't know or understand or do not have to care about N's "end"—it is not up to you to see whether there may be some alternative way.

(B) If you do not do x, there will be a harm, a penalty, a bad situation for somebody and for you, something to be avoided.

This is why, in order to make the impinging goal more prescriptive and coercive, we make commitments and pacts even with ourselves. So we are bound, we are "in debt" and bound by duty to ourselves.

3.3 Meta-goals

N introduces a goal to do or not to do a given action, a goal that has to be "adhered to," that is, adopted because the source has the goal that you adopt it. So Ns are *meta-goals*: goals about a goal of yours, about a goal you *have to* adopt and pursue. And they want and have to win/prevail against your *possible* conflicting goals; there is a presupposition (like in any influencing action) of *possible* (and assumed) conflict with your autonomous goals; we do not suppose that you already and independently have that goal and would pursue it independently of N. However, this is not so crucial, since it makes a difference whether you do something for your personal motives or whether you engage in the same outward behavior but as an application of and respect for an N. The second case is markedly different, not by reason of possible sanctions but in virtue of the real value and functioning of N. N is not a statistics of a given behavior; N is effective when that behavior is *olive to* the recognition of N as an N, to respect for the will of an authority, and you are also motivated by that (for moral principles *or* to avoid its sanctions).

So the goal of N is not just that you do something, but that you perceive/receive the N imperative, recognize it, and adopt it, and that you conform to it, and for that reason behave accordingly. Sometimes what matters is the subject's *obedience*, not the real content of the "order": to see whether he is submitted.

Ns do not necessarily entail a conflict in the subject. At any rate, the N usually creates some problem, being in conflict with other goals[14]: N can be in conflict with desires ("Do not covet your neighbor's wife"), with drives and needs ("Do not steal even if you are hungry"), with impulses ("Do not kill, even if you are furious"), or with other N-goals (conflict between Ns or their instantiations).[15]

Even in this case—involving N against other goals or one N against another N—the conflict can be of two kinds, or origins: logical contradiction (opposite goals) and resource scarcity (competing goals).

Moreover, in the decision-making process, the N-goal is subject to exactly the same scrutiny and steps as the other goals in what concerns the decision. For example, "I prefer to conform, but… are there the conditions and resources for that? Am I skilled, able? Are there possible side effects in this context to be avoided?"

3.4 Origin and Base of Norms: Norms Come from the Social Goals to Be Adopted

To better understand the meta-goal and adhesion-based nature of N, let us reflect on their forerunners. The origin and forerunner of Ns is not the frequency of a behavior or a hierarchy and authority, etc., but is a social goal: *my goal about your behavior*, and in particular (with socio-cognitive agents) *my goal about your goal*—a goal that impinges on you in that *you are expected to "adopt"* it (or, better yet, to "adhere" to my "request/prescription," a meta-goal).

The real origin of an *N as an N* is not just an expectation, a prediction (this is the origin of trust): It is *my social meta-goal* and our searching for some tool for influencing you, for inducing you to adopt that goal. The transition from mere expectation to (implicit or explicit) *prescription* is the step preparing "norms" in a deontic sense (Castelfranchi and Tummolini 2003; Castelfranchi et al. 2007). Real "authority" (and not just "force" based on fear or submission) is built on Ns, not the other way around.

[14] Or, better yet, the content goal derived from the N is in conflict with other goals I have.

[15] Notice that conflict is additional proof that an N is a goal and a goal source. In fact, conflict is a specific property of goals (of any kind: desires, drives, impulses, needs, projects, plans, intentions, values, etc.).

That is why *for an N be an N* (perceived and treated as such), I have to perceive the "prescription/request," not as your "personal" (private) request or goal, but as anonymous, impersonal, coming from the "collective" or from the representative or institutions of the collective, and it should not be addressed to *me*, as an individual, a specific person, but to a class of subjects, who may happen to consist of only one person, me; or to me as a role player.

4 The Relationship Between the Mental and External Goals of Ns

4.1 Norm Functions and Goals

As we said, frequently, the "functions" of an N are not fully understood (we may not be fully aware of them) and thus may not be intended (or at least may be "passively" intended or "accepted"): They may be unconsciously or unwillingly "pursued." Sometimes, there is even a paradoxical function (a kako-function, on which see note 2): a bad result, contrary to our subjective (individual or collective) goals, and even in contrast with the official objective of the N, but contributing to its reproduction and not just occasional or accidental.

There is also an important distinction here between two different kinds of Ns: ones that are explicitly issued, "deliberated" (by appropriate procedure and roles), like legal Ns or the official "rules" in an organization, versus conventional norms, emerging and established by tacit negotiation among participants. In the first case there is surely an intended result, a subjective goal of N. In *conventions*, this is less clear: Since nobody "decides" about N, and nobody necessarily decides its aim, or what it should guarantee, G is not explicitly in the minds of the agents.

Nevertheless, sometimes the G is understood (or supposed to be) and may also be approved, and we do not just obey N but "collaborate" with it. For example, the N prescribing that you do not stick our fingers in your nose in public has the G of not disturbing or disgusting others; we understand the meaning and aim of that N and respect it, or we respect it because we agree about that G or we do so in order not to be blamed, or simply for the sake of obedience and conformity with Ns, or both.

There is also a nonintended function of these kinds of Ns, which is to establish manners and rules of politeness, in such a way as to select people, give them a status and a membership, distinguish between levels in the population, preserve traditions, etc. Frequently, this function is really more important than the specific contents of politeness Ns, which can lack any sense, being just a ritual behavior.

An example of a nonintended (good) F of social Ns and customs[16] is the *reduction of uncertainty*, the right frame for reading events and activating the right (expected)

[16] These are not just frequent and regular behaviors but have a prescriptive component: People not only expect but want us to behave conformingly, and they critically react to any "violation" we may commit (Castelfranchi and Tummolini 2003).

behaviors (Garfinkel 1963); this reduction of subjective uncertainty, but also of cognitive and decision-making costs, and of negotiation costs, this "coordination" is an F of any social convention, habit, or script. Any community enjoys these benefits, and its customs, games, and scripts remain and are inherited and replicated for that benefit, too. But we do not *intend* that: We play that game for the specific results we obtain for our goals, not for maintaining the social order, trust, or uncertainty reduction.

Another F of any N—also in a strict sense (laws, moral Ns, social Ns, formulated in terms of an obligation or prohibition)—is *to teach and learn to "obey," as such*, any authority's will, any recognized N. It is a fundamental mental attitude, which makes us "social," acculturated: It is a fundamental cognitive and motivational basis of collective activity. We have to obey N and authority *as N* and *as authority*, not on condition that and because we understand and agree about N's goal. Our task is to "recognize" N and thus obey it, and by doing so, we send out a "signal," a message that that is an N, that we respect N and the authority, and that they should be respected. We do not intend that F of our behavior, but it is there and it reproduces and spreads it.

There are also *intended functions* of Ns; for example, we intend the (illusory?) effect of a stronger sanction for a given crime in order to dissuade people from that crime. N as the explicit goal of establishing that "if X commits that crime, s/he has to pay x amount of money, or has to serve a prison sentence of x years," and that these provisions be applied. But we also wish that an effect of this N be the reduction of that crime, and we change the law with this objective. This objective/G is not something that somebody "has to do," what is prescribed by N: The objective is some expected consequence—that is, the real aim of N. The punishment is increased only instrumentally in view of that outcome. Analogously, we establish Ns for giving right of way in traffic, etc., but we intend the general (expected) emerging effect: a regulation of traffic, reducing accidents and conflicts, a safer speed, making clear who is "in the right" in the event of an accident, etc. This is not what is prescribed (first-level goal) but is an *intended* F of it, and is also a G of it (at least in the mind of the issuer).

As noted there also are kako-functions: bad results that systematically reproduce N and a given behavior. What is (supposed to be) the goal and/or function of laws (and thus of legislators, of government)? (i) Social order and its maintenance or (ii) the "common good" and protection of the "commons." Imprimis (ii); (i) only if/since it is a "common good," but not if it favors a minority, some privilege, or class domination. Objectively, however, (i) the normative order (social, economic) is a subordination in favor of the dominant interests and groups. Ns are introduced or changed *in order to protect interests*: Every political, economic, or civil N protects some interest and subject. Who has the power to obtain the "right" N from the authority?

4.2 Subgoals

There is an obvious relationship between N's functional aim (F), its public effect/mission, and the G that Y (the subject) had to internalize and pursue (Gn): Gn is necessarily a subgoal for realizing F; the expected outcome of an action realizing Gn is F or part or a condition for it. Ns are aimed at planting in our mind—for regulating our behavior—a subgoal of the intended outcome (which is supposed to be a public good). For the formal (enounced, proclaimed) and official N (legal or not), there is usually a good overlap between N's intended goal and its F in the group/community, and we can even "adjust" N to its (ascribed) results. There is frequently a partial overlap, a partial explicitation of N's goal hierarchy: We issue N (and respect it) "for" G, which is a subgoal in the full goal "chain": goals for good Fs.

Society acts like evolution! On account of both our bounded rationality and limited computational resources and our personal preferences, we cannot understand and calculate the final, very high, and long-term and "complex" outcomes of our behaviors: their "fitness." This is why evolution and culture "terminalize" (as final motives) some subgoals of the real F, like in the relationship between evolutionary fitness and our internal "motivations," like in the creation of social "values" that had to be noninstrumentally justified as ends in themselves. For the same reason, Ns have to be obeyed (even) without understanding or sharing their aim. Society and culture reproduce the evolutionary trick.[17]

4.3 The Subject and N's Aim

As noted, N does not presuppose our understanding and sharing (pursuing) the *aim of N* (for the "subjects," it is just a presupposed F). But it does presuppose, or require, some belief that there is a goal, and more precisely a common end or value, and that the N authority (the issuer) is pursuing that goal. We have to *trust* the issuer and his playing a "tutoring" role.

We are "obliged" to obey even if we do not agree on N's goal. But we have to trust the issuer for his *intention* and role to do something "good," not self-interested. He may be wrong, but he should not be abusing of his power. If this is not the case, our impulse is not just to violate N but is stronger, because he is not playing his role, and his harming me/us. We feel entitled to rebel.

This "agreement" (common goal) about the Polis, Ns, authority, etc., is the background for issuing and respecting Ns. Some (not fully conscious) delegation, or

[17] N's other generic "function"—the restatement of a normative system, of authority, of submission—is also ensured by an internalized subgoal: the goal of adopting N (given its recognition as an N, which is another goal and subfunction of N).

empowerment (Gelati et al. 2004), and "alienation"[18] is intrinsic in a real normative process.

5 Concluding Remarks

Norms are artifacts, tools for manipulating human conduct through the manipulation of our goals and preferences/choices. It is impossible to understand the efficacy and working of norms without a modeling of *how Ns work in our mind, how they succeed in regulating our behavior from within*, and how do they give us goals. They are built for that.

However, Ns also have goals (they are aimed at achieving certain social outcomes), have effects (including unintended ones) and have functions. We do not understand and intend all the functions of Ns, and the subject is not supposed or requested to understand even all the goals of Ns and to obey on condition that she agrees and cooperates.

Let us conclude with a nice paradox of Ns.

N's goals/functions and their possible reversal. What is the real goal or function of a given N1? That we do not violate it, that our behavior conforms to N1; or that we be punished, as by paying a fine (norm N2)?

N2 (and its application) should in principle be only a means, a secondary and instrumental N designed to protect the realization of N1, that is, its goal. It is a meta-norm about the possible violation of another N, but functional to its *effectiveness*, not to its violation. But this *instrumental relation* can be reversed, and the means become the end, and the end merely a means, an excuse. I can issue N1 with the expectation (and goal!) that you violate it, so I can punish you. This is the attitude of some bad parents, but this is also, for example, the use of speed limits in some local government in Italy: They set an unreasonable speed limit *in order* to have a lot of people (not local citizens) violate it while traveling through their territory, *in order* to gain a hefty income from their fines. Hence, *N is issued in order for it to be violated!*

Even the subject can in some sense reverse the right goal relation, by interpreting and using the punishment (such as a fine) as just an additional cost for the possibility of exploiting the violation of N, and they decide to pay in order to be free to violate N1.[19]

[18] Meaning that the subject alienates his own intellectual evaluative, problem-solving, decision-making capabilities by "delegating" them to others, along with the power and the solution. Moreover, he is not in a condition to realize that, to understand this process, and behaves *without recognizing* his own alienated powers and without the possibility of reappropriating them. He has to be blind and to adopt N blindfolded.

[19] These are the famous findings of Uri Gneezy and Aldo Rustichini (2000).

References

Bargh, J., P. Gollwitzer, A. Lee-Chai, K. Barndollar, and R. Trötschel. 2001. The automated will: Non conscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology* 81: 1014–1027.

Bicchieri, C. 2006. *The grammar of society: The nature and dynamics of social norms*. Cambridge: Cambridge University Press.

Bratman, M. 1987. *Intention, plans, and practical reason*. Cambridge, Mass.: Harvard University Press.

Castelfranchi, C., and L. Tummolini. 2003. Positive and negative expectations and the deontic nature of social conventions. In *Proceedings of the 9th international conference on artificial intelligence and law*. Edinburgh: ACM.

Castelfranchi, C., F. Giardini, E. Lorini, and L. Tummolini. 2007. The prescriptive destiny of predictive attitudes: From expectations to norms via conventions. In *Agenti software e commercio elettronico: profili giuridici, tecnologici e psico-sociali*. Ed. Sartor, G., C.Cevenini, G. Quadri di Cardano. 43–55. Bologna, GEDIT.

Castelfranchi, C. 2012. Goals, the True Center of Cognition. In *The goals of cognition*, ed. F. Paglieri, L. Tummolini, R. Falcone, and M. Miceli. London: College Publications.

Castelfranchi, C. 2013. Cognitivizing norms. Norm internalization and processing. In *Law and computational social science*, ed. S. Faro and N. Lettieri. *Informatica e Diritto*, vol. XXII, 75–98.

Castelfranchi, C., and F. Paglieri. 2007. The role of beliefs in goal dynamics: Prolegomena to a constructive theory of intentions. *Synthese* 155: 237–263.

Conte, R., and C. Castelfranchi. 1995. *Cognitive and social action*. London: UCL Press.

Conte, R., G. Andrighetto, and M. Campennì. 2010. Internalizing norms: A cognitive model of (Social) norms' internalization. *International Journal of Agent Technologies and Systems* 2: 63–72.

Elliot, A. 2006. The hierarchical model of approach-avoidance motivation. *Motivation and Emotion* 30: 111–116.

Garfinkel, H. 1963. A conception of, and experiments with, 'trust' as a condition of stable concerted actions. In *Motivation and social interaction*, ed. O.J. Harvey, 187–238. New York: The Ronald Press.

Gelati, J., A. Rotolo, G. Sartor, and G. Governatori. 2004. Normative autonomy and normative co-ordination: Declarative power, representation, and mandate. *Artificial Intelligence and Law* 12: 53–81.

Gneezy, U., and A. Rustichini. 2000. A fine is a price. *Journal of Legal Studies* 29: 1–18.

Higgins, E.T. 1997. Beyond pleasure and pain. *American Psychologist* 52: 1280–1300.

Rao, A., and M. Georgeff. 1995. BDI-agents: From theory to practice. In *ICMAS-95.* In *Proceedings of the first international conference on multiagent systems,* ed. V. Lesser, 312–319. Menlo Park: AAAI Press.

Authority

Kenneth Einar Himma

1 Two Kinds of Authority: Epistemic and Practical

There are two kinds of authority: epistemic and practical authority. A person, *P*, has *epistemic authority* over a person, *Q*, with respect to the proposition *x* if and only if *P* sincerely asserting that *x* is true is, by itself, a reason for *Q* to believe that *x* is true.[1] An example of an epistemic authority with respect to cancer would be an oncologist. *P* has *practical authority* over *Q* with respect to the performance of act *a* if and only if *P* directing *Q* to perform *a* provides *Q* with a novel reason for doing *a*. Examples of practical authority include parents and judges.

Although practical authority differs from epistemic authority, it is worth noting, at the outset, two features common to each form of authority. First, each form of authority is characterized, at its conceptual foundation, as involving the capacity to create in another person *new* reasons of a theoretically significant kind. Second, in each case, the reasons that an authority's directive or statement provides to a subject are content-independent in the following sense: It is the source—and *not* the content—of the directive or statement that provides the subject with a reason.[2]

Nevertheless, practical and epistemic authorities are different in an important sense. Although both are properly characterized as authorities in virtue of their ability to give subjects new content-independent (or source-based) reasons, they are reasons of a different kind. An epistemic authority's opinion provides the subject with a novel reason to *believe* that opinion—provided that the content of the opinion falls within the scope of the authority's expertise. For example, a physician who tells me

[1] For a couple of notable discussions of epistemic authority, see Hurd (1999) and Zagzebski (2012).

[2] The notion of a content-independent reason will be explained in more detail below at pp. 9–10.

K. E. Himma (✉)
School of Law, University of Washington, Seattle, WA, USA
e-mail: himma@uw.edu

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_8

I have the flu has given me a new content-independent reason to believe I have the flu. In contrast, a practical authority's directive gives rise to reasons for *action*—in particular, a reason to do what the authority has directed the subject to do. For example, a judge who orders me to pay damages to *P* gives me a new content-independent reason to pay damages to *P*.

Indeed, it is commonly thought that legitimate (or morally justified) practical authorities have the capacity to bind subjects by providing reasons that give rise to a moral obligation to obey. It would be rather odd to think of epistemic authority as giving rise to reasons that *obligate* (as opposed to *oblige*) the subject to either believe what the epistemic authority opines or do what the epistemic authority recommends.[3] For this reason, an opinion from an epistemic authority should function cognitively in a different way from a directive by a practical authority.[4]

Again, while the notion of epistemic authority is relevant in addressing many issues, descriptive and normative, that arise in connection with the law, the concept of authority that figures most prominently in law is practical authority. Law is contrived, by nature, to issue directives that create new reasons for the subject to do what such directives require; after all, law, by nature, regulates behavior—and not thoughts. Accordingly, the remainder of this essay is centrally concerned with discussing practical authority.

2 Power, de Facto Authority, and Legitimate Authority

The most influential theorist on the nature of practical authority, Joseph Raz, distinguishes between de facto authority and legitimate authority (or authority per se). A legitimate authority is morally justified in issuing directives that tell subjects what to do and hence provide subjects with new reasons for doing what the authority directs them to do. In contrast, a de facto authority "either (1) claims to be legitimate or (2) is believed to be so" (Raz 1994, 211). As is readily evident, not every de facto authority is legitimate; some practical authorities might claim legitimacy without being legitimate, as is typically the case with totalitarian states. A de facto authority

[3] Although epistemologists have sometimes talked of the existence of epistemic obligations, it is not clear, given the seemingly involuntary character of belief formation, how a person could be *bound* to accept an epistemic authority's opinion on a matter within the authority's expertise; after all, we do not seem to choose what to believe. Further, while it is true that my consulting a physician, for example, acknowledges the physician's superior expertise on the matter, it is hard to see how that acknowledgment alone could *bind* me to accept her opinions—although her greater expertise surely gives me a reason to believe something.

[4] So much so that one might reasonably question whether the concept of "authority" applies in the epistemic context—if, as seems reasonable to think, the concept applies only to matters over which the subject of authority has direct volitional control. Of course, we are reason-responsive when it comes to belief, but that response is not determined by a free choice; that response is determined by a committed trust to the person's epistemic authority and is sometimes conditioned by our own sense of what is intuitively plausible, which is also not a matter of direct free choice.

is legitimate when either its claim that it has legitimate authority or its subjects' belief that it has legitimate authority is true.

Both de facto authority and legitimate authority should be distinguished from political *power*. A person can have power over another person without having authority. *P* can be an authority over *Q* only insofar as *P* is generally accepted as an authority in the relevant community (which might, or might not, include *Q*'s acceptance). All that is needed for *P* to have power over *Q* is that *P* has some reliable coercive means for inducing *Q* to comply with *P*'s commands. As is evident, the claim that *P* has power over *Q* is purely descriptive.

It is worth noting that, like the notion of power, the notion of de facto authority is purely descriptive while the notion of legitimate authority is normative.[5] As Raz puts the point, again, a de facto authority is someone who "either claims to be legitimate or is believed to be so, and is effective in imposing its will on many over whom it claims authority" (Raz 1994, 211). There are a number of ways one might conceive legitimacy: One might claim an authority is legitimate (1) when it has a "moral right to rule," as Raz sometimes puts it; (2) when its directives give rise to content-independent *moral* obligations to obey[6]; or (3) when its directives, in the case of coercive authorities, are justifiably backed by coercive enforcement mechanisms, from the standpoint of political morality.[7] But, however fleshed out, the notion of legitimacy is a morally normative notion insofar as it makes reference to someone's rights, obligations, or justified use of coercive enforcement mechanisms.

3 Practical Authority as Personal

The analysis of legitimacy has some interesting implications for the nature of practical authority. Insofar as legitimate authority has a moral right to rule or generates moral obligations to obey that are owed to the authority, it is a conceptual truth that practical authority is *personal* in the sense of being a subject of conscious experience. Only beings that are personal in this way can have a moral right to rule or can be owed moral obligations.

[5] For an extremely helpful discussion of the points sketched in this paragraph, see Christiano (2013).

[6] One interesting issue here concerns to whom the obligation to obey is owed. One could argue that the obligation is owed to the authority, which might facilitate justifying a coercive authority's use of enforcement mechanisms. Alternatively, one could argue that the obligation is owed to other subjects of the authority, as in, say, a fair-play argument that grounds an obligation to obey in one subject in the benefits that she receives from the compliance of others. But if Raz is correct in thinking that legitimate authority confers a moral right to rule on the part of the authority, then the corresponding moral obligations to obey are owed by subjects to the authority. Of course, this is compatible with the authority's generating obligations to obey that are owed to other subjects. Indeed, it seems quite plausible to think that subjects of morally legitimate authority have a moral obligation to obey that is owed to both subjects and the authority.

[7] These three conceptions of legitimacy do not necessarily coincide. As we will see, the differences between justifying coercive authority and justifying non-coercive authority are salient in giving adequate justifications for each. See Sect. 7, below.

Further, given the nature of practical authority and personal beings, an authority must be rational in the sense of being able to deliberate upon reasons in reaching a decision. For example, a lion might be able to direct a pride of lions, but it would be silly to characterize a lion as having practical authority over the other lions. Lions do not possess the relevant capacities to be capable of deliberations culminating in directives that give others reasons to act. This is not to affirm or deny that lions can have reasons to do things. It is merely to deny that lions can deliberatively issue directives in a sense that requires the sort of rationality that authorities must, as a conceptual matter, possess.

Similarly, a computing machine, given the technological limitations at this time, cannot have practical authority over a person, although it might well have epistemic authority over a person. A satellite navigation system might be an epistemic authority in that its output gives a person a new reason to believe that the system's directions are correct, but it seems implausible to think that it, by itself, gives rise to a new reason for action. Believing the directions are correct gives rise to a reason to comply with them, but that is different from claiming that the navigation system exercises practical authority. Of course, computing machines might someday have sufficient artificial intelligence capacities (which would presumably give rise to consciousness) that render them conceptually capable of being practical authorities. But those technologies are not sufficiently developed, at this juncture, to give rise to an example of artificial practical authority.

Likewise, an authority must be able to receive and issue communications to a subject regarding the authority's view about what the subject should do.[8] As the capacity for rationality is incorporated into the notion of authority, authority must be able to communicate in something that counts as a language. What properties something must have to count as a "language" is not entirely clear, but a language must surely contain semantic mappings from symbols to meanings and syntactic rules for articulating well-formed expressions in the language. Accordingly, it would seem that an authority must be capable of communicating in a language defined by semantic and syntactical rules.

If this is correct, this raises some puzzles about the logical connections between authority and authoritative directives. Any directive issued by an authority within the scope of her authority is clearly *authoritative*; thus, it is a sufficient condition that a directive issued by an authority within the scope of her authority is authoritative and provides subjects with a reason for action. But it is not obviously a necessary condition for a norm to be authoritative that it is directed or promulgated by an authority. If morality is objective and manufactured by a personal God, then the norms of morality are authoritative in virtue of being issued by a divine authority. Arguably, objective moral norms are no less plausibly characterized as "authoritative" if a personal God

[8] As Raz puts the point, "what cannot communicate with people cannot have authority over them" (Raz 1994, 217).

does not exist.[9] Either way, moral norms seem to provide subjects with the right kind of reasons for action.

More to the point, similar puzzles arise in connection with the authority of *law*. It is clear that individual valid laws—at least, those of a *legitimate* legal system—are directives that are authoritative over subjects who are within the jurisdiction of the relevant legal system.[10] What is not so clear, however, is whether the legal system that produces those authoritative directives should be considered as an authority.

The issue arises because a legal system, unlike a person's parents, is not *personal* in the sense usually thought requisite for being an authority. A legal system is neither an individual person nor some kind of compound person if there are such things. Rather, a legal system seems to be an abstract object that is constituted by various elements. Some of these elements are persons, such as individual executives, legislators, and judges; some of these elements are themselves abstract objects. Laws taking the form of rules, principles, or standards are normative propositions and are hence abstract objects.

It is important to understand that the defining property of abstract objects is that they are incapable of causally interacting with the world. There is nothing controversial about this among theorists working in the philosophy of abstract objects.[11] Consider, for example, the object denoted by the symbol "2". That object is the number for which "2" stands. Cursory reflection is sufficient to recognize that the number denoted by "2" cannot be seen, heard, touched, smelled, or tasted. In fact, the number denoted by "2" cannot be directly considered in thought; a symbol for that object is needed to represent that object in thought—although what that

[9] It is not clear whether Raz would characterize the norms of morality in such a case as "authoritative." One might deny, I suppose, that such norms are authoritative because they are not directives of an authority. I do not find that position plausible, but I cannot address the issue of whether there can be impersonal sources of authority—such as would be the case if moral norms are objective without having a personal being as author—in detail here. All I can do is raise the relevant concerns.

[10] As we have seen, a legal system might be believed to be authoritative, or claim authority, without actually being legitimate. It is not exactly clear whether, as a conceptual matter, the laws of an illegitimate legal system provide any kind of reason to act that is not connected with a prudential desire to avoid coercive enforcement mechanisms. One possible view is that such law merely "purports" to provide certain kinds of reason (usually thought to exclude the reasons provided by the authorization of coercive enforcement mechanisms), although what this would involve is somewhat unclear.

[11] As Gideon Rosen puts the point: "Concrete objects, whether mental or physical, have causal powers; numbers and functions and the rest make nothing happen. There is no such thing as causal commerce with the game of chess itself (as distinct from its concrete instances). And even if impure sets do in some sense exist in space, it is easy enough to believe that they make no *distinctive* causal contribution to what transpires. Peter and Paul may have effects individually. They may even have effects together that neither has on his own. But these joint effects are naturally construed as effects of two concrete objects acting jointly, or perhaps as effects of their mereological aggregate (itself a paradigm concretum), rather than as effects of some set-theoretic construction. Suppose Peter and Paul together tip a balance. If we entertain the possibility that this event is caused by a set, we shall have to ask which set caused it: the set containing just Peter and Paul? Some more elaborate construction based on them? Or is it perhaps the set containing the molecules that compose Peter and Paul? This proliferation of possible answers suggests that it was a mistake to credit sets with causal powers in the first place" (Rosen 2014).

representation amounts to, beyond its being ideational in character, is not clear. Thus, the propositions that constitute law are abstract objects that cannot causally interact in any way with subjects—although subjects can apprehend them through mental symbols that represent these abstract objects in a way that can be processed by rational beings in deliberating what to do.

Notice that a legal system is not an abstract object solely in virtue of including abstract objects, like the normative propositions expressed by sentences expressing laws; the legal system is an abstract object also in virtue of its being a collection of different types of object—or, otherwise put, a *set* of some kind. We cannot causally interact with sets if, as is commonly believed, sets are abstract objects. We can sometimes pick up members of a set if they are concrete individuals—such as, for example, an apple and an orange; what we cannot do is pick up the set made up of that apple and that orange with nothing more. If we put the apple and orange in a bag, we can, of course, pick up the bag; however, the bag containing the apple and orange is a concrete object that enables us to pick up the apple and orange together—and larger collections of fruits and other objects. But, strictly speaking, we are picking up the bag with the apple and orange, and not the set consisting of the apple and orange. Abstract objects cannot be picked up.

Like a set of any objects, a legal system is not something that can be touched, seen, smelled, tasted, or heard—although various elements of a legal system can be perceived in one of these ways. Indeed, like a set of fruits, a legal system is not the kind of thing that can be lifted and carried, and it is not because a legal system weighs too much. A legal system is simply not the kind of thing that can be lifted; not even an omnipotent God could lift either the number "2" or a legal system.

At first glance, this seems to create a problem for the view of a legal system as an abstract object. How could a law be authoritative if an authority does not issue or promulgate it? There are two potential answers here. First, as seen above, one could argue that it is not a necessary condition for a norm to be authoritative that some authority has promulgated it. Again, if morality is objective and not manufactured by God or some particular person, then the norms of morality can be thought of as authoritative despite not being issued by an authority.

Second, one could attempt to explain the authoritative quality of valid legal norms in terms of the authority of some particular official or perhaps the collective authority of *multiple* authorities, as opposed to a set of authorities. It is, of course, the cooperative work of these personal beings that culminates in the production of laws that are authoritative—and hence the authoritative quality of valid laws might well be inherited from the authority of these personal beings. If so, the authoritativeness of laws would be derivable from the actions of personal authorities in a way that is much more complicated than initially appears.

But it would be a mistake to try to suppress these complexities by attributing the requisite qualities of a personal being to something that is, as an uncontroversial conceptual matter, not capable in principle of doing what personal beings do. It is true that a legal system contains many objects that are capable of guiding a subject. Officials, as personal beings, can express a view about what subjects ought to do, and laws, as propositional objects, express content dictating what subjects ought to do. Nevertheless, the legal system is constituted by the set of officials, norms, etc., and is hence as much an abstract object as any other. Accordingly, if a legal system is an abstract object and *if* only personal beings can be practical authorities, then a legal system cannot be an authority because it is not a personal being.[12]

4 Practical Authority and Its Reason-Giving Capacity

Legitimate practical authority is characterized by the ability to provide a subject with a *new* reason for doing what the authority directs the subject to do.[13] That is to say, a legitimate authoritative directive provides the subject with a reason for action that she does not have absent the directive (or norm). Thus, a legitimately authoritative directive provides a reason for doing what the directive requires that alters the subject's reasons for acting with respect to the relevant action.

This new reason is *moral* in character. As discussed above, to say that an authority is legitimate is to say, among other things, that the authority is morally justified in issuing directives that bind subjects. As such, the directives of a legitimate authority give rise to moral obligations. As moral obligations give rise to moral reasons, the new reason to which a legitimately authoritative directive gives rise is a moral reason.

It bears reiterating here that not all authorities are legitimate. A merely de facto authority does not have the capacity to give subjects a *moral* reason to do what the authority directs. A de facto authority might have sufficient coercive power that a subject has a *prudential* reason to do what the authority directs—in the form of a reason to avoid being subject to coercive sanctions. But a merely de facto authority (or illegitimate authority) has no general capacity to issue directives that *provide* subjects with a new moral reason to do what the authority directs.[14] At most, as the matter is commonly put, a de facto authority "purports" to provide moral reasons for action—although purporting seems, at first glance, to require certain personal communicative capacities lacking in abstract institutional authorities, such as law.

[12] This raises an interesting issue with respect to Razian positivism. Raz takes the position that it is a conceptual truth that law "claims" legitimate authority. On this view, while law's claim to legitimate authority can be, and often is, false, it is a conceptually necessary condition for a something to count as a legal system that it makes such a claim. If law is an abstract object of some kind, then it is conceptually impossible for law to make claims. See Himma (2001) and, for a very similar subsequent argument, Dworkin (2002).

[13] As Scott Hershovitz convincingly explains: "When one makes a request, one gives the addressee a reason for action that she did not have before … Countervailing reasons may outweigh the [reason provided by the] request. If I request that you help me carry my groceries, I expect you will consider my request along with all the other reasons you have for action. I expect you to act upon my request only if it tips the balance of reason in favor of doing so" (Hershovitz 2003, 204). Although he is expressly concerned with requests in this passage, the same considerations, as Hershovitz observes, apply to authoritative directives.

[14] Of course, subjects might have a moral reason to do what a de facto authority commands if the command reflects the requirements of morality, but that reason would not be a new reason that is explained by the authorities issuing the relevant directive. See, below, at p. 10.

Further, insofar as the directives of a legitimate authority are, as a general matter, morally justified, the new moral reason to which a particular directive gives rise is content-independent in the following sense: that a legitimate authority commands that I do φ gives me a moral reason to do φ regardless of what φ is. If "φ" stands for "cross the street," then the directive to φ gives me a moral reason to cross the street; if "φ" stands for "do not cross the street," then the directive gives me a moral reason not to cross the street. The directives of a legitimate authority give rise to content-independent moral reasons to do what the directive directs. Otherwise put, the directives of a legitimate authority give rise to *source-based* reasons that do not depend on the content of the directive.

An illegitimate authority's directives might also give rise to content-independent reasons for action, but they need not be moral in character. As noted above, an authority with *sufficient* capacity to coerce subject behavior might have a capacity, other things being equal, to give subjects a content-independent reason to act, but that reason will be prudential and not moral.[15] The reason will be grounded in the subject's prudential interest in avoiding being subject to coercive enforcement mechanisms, rather than necessarily in any content-independent moral reasons.[16]

One might think that an illegitimate authority can, under certain circumstances, issue directives that give rise to moral reasons for action. Insofar as there are moral reasons to follow a directive of an illegitimate authority, these reasons will be content-dependent in the sense that it is the moral content of the directive that provides the reason. For example, there is surely a moral reason to conform to the content of an illegitimate authority's directive not to kill innocent persons.

But to say that the authority's directive in this case gives rise to a *new* moral reason to act misrepresents the situation. The subject's moral reasons derive from the moral content of the directive and not from the fact that the authority has issued the directive. Since the directive of an illegitimate authority cannot give rise to the right kind of source-based reason to obey, it cannot give rise to a new moral reason to obey. What moral reason subjects might have to obey an illegitimate directive has nothing to do with the ostensible status of the directive's source as an authority.

Thus, while the directives of a legitimate authority give rise to content-independent moral reasons to act, an illegitimate authority, on Raz's view, merely "claims" (or "purports") that its directives give rise to content-independent moral reasons to act. An illegitimate authority's directives do not necessarily give rise to reasons that are either content-independent or moral. While it is clear that a morally illegitimate authority cannot issue directives that give subjects a content-independent *moral* reason to comply, it should also be clear that any illegitimate authority that lacks sufficient coercive ability or power over a subject lacks even the general capacity to give rise to content-independent *prudential* reasons for acting.

[15] As a merely prudential reason, this is a reason that can be outweighed by moral reasons.

[16] To say that a person is "subject" to coercive enforcement mechanisms is not to make any claim about the probability of incurring liability under such mechanisms. It is rather to say that the mere authorization of coercive enforcement mechanisms backing the directive gives a person some content-independent reason (though possibly quite weak) to comply with the directive.

5 Practical Authority and Its Capacity to *Bind* Subjects

The *new* moral reasons created by legitimately authoritative directives have a notable quality: Those reasons are sufficiently strong to create a *moral obligation* to comply with the authority—or so it is commonly thought. One of the most intuitively conspicuous features of authority is that its legitimate directives *bind* subjects. It is not merely a matter of a legitimate directive being something a subject *should* obey; it is rather a matter of a legitimate directive being something a subject *shall* or *must* obey. In some very difficult sense to specify (which does not in any way implicate the subject's capacity for free will), the subject of a legitimately authoritative directive is not "free" to disobey.

This is not limited to sources of authoritative directives that take the shape of a legal system. Any legitimate authority has the capacity to morally obligate subjects with a directive that falls within the legitimate scope of the authority. The purpose of an arbitrator, to take one of Raz's most influential examples, is to resolve a dispute between two persons regarding what should be done based on the reasons that antecedently apply to them. If the arbitrator may legitimately weigh the reasons and resolve the dispute *for the subjects*, then they are bound, morally (and possibly legally, depending on the facts of the particular case), to comply with the arbitrator's decision.

In practice, most instances of arbitration are reasonably thought to be legitimate because grounded in mutual promises of the parties to abide by the decision; even when ordered by a court, the directive is typically grounded in some kind of antecedent agreement between the parties. While the existence of a contractual obligation might not be a necessary element of the arbitrator–subject relationship, an exchange of promises to obey an arbitrator, regardless of what she decides, gives rise to at least a prima facie moral obligation to obey the arbitrator. But it is surely possible for arbitration that is imposed on the subjects to be legitimate and hence create moral obligations to obey in the relevant subjects.

In cases of coercive authority, the legitimacy of an authority has one very important implication: It would appear to morally justify the use of coercive enforcement mechanisms by the authority to ensure compliance. Insofar as coercion presumptively infringes upon a person's moral interests in acting autonomously (construed to include a moral interest in not being coercively required to perform, or abstain from, a particular act), it is morally problematic and requires some kind of moral justification. If a subject has a content-independent moral obligation to obey the authority's directives, then there is some reason to think that it is morally permissible for the authority to resort to coercive enforcement mechanisms as a response to non-compliance.

Of course, the issue is somewhat more complicated than that in both directions. The existence of a moral obligation to obey authority is not obviously a necessary condition for the authorities being morally justified in coercively imposing certain consequences for non-compliance. If a parent's power to direct his or her children includes a morally justified capacity to resort to (mildly) coercive sanctions to induce compliance, that capacity cannot always be explained by the existence of a moral obligation to obey. Children do not come into the world with a developed capacity for moral agency that renders them morally accountable for their behavior; it is simply nonsense to think that, absent extraordinary circumstances, a three-year-old child has any moral obligations whatsoever.

Indeed, the example of parental authority creates a problem for the common view that the reasons created by a legitimately authoritative directive are moral in character. Although it is probably true that there are degrees of moral agency and moral accountability and that children become full moral agents by gradually acquiring the properties giving rise to moral agency, there will still be some children who seem subject to legitimate parental authority without any degree of moral agency (e.g., a two-year-old child). While other older children will be increasingly subject to acting according to moral reasons over time, very young children, fully lacking the relevant capacity, will not be subject to moral reasons. Indeed, it is reasonable to think that—if the language of reasons does not apply to very young children—such children are likely to comply out of some sense of prudential interest. At the earliest stage, rearing a child is a matter of developing certain stimulus-response mechanisms that will eventually culminate in views that will become moralized through the socialization process.[17]

Moral agency requires both the capacity for free action and a capacity for rationality that is sufficiently developed to support some threshold level of understanding of core moral requirements. Children may come into the world able to *choose* freely—and that much is incorrect if, as seems reasonable, free choice requires the ability to rationally weigh reasons—but they clearly do not come into the world with a sufficiently developed capacity to understand core moral requirements that would warrant either imputing obligations to them or holding them accountable for breaching such obligations. If this is correct, then authority can be morally justified in imposing coercive consequences for non-compliance on a subject without her having a moral obligation to comply.[18] Thus, the existence of a content-independent moral obligation to obey on the part of a subject is not a necessary condition for an authorities being morally justified in coercively enforcing a directive.

Nor is the existence of a content-independent moral obligation to obey an authority's directives a sufficient condition for a justified application of coercive enforcement mechanisms. Inducing compliance in a subject by coercive means remains

[17] Indeed, this description conforms to the first level of moral development in the theories of both Lawrence Kohlberg and Carol Gilligan. See Kohlberg (1984) and Gilligan (1982).

[18] In the case of parenting, these coercive consequences are not properly characterized as being "punishment" in any sense that includes a retributivist notion that the relevant unpleasant consequences are, as a moral matter, *deserved*. Parental discipline of young children can be characterized as "punishment" in a less robust sense that does not involve moral connotations of the disciplinary actions being deserved by the child. Parental discipline might, of course, be morally warranted by the parent's moral duties to rear a child to have certain character traits and behave in certain ways. But "punishment," in the robust sense of the word, connotes that the unpleasant consequences are morally deserved by the non-complying behavior of the subjects. As can be seen, the issues that arise with respect to authority and coercive enforcement mechanisms are quite complex—and, for that reason, cannot be addressed in more detail here.

morally suspect in the sense that we lack a theory that shows clearly that authorities are morally justified in enforcing their directives. If the content-independent moral obligation is owed to someone other than the authority, then the authority is not the one who is wronged by non-compliance. While merely being wronged by an act does not necessarily justify imposing coercive sanctions or inducements to act, it is not a trivial matter to see how someone could be justified in imposing such measures for non-compliance if non-compliance does not result in a wrong to the person seeking to impose such measures. While the law frequently allows for such practices, the question is whether and how those practices are justified.

Of course, there are other elements that might be present in the authority–subject relationship that could justify the use of coercion. In the case of a private arbitrator, the subjects might contract with each other to abide by the arbitrator's decision subject to certain coercive penalties. In this case, the arbitrator does not seem fairly characterized as having been wronged by a non-complying party and yet is justified in imposing the penalty. What does the necessary moral work in this case is the mutual exchange of promises supported by consideration (i.e., the contract); these features give rise to morally protected reliance interests that might justify coercive enforcement mechanisms. The existence of a content-independent moral obligation to obey the authority does not suffice, by itself, to morally justify the authority's imposition of coercive penalties for non-compliance.

6 The Kind of Reasons to Which Legitimate Directives Give Rise

As we have seen, legitimately authoritative directives are commonly thought to create content-independent moral obligations to obey and hence provide some type of special reason for action.[19] That an act is morally good provides a reason to perform that act, but it is not a reason that *binds* the subject. Moral obligations bind subjects and hence provide reasons for action that are considerably more robust in the sense that the subject of an obligation does not have an option not to comply; a subject, in contrast, should, but need not, perform an action that is morally good.

Moral obligations can thus be seen as providing a reason that is "final" in the sense that it either *is not* or, depending on one's metaethical view, *cannot* be defeated by other reasons.[20] Each possibility requires a different formulation. Raz's notion of a

[19] See the discussion on parental authority and the moral incapacities of young children, above, at p. 11.

[20] According to William Frankena, morality is "supremely authoritative"; on this plausible view, moral obligations claim supremacy over all other obligations—including legal: When moral obligations come into conflict with other obligations and practical considerations, the moral obligations win; the only thing that can defeat a moral obligation is another more important moral obligation (Frankena 1966, 688–696). Similarly, Bernard Gert describes this feature of morality as follows: "Among those who use 'morality' normatively, all hold that 'morality' refers to a code of conduct that applies to all who can understand it and can govern their behavior by it. In the normative sense,

conclusive reason captures the weaker idea that what I have called a final reason is one that is not, as a contingent matter, defeated by other reasons. As Raz defines the notion, "*p* is a conclusive reason for *x* to ϕ if, and only if, *p* is a reason for *x* to ϕ (which has not been cancelled) and there is no *q* such that *q* overrides *p*" (Raz 1975). It is crucial to note that, according to Raz's definition, a conclusive reason is one that, as a matter of *contingent* fact, *is not* outweighed or overridden by other countervailing reasons. That is, a merely conclusive reason would be final with respect to outweighing any other reasons obtaining in *the* actual (and thereby in one specified contingent) world.

In contrast, Raz's notion of an absolute reason captures the stronger idea that a final reason is one that *cannot* be defeated by other reasons; that is, that there is no logically possible world in which an absolute reason is outweighed or overridden by countervailing reasons. As Raz defines this notion, "*p* is an absolute reason for x to ϕ, if and only if, there *cannot* be a fact which would override it; that is to say, for all *q* it is never the case that when *q*, *q* overrides *p* (Raz 1975, 27; emphasis added)."[21]

This much about Raz's view seems uncontroversial. Morality is thought to trump all other considerations in the following sense: Only a more important moral obligation can provide a reason for action that defeats the reasons for action provided by a less important moral obligation. Accordingly, if *P* has a moral obligation to do φ, then *P* is bound to do φ, regardless of other considerations—prudential or otherwise. Morality is supreme in terms of the force of the reasons moral obligations provide. Insofar as this is so, it follows that the reasons morality provides are final in the sense described above.

Accordingly, each of the types of reason Raz defines is a potentially accurate description of the final reasons morality provides depending on whether or not morality is objective or conventional in character. If conventional in character, then the truth-value of any moral principle is contingent, since it depends on the contingent views or practices of those who determine the content of the relevant convention; in this case, the appropriate sense in which reasons are final would be that they are *conclusive* in character. If objective in character, then a moral principle is necessarily true, if true at all[22]; in this case, the appropriate sense in which reasons are final would be that they are *absolute* in character.[23]

One of the most influential features of Raz's theory of practical reasoning is an account of the nature of the final reasons that apply in morality and of the way in which they either characteristically or should function in moral deliberations. Raz

morality should never be overridden, that is, no one should ever violate a moral prohibition or requirement for nonmoral considerations" (Gert 2012).

[21] There is an ambiguity in Raz's formulation of the two kinds of reason. While the first clause uses the modality "cannot," the second employs only a variation of a universal quantifier ("never"), suggesting that there is, in fact, no overriding *q*. To avoid replicating the notion of a conclusive reason, the notion of absolute reason should be construed as intending the modal clause. Otherwise, there is little difference between the two concepts.

[22] For example, on an objectivist view, it *cannot* be morally permissible to torture infants for fun.

[23] One can, of course, disagree that moral objectivism implies that moral judgments are necessarily true, if true at all. If so, then Raz's notion of a conclusive reason would apply to an objective morality.

calls this specific type of reason a *protected reason*, which consists of a first-order reason (i.e., a reason having to do with actions or beliefs) to do what the authoritative directive or valid moral norm requires, together with a second-order reason (i.e., a reason for acting, or not acting, on some set of first-order reasons), which he calls an *exclusionary reason*.[24]

As the notion of a first-order reason to do what a norm or directive prescribes is straightforward, the remainder of this section will be concerned with explicating the notion of an exclusionary reason. To begin, note that the notion of an exclusionary reason is compatible with each conception of morality and each of Raz's conceptions of a final reason. A reason, for example, could be exclusionary in all logically possible worlds, or it could be exclusionary in some worlds, but not others, depending on the specific circumstances of the possible world.

According to Raz, an "exclusionary reason is a second-order reason to refrain from acting for some reason" (Raz 1975, 39). Normally, practical rationality requires of a person that she weighs all of the relevant reasons and acts on her assessment of the balance of reasons.[25] Exclusionary reasons operate to exclude certain reasons from the reasons on which a person can rationally act. They do not prevent her from deliberating to determine what the balance of excluded reasons would require by way of acting; they simply preclude her acting on the basis of those reasons or her assessment of those reasons.

On this view, for example, a moral obligation not to kill an innocent person provides an exclusionary reason that precludes the agent's acting on the basis of certain reasons she might take herself to have to kill an innocent person, such as a desire to kill the person for the purpose of taking her belongings. This element of

[24] Hershovitz does a characteristically elegant job of explaining the notion of a second-order reason: "What does it mean to have a second-order reason, a reason to act for or not act for another reason? An illustration will help. Suppose Aaron's grandmother is in the hospital and that this provides Aaron reason to visit her. Suppose further that Aaron goes to the hospital and visits his grandmother, but only because he was hoping to run into Michelle, whom he has a crush on. In this case, Aaron conforms to his reason to go to the hospital to visit his grandmother but he does not comply with it. Does Aaron have reason to comply with his reason to visit his grandmother rather than just conform with it? Raz suggests he does, and I agree. Because Aaron went to the hospital to see Michelle and not his grandmother, his actions do not embody appropriate respect for his grandmother. Aaron had a reason to comply with his reason to visit his grandmother, that is, he had a second-order reason to act for a reason: Only through visiting his grandmother for the sake of visiting her could he show her proper respect" (Hershovitz 2003, 202).

[25] Heidi M. Hurd, for example, argues it can never be practically rational to accept exclusionary authority because it violates the principle that an agent should always act on the balance of reasons available to her (Hurd 1991). As Thomas May makes the point: "Acting on what the authority judges ought to be done appears to circumvent one's own evaluational judgement, and thus autonomy. By circumventing the evaluational judgement of the subject it seems the subject is *prevented* from acting on her own determination of what ought to be done. The subject seems to be eliminated from the determination of her behavior" (May 1998, 130). It is worth noting that even if it is practically irrational to accept authority in the sense of providing exclusionary reasons, it does not follow that authority is necessarily illegitimate. This could simply be taken to imply that consent would be no part of a successful theory of state legitimacy and that other moral considerations would be sufficient.

exclusionary reasons is primarily negative in character: It simply precludes acting on the basis of some specified set of reasons.

Of course, a moral obligation not to kill also provides a *new first-order* reason not to kill, which makes it another example of a protected reason, but its exclusionary character distinguishes it from other reasons having to do with whether to kill in the following respect: It is not a reason to be weighed in the balance with other reasons for or against killing; it is a first-order reason not to kill coupled with a second-order reason not to act on a specified class of reasons that would include, for example, any prudential reasons.

On Raz's view, a legitimately authoritative directive requiring P to do φ creates a first-order reason for P to do φ and a second-order exclusionary reason not to act on a specified class of other first-order reasons. Again, exclusionary reasons do not preclude an agent from deliberating to determine what the balance of excluded reasons require; they simply preclude the agent from acting on those reasons or on her assessment of what those reasons require.

In contrast, H. L. A. Hart took the position that legitimately authoritative directives provide *peremptory* reasons that preclude the agent from even deliberating on the class of excluded reasons (Hart 1982, 253).[26] On this view, an agent who has a peremptory reason to do φ would be violating norms of practical rationality (which presumably incorporates norms of both morality and prudence) simply by deliberating on—or weighing for herself—the class of excluded reasons.

Hart's view is problematic. Neither the norms of morality nor the norms of practical rationality are directly concerned with an agent deliberating on reasons. Morality is chiefly, if not exclusively concerned, with what an agent does in the world—not with how an agent deliberates on reasons that do not ultimately culminate in her performing some act.[27] Practical rationality is concerned with evaluating the rationality of a person's actions, and this is done by assessing whether the agent's acts conform to the balance of what Raz sometimes calls "right reason."

There might be exceptional cases in which norms of rationality or morality would condemn an agent for merely having a thought or considering a reason—even if she does not act on that reason. Consider, for example, a person P who consciously harbors a white supremacist worldview that is grounded in the thought that persons of other races are "subhuman" and undeserving of moral respect. Suppose P never acts either on those views or in any way that would express those white supremacist

[26] As Hart puts the point: "[T]he commander characteristically intends his hearer to take the commander's will instead of his own as a guide to action and so to take it in place of any deliberation or reasoning of his own: the expression of a commander's will that an act be done is intended to preclude or cut off any independent deliberation by the hearer of the merits pro and con of doing the act…. This, I think, is what is meant by speaking of a command as 'requiring' action and calling a command a 'peremptory' form of address" (ibid., 253).

[27] It is true, of course, that the moral evaluation of a person's action will also include consideration of her motive for acting. Giving to charity for the reason that it will help the poor, for example, is a morally valid reason to give to charity; doing so to enhance one's reputation in the community is not. But a person's motive has to do only with the actual reason on which she acted and not on how she arrived at that reason. There might be instances in which the reasoning comes into play, but this is not how one's mental states are characteristically evaluated from a moral point of view.

views in a harmful fashion. Indeed, assume that *P*'s outward behavior evinces equal treatment and respect for all persons—white and nonwhite—and that *P*'s outward behavior is so benign when it comes to differentiating persons on the basis of race that no one would ever think to characterize *P* as a racist.

It is reasonable to think that the mere holding of such views is morally culpable, but this is not an unproblematic position. Again, morality seems characteristically concerned with outward behavior and not inner mental states, in part, because it is not clear to what extent the relevant inner states—our beliefs and our desires—are within a person's volitional control. It is plausible to think that some mental states are subject to moral evaluation; hate might be one such example, along with the racist views considered above. But if these are subject to moral evaluation, these states would seem to constitute either borderline or otherwise exceptional cases.

If it is somewhat unclear whether morality or practical rationality applies to the mental states and acts discussed above, there is little reason to think that either precludes merely deliberating on a set of excluded reasons. Perhaps norms of morality and practical rationality would condemn being tempted to act on one's deliberation on excluded reasons after one has arrived at a view about the balance of those reasons, but it seems implausible to think that those norms would condemn deliberating on excluded reasons out of curiosity or out of a desire to see whether the directive lines up with the outcome of one's own deliberations. If so, Hart's view that legitimate directives give rise to peremptory reasons is false.

Raz's view is, clearly, not subject to such objections. Raz's view harmonizes much more closely with morality's significantly greater concern with actions than with thoughts. Raz's account of an exclusionary reason does not preclude deliberating on the balance of applicable and excluded reasons; it merely precludes acting on the basis of those reasons.

This certainly conforms to our intuitions about authority in general. If an arbitrator legitimately decides a contested issue between *P* and *Q* with a directive that binds them, there seems nothing either practically irrational or morally problematic with either *P* or *Q* deliberating on the excluded reasons. Certainly, neither the parties nor the arbitrator seems to have any grounds for complaint or criticism in such a case.

Nevertheless, Raz's view is subject to some questions and concerns. To begin, there are other mechanisms than that of an exclusionary reason for capturing the idea that authoritative directives provide reasons that are final in the relevant sense. As Stephen Perry has persuasively observed, an authoritative directive need not function as a second-order reason to do the conceptual work authority is thought to do (Perry 1989). If sufficient weight is assigned to the directives of authority, it might simply outweigh all the other applicable reasons in the vast majority of cases—which would be enough for authority to be fairly characterized as capable of performing its conceptual function of telling people what to do by issuing directives making certain behaviors mandatory.[28]

[28] This is a view of authoritative reasons that would not run afoul of Hurd's view that it can never be rational for a person to accept *exclusionary* authority. Authority, as Perry conceives it, is not exclusionary, as it provides only strongly weighted reasons, rather than exclusionary reasons.

Further, it is not clear that Raz's account correctly applies to all forms of authority. One particularly salient form of authority that does not seem to cohere to the Razian account is law. It is clear that, on Raz's view, the valid legal norms of a *morally legitimate* legal system—i.e., a legal system that is an "authority," rather than just a "de facto authority"—would give rise to exclusionary reasons for action; morality seems to be a paradigm for systems of norms that give rise to exclusionary reasons in the sense that if any system gives rise to such reasons, morality does. What is not as clear is whether a legal system that has purely de facto authority—i.e., one that is not morally legitimate—gives rise to exclusionary reasons.

The question, then, is whether it is a conceptual truth that law gives rise to exclusionary reasons or, otherwise put, whether law as such gives rise to such reasons, which would entail that even illegitimate legal systems provide such reasons. Raz seems to answer the question in the affirmative:

> The legal point of view and the point of view of any other institutional system is an exclusionary point of view. Legal norms may conflict and in deciding what, according to law, ought to be done one may have to balance different conflicting legal considerations, but law is an exclusionary system and it excludes the application of extra-legal systems (Raz 1975, 145).

Likewise, Raz states that "an authoritative determination of a primary organ to the effect that x has a duty to perform a certain action is an exclusionary reason for x to perform that action" (Raz 1975, 145).

One issue that arises in connection with Raz's view here is what the source of the exclusionary reason would be. On Raz's view, prudential concerns to avoid being subject to coercive enforcement mechanisms define reasons for action but the reasons are "of the wrong kind" (Raz 1975, 145); other things being equal, prudential concerns are first order in character. Sanctions, thus, define, as Raz puts it, "auxiliary" reasons and not exclusionary reasons (Raz 1975, 145).

On Raz's view, it is a conceptual truth that law provides reasons that are unrelated to either the authorization of coercive enforcement mechanisms or morality. On this view, "it is the fact that those actions are required by law … [that] is the reason for performing them" (Raz 1975, 155). Otherwise put, the claim seems to be that one should comply with the law because it is law.

At first glance, if this is Raz's view, it is puzzling. It is hard to understand—and especially under a positivist view of the sort Raz holds—how law *as such* could give rise to reasons of any kind not related to the possibility of incurring sanctions. From the standpoint of practical rationality, it is hard to see why the mere fact that a norm was promulgated according to a social rule of recognition (i.e., has a social "source") would give rise to a reason of any kind. Once sanctions are subtracted from the picture, it is not clear what features of law would do the necessary reason-giving work.

As it turns out, Raz seems to take a weaker and more plausible position. On his view, law *as such* provides reasons only to those who have accepted the law and thus take what he calls "the legal point of view." As he describes the notion:

> The ideal law-abiding citizen is the man who acts from the legal point of view. He does not merely conform to law. He follows legal norms and legally recognized norms as norms and accepts them also as exclusionary reasons for disregarding those conflicting reasons which they exclude (Raz 1975, 171).

Further, he holds that it is not conceptually necessary for the existence of a legal system that citizens take the legal point of view—or even that they, from the standpoint of morality, should take the legal point of view; as Raz puts the point, "[i]t is not necessary for a legal system to be in force that its norms subjects are ideal law-abiding citizens or that they should be so (i.e. that legal norms are morally valid)" (Raz 1975, 171). Rather, he holds "it is necessary that its judges, *when acting as judges*, should on the whole be acting according to the legal point of view" (Raz 1975, 171).

It should be noted that the above quote calls attention to an ambiguity in the claim that law *as such* provides exclusionary reasons. On one interpretation, the claim states that every valid *legal norm* provides an exclusionary reason. On another interpretation, the claim states that every *legal system* provides some exclusionary reasons. The above quote suggests that Raz has in mind the second interpretation, which makes a much weaker conceptual claim about law than the first. Here, it is important to note that the second quoted sentence in the last paragraph states only that judges "should on the whole" regard the law they apply as exclusionary reasons; it is not conceptually necessary that they regard every such law as exclusionary.

Raz's view, as expressed above, is weaker in a second sense. According to the above quote, what is conceptually necessary to the existence of a legal system is not that it provides exclusionary reasons for citizens. Rather, what is essential is that law provides exclusionary reasons to a particular subclass of *officials*—namely judges. What is notable here is that, on Raz's view, judges must *accept and treat* not only duty-creating recognition norms as exclusionary, but also the first-order norms they apply to subjects.

As regards citizens, then, one might take the view that it is a conceptual necessity that law as such merely "purports" to provide citizens with exclusionary reasons insofar as law is enforced in an exclusionary manner. In contrast, if morally legitimate legal systems give rise to content-independent moral obligations, then the laws of a morally legitimate legal system would provide citizens with exclusionary reasons that would be moral, and not legal, in character.

Although Raz's claim is frequently understood to be that law provides citizens with exclusionary reasons, there is good reason to reject that claim, as Raz appears to. There are many valid laws of legitimate legal systems that do not seem plausibly characterized as giving *citizens* exclusionary reasons. For example, I habitually park illegally because it is profitable to do so. The law prohibits parking without paying a fee and imposes a fine for non-compliance. The fine is $45, compared to the cost of parking, $10. I park illegally five times a week and get a ticket, at most, once a month. Assuming a month has four weeks, my net saving is $155—a profit that, on my prudential calculations, makes parking illegally the rational thing to do.

Illegal parking seems neither necessarily morally wrong nor necessarily practically irrational, but it can be problematic from either standpoint in certain cases. If one illegally parks in front of a fire hydrant, one is creating a risk of harm to others—and that seems problematic from the standpoint of both morality and practical rationalities (conceived of, as Raz's account does, as concerned to identify what right reason requires). Similarly, if one takes up a parking space for significantly longer than the time allowed for that space, one arguably acts in a way that is problematic from each vantage point. Finally, if one parks illegally in a space reserved for disabled persons, then one is acting in a morally problematic way.

But more mundane instances seem consistent with both morality and practical rationality. If I park illegally for half the time that anyone is allowed to park in the space, it is far from obvious that I have done any moral wrong or violated norms of practical rationality. Parking laws are enforced by comparatively minor fines, and there is little stigma, moral or otherwise, attached to such mundane instances of illegal parking.[29]

Nor is it any more plausible to think that officials regard such laws as giving rise to either moral obligations or exclusionary reasons. While the model of conceiving certain laws as simply defining costs for non-compliance surely cannot be generalized across all areas of law (e.g., criminal law), it seems eminently plausible in the case of "municipal offenses." Indeed, I have frequently come back to my car just as a police officer was about to write a ticket and persuaded her to let me go. When an officer does let me go, she does so with just a lecture and warning—suggesting that, at some level, she understands and *condones* my reasoning.

This is reason enough to reject, as Raz seems to, the idea that law must provide citizens with exclusionary reasons, but there are other laws than those defining municipal offenses that are not plausibly regarded as giving citizens exclusionary reasons to comply. Consider, for example, contract law, and suppose that P and Q make a contract for P to do φ. Suppose that it is far more profitable for P not to do φ and to pay what a court requires as damages for breach. Suppose, further, that Q is indifferent with respect to whether P performs or pays court-ordered damages. Finally, suppose that P and Q know all the facts and that Q will choose to litigate if P breaches, rather than to settle.

From the standpoint of practical reason, the rational thing for P to do—from each party's vantage point—seems to be to breach and pay damages. Since Q is not made worse off by the breach and P is made better off by the breach, it would seem rational for P to breach, unless there is some *moral* reason requiring P to perform, a claim that seems implausible. However, the reasons adduced above seem to exhaust the class of relevant reasons; it seems silly to claim that an overriding reason would be "because the law says so." Thus, the rational thing to do seems to be for P to breach—and, notably, for reasons that are *prudential* in character.

Looking even at official practice, the rational thing for P to do seems to be to breach. From the standpoint of the legal system, there is nothing that would seem to entail that officials expect the parties to perform under the contract. If officials conceived of contract law as providing exclusionary reasons, then the appropriate legal remedy for breach would be specific performance. But courts in every

[29] Similar claims can be made about crossing against a red light in the middle of the night when no one is on either of the roads intersecting at a crosswalk.

jurisdiction of the USA prefer ordering payment of money damages to ordering specific performance. A special showing that money damages cannot adequately compensate for the breach is a legal requirement for ordering specific performance. But if a special showing is required to justify a court order of specific performance of the very thing *P* contracted to do, then the law itself seems not to view the mandatory norms requiring mutual performance of contractual duties as giving rise to exclusionary reasons.

Indeed, it seems implausible to think that anyone *should* have any other attitude toward the violations of such laws. It is surely reasonable to think that criminal laws provide exclusionary reasons—if any do—but the whole point of cutting the law up into the different areas of criminal and civil, with the different enforcement mechanisms, suggests that officials themselves believe that different types of offenses give rise to different levels of normative force.

It is true, of course, that valid legal norms—even those of the civil law—tend to be enforced in a way that excludes certain legal excuses or justifications, but this does not imply that the law provides, or even purports to provide, exclusionary reasons. Courts typically, though not universally, enforce some kind of remedy for violations of the law that are motivated only by prudential reasoning—suggesting that the law "excludes" certain justifications for non-complying behavior as excuses that will preclude application of the law's coercive enforcement mechanisms. Even so, the fact that law is enforced in such a way entails nothing as to the structure and functioning of the reasons law provides. In particular, exclusionary enforcement of law does not imply either that subjects characteristically do, or should, regard the law as providing exclusionary reasons.

7 The Justification of Practical Authority

To the extent that the exercise of practical authority appears to infringe autonomy rights, it raises a number of normative issues. First, insofar as the exercise of practical authority seems to infringe on autonomy rights, it raises a moral issue: What conditions must practical authority satisfy to be morally justified? Second, insofar as it is inconsistent with the principle requiring that one acts on the balance of reasons, it raises an issue of normative practical rationality: Under what circumstances is it practically rational to accept an authority and act on its directives?[30]

The importance of justifying practical authority—especially state authority—arises because there are two different forms of philosophical anarchism challenging the idea that many states are legitimate. According to *strong philosophical anarchism*, no state can be legitimate unless subjects rationally consent to its authority; however, it is not rational for subjects to consent because people have a moral duty not to allow an authority to substitute her judgment for theirs in deciding what to do (Wolff 1970). According to *weak philosophical anarchism*, it is possible for a state to

[30] See, e.g., Hurd (1991), and note 25.

be legitimate on the basis of subject consent because people have a moral *right*, which can be waived by consent, not to be bound by the state's commands—rather than a moral *duty* not to allow the state to preempt their own judgment (Green 1989; Simmonds 2001).

Complicating the task of justifying legitimate authority is that practical authority can take different forms that have morally salient different features. Some practical authorities lack the capacity to coercively enforce their directives. For example, one might think a physician is a *practical* authority in the areas of her expertise (i.e., *epistemic* authority) and can issue what appears to be authoritative directives (e.g., "take two aspirins and call me in the morning"), but a physician lacks any capacity to coercively enforce her directives.[31] In contrast, many practical authorities have the capacity to coercively enforce their directives. For example, an arbitrator to a dispute typically has some coercive mechanism at her disposal to enforce her resolution of the dispute.

Of course, the most important example of an authority with the capacity to coercively enforce its directives is a legal system. Regardless of whether or not the authorization of coercive enforcement mechanisms is a conceptually necessary feature of a legal system,[32] every existing legal system of which we know is authorized with the capacity to coercively enforce the law—and does so, sometimes in a ruthless and discriminatory way. Tragically, this is happening in the USA with increasingly frequent and unjustified shootings of unarmed black persons by police with little apparent appreciation of the fact that black lives matter.

This feature of law gives rise to what is the defining problem of normative political philosophy. The problem is how to justify the state's doing what no one else is permitted to do—namely issue commands backed by the threat of violence. In this respect, the state resembles an armed robber whose demand for the victim's money is backed by the threat of force. There are, of course, many philosophical attempts to state the conditions that determine when the state is morally justified in using coercive enforcement mechanisms to induce compliance with the law or punish noncompliance—the theories of John Rawls and Robert Nozick being two of the most highly influential in contemporary political philosophy.

As interesting as these theories are, they are not relevant here, as they would not apply to all forms of practical authority. Neither Rawls's theory of justice (i.e., the principles of justice he believes would be selected from the original position) nor Nozick's libertarianism is even remotely relevant with respect to the normative issues involved in accepting and complying with a physician's practical authority—or, if

[31] The issue of whether a physician has practical authority is a difficult one, but not much turns on it here if, as many theorists believe, there can be practical authorities that lack the authorization to coercively enforce directives. Joseph Raz, for example, argues that law—and hence legal authority—would be needed to resolve certain disputes in a society of angels who are always conclusively motivated to obey the norms that resolve those disputes (Raz 1975, 159–160; Shapiro 2011, 169–170). For a response to this argument, see Himma (2016).

[32] I argue that the authorization of such mechanisms is a conceptually necessary feature of law. See Himma (2017, 593–626).

one is skeptical that physicians have practical authority over a patient, the practical authority of any agent that lacks authorization to coerce compliance.[33]

In contrast, Joseph Raz has a more comprehensive normative theory that is intended to state the conditions under which practical authority is justified—one that would cover the practical authority of someone who lacks the ability or authorization to coerce compliance. According to the normal justification thesis (“NJT”), authority is justified to the extent that the subject is more likely to do what right reason requires by following authoritative directives than by following her own judgment:

> The normal and primary way to establish that a person should be acknowledged to have authority over another person involves showing that the alleged subject is likely better to comply with reasons which [objectively] apply to him (other than the alleged authoritative directives) if he accepts the directives of the alleged authority as authoritatively binding, and tries to follow them, than if he tries to follow the reasons which apply to him directly (Raz 1994, 214).

Given the mediating function of authority in Raz’s service conception of authority, it is natural to suppose that authority is justified only insofar as it does a better job than its subjects of deciding what the reasons that antecedently apply to the subjects require by way of action.

To see the motivation for NJT, it is helpful to compare the justification for taking someone’s advice. Consider a case in which one person *P* will be hurt if her friend *Q* does not accept *P*’s advice. The desire to spare *P*’s feelings might, depending on the circumstances, be a reason for accepting *P*’s advice; if the matter were sufficiently inconsequential and the advice were harmless, *Q* might be justified in accepting *P*’s advice to avoid hurting her feelings. But, as Raz points out, that is not the normal reason for accepting advice: “The normal reason for accepting a piece of advice is that it is likely to be sound advice” (Raz 1985, 19). Likewise, it seems natural to accept and follow a practical authority’s directives because they are likely to require subjects to do what, as a matter of objectively right reason, ought to be done.[34] If so, NJT states, at the very least, considerations that are clearly relevant with respect to justifying authority.

Raz describes NJT as a “*moral* thesis about the type of argument which could be used to establish the *legitimacy* of an authority” (Raz 1985, 18; emphasis added). Further, on Raz’s view, the legitimacy of an authority entails what he calls “a [moral] right to rule,”[35] and thus Raz believes that it entails the existence of a content-independent moral obligation on the part of subjects to obey the authority’s directives

[33] This strikes me as a difficult issue. While a physician is clearly an epistemic authority, it is not clear whether a physician is a practical authority. On the one hand, a physician’s recommendations do not seem to be directives in the relevant sense; on the other, it does not seem to be irrational for a patient to accept a physician as being a practical authority. But deciding this issue is not important here insofar as it is commonly accepted that there can be practical authorities that lack the capacity to coerce compliance or punish disobedience.

[34] For a critical discussion of NJT, see Himma (2007).

[35] One reason to think that physicians lack practical authority is that it is implausible to think that a physician has a “moral right to rule.” Whether that is true depends, I suppose, on the character of a patient’s acceptance. Nonetheless, it is worth noting that if NJT is construed as a sufficient condition for the justification for accepting someone as a practical authority, it would justify accepting

(provided, of course, that the content of the directives falls within the scope of the authority's legitimacy).[36]

There are a number of reasons to think that NJT fails as a general principle of moral legitimacy. To begin, even if NJT succeeds as a principle of moral legitimacy with respect to practical authority lacking authorized access to coercive mechanisms, it is reasonable to think that something more would be needed to justify the legitimacy of coercive state authority. The introduction of coercive enforcement mechanisms complicates the moral issue considerably because persons have a presumptive moral right to be free of coercion. The exercise of all practical authority raises a moral issue because it infringes a person's moral right (or duty, as Wolff would have it) to make and execute her own decisions. But coercion introduces the threat of violent repercussions for non-compliance that would appear to require much more by way of moral justification than a principle that merely requires that the authority be better than the subject at deciding what the subject should do according to right reason. Even if any person who functions as a practical authority is passively accepted or acquiesced to by the subject, it is not enough to justify the threat of violence that the authority knows better than the subject what the latter should do according to right reason.

But the potential problems do not end here. Satisfaction of NJT seems neither sufficient nor necessary to give rise to content-independent moral obligations to *obey* the directives of a de facto authority; otherwise put, satisfaction of NJT does not seem to be either sufficient or necessary for authority to be morally legitimate. Clearly, it is not sufficient. The mere fact that complying with the directives of an authority is more likely to conduce to my interests (assuming right reason dictates that I should do what conduces to my interests) than not complying can, to borrow from Hart, *oblige* me to do what the authority directs—"obliging" being a notion that is prudential in character; if complying would conduce to my interests, I would be foolish not to accept the authority's directives. But the mere fact that following an authority is in my interest cannot, in and of itself, morally *obligate* me to obey the directives of the authority.

The deeper problem here is a familiar one. Raz's NJT seems to take the satisfaction by *P* of whatever standards confer epistemic authority over *Q* as sufficient to confer practical authority on *P* over *Q*.[37] But if this is Raz's considered view, it is mistaken. Epistemic authority is concerned with providing truth-conducive reasons to believe

physicians as practical authorities. These conflicting considerations highlight the difficulties concerning the issue whether physicians are practical authorities. See note 33, above.

[36] The idea that a patient has a content-independent moral obligation to obey a physician seems highly implausible, which, of course, casts doubt on the idea that physicians are practical authorities. Although I think it useful to assume physicians are practical authorities to save space, nothing turns on this issue—if, as many theorists seem to think, there can be practical authorities that lack coercive authority.

[37] Heidi Hurd takes a somewhat different position—although its ancestral lines to Raz are clear; as she puts her position: "law can at best provide us with reliable moral advice, but cannot provide us with any reasons to do what morality otherwise prohibits" (Hurd 1999, xiii). She goes on to claim that law can have a form only of theoretical authority (a species of epistemic authority) and not practical authority. Whereas Raz seems to want to infer justified practical authority from justified

and does not obviously imply practical authority. *P*'s epistemic authority with respect to what *Q* should do according to right reason gives *Q* a reason to believe what *P* has said *Q* should do, but it does not necessarily give *Q* an exclusionary reason to do what *P* has said—and it certainly cannot, by itself, give rise in *Q* to a justified expectation that *P* complies. The mere fact that a physician, who has epistemic authority with respect to my physical health, informs me that (1) I should take an antibiotic because (2) I have a bacterial infection gives me a reason to believe both (1) and (2). But while that fact, *by itself*, might give me a reason to take the antibiotic, it is implausible to believe that the character of the reason is an exclusionary reason. Further, the idea that the physician's epistemic authority, alone, creates in the physician a morally justified expectation that I comply seems straightforwardly false. If a physician has practical authority over a patient, it cannot be explained in terms of the physician's epistemic authority alone. Something else in the physician–patient relationship will also be needed to explain the physician's practical authority.

Stephen Darwall makes a similar argument, although his reasoning differs in that it relies on his problematic conception of second-personal reasons. As he puts the point:

> Meeting the standards of the normal justification thesis is not, however, sufficient to establish practical authority. There are cases where one person might very well do better to follow someone else's directives where it seems clear that the latter has no claim whatsoever on the former's will and actions and consequently no practical authority with respect to him. And cases where an "alleged subject" would do better in complying with independent reasons where genuine authority does seem to be involved also seem to involve some assumed background accountability relation that gives the authority's directives standing as second-personal reasons. In these cases, it is the latter that establishes the directives' authority, not the former (Darwall 2009).

Darwall argues, in effect, that more is needed than just epistemic authority to give rise to practical authority over another person; in particular, what is needed is that the subject must be, for other reasons, antecedently accountable to the person who is the epistemic authority. Lacking those other accountability relations, epistemic authority is insufficient to give rise to practical authority. That much is surely correct.[38]

Darwall is surely correct in thinking that NJT fails as a sufficient condition for justifying practical authority. But his remarks above are nonetheless problematic inasmuch as they rely on a deeper account of practical authority that more accurately explicate the notion of *standing* than the notion of practical authority. The problem is that Darwall believes that the norms of morality, insofar as they protect us from certain acts, make each of us a practical authority over every person who is ultimately accountable to us. He believes that it is our status as practical authorities that explains

epistemic (or theoretical) authority, Hurd argues that the best law can provide is theoretical authority. It cannot have the kind of practical authority that would provide a content-independent exclusionary reason to do something that violates moral requirements.

[38] Scott Hershovitz argues that satisfaction of NJT is not a sufficient condition for legitimacy because matters having to do with whether a state employs democratic procedures are also relevant. See Hershovitz (2003), and note 14, above.

why we have a justified demand that obligations owed to us be satisfied and can make justified claims to that effect.

But this stretches the notion of practical authority well beyond its intuitive boundaries. We are not accurately characterized as practical authorities over others with respect to obligations owed to us insofar as we are not the source of those obligations—at least not in the sense of "practical authority" that conceptual jurisprudence wishes to explicate. If God, for example, manufactures morality, then God is a practical authority and God's directives are authoritative; that we enjoy the protections of these directives does not, on any remotely plausible conception of authority, entail that we are all practical authorities with respect to claiming satisfaction of those obligations. But if morality consists of objective truths that are independent of God's willings or commands, then the norms of morality are authoritative without there being a practical authority, if practical authority is, by nature, personal in character. As we have seen in the case of a legal system, there can be authoritative directives that do not originate with a practical authority; however, if some person's directives are authoritative, it must be because that person is the source of the directive and is a practical authority. While Darwall is correct in thinking that P's epistemic authority over Q cannot give rise to practical authority over Q in the absence of preexisting norms making Q accountable, his views about what gives rise to these accountability relations—which stem from his problematic account of second-personal reasons and practical authority—are unhelpful in securing the point.

NJT is even less successful, given its content, in providing a moral justification for using coercive means to enforce those directives against me. No matter how much better a de facto authority might be than I am in discerning the requirements of right reason, that fact alone is not sufficient to justify using coercive force against me to ensure that I obey her judgments.

Nor does satisfaction of NJT seem necessary for an authority to be morally legitimate. Although consent-based theories of legitimacy generally hold that subject consent to authority is, by itself, sufficient for the legitimacy of an authority over that subject and hence that NJT is not necessary for legitimacy, this is too strong. My consent to follow authority, in and of itself, is not enough to give rise to a content-independent moral obligation to follow that authority. Insofar as the legitimacy of an authority over a person is wholly grounded in that person's consent to it, its continuing legitimacy depends on that person's continuing consent. The problem is that, on any ordinary conception of unilateral consent, a person's consent, since voluntary, can be withdrawn at any given moment—and it is as implausible to suppose that a person can give irrevocable effective consent to be bound by a state *for her entire life* as it is to suppose a person can give effective consent to being a slave.

Accordingly, if my moral obligation to follow the directives of an authority is grounded *entirely* in my unilateral consent, that obligation can be extinguished at any time by my withdrawal of consent. Indeed, on such an account, it would be difficult to distinguish the obligation that authority gives rise to form non-obligatory behaviors that I am free to engage in or refrain from at my disposal because I am always free to withdraw and renew my consent to authority at my discretion.

Consider, for example, the authority of an arbitration (which Raz takes as paradigmatic of practical authority): My unilateral consent, without more, to follow the directive of an arbitrator regardless of its content cannot morally obligate me to follow it. For if the authority of the arbitrator over me is *wholly* grounded in *my* consent, then the arbitrator's authority over me is terminated simply by my withdrawal of that consent.

What is missing is the reliance of others on my consent—and such reliance typically takes the form of a corresponding commitment to obey the directives of an authority. If you and I both consent to abide by an authority's directives, then we are both forgoing options that would otherwise be available. There are different ways to explain how this gives rise to a moral obligation on the part of each of us to obey the directives. One might take a strict contractarian view and conceptualize our joint consenting as constituting a contract that gives rise to the obligation. Or one might argue that it would be unfair to allow someone to reap a benefit from disobedience given that other people have abdicated such a benefit. But, however, this is done, a key element in the legitimacy of authority is typically thought to rest on the express or implied consent of all persons over whom the authority is legitimate.

Raz is, of course, aware of the importance of consent to most theories of legitimate authority, but he rejects the idea that it is consent that binds the individual: "[a]greement or consent to accept authority is binding, for the most part, only if conditions rather like those of the normal justification thesis obtain" (Raz 1994, 214). The key element of his normative theory of legitimacy is the superior expertise of the authority in determining what subjects ought to do according to right reason; subject consent is of little to no importance on Raz's view.

It is worth nothing that Raz gives little, if anything, by way of argument for this conclusion. Contractarian theories of moral legitimacy have been widely regarded as plausible for centuries. These theories go back as far as Hobbes and Locke, and they remain of considerable contemporary influence in theories that are as different as the theories of Rawls and Nozick. Given the tremendous influence that such theories have enjoyed, Raz needs more argument than he gives to reject the central importance of consent to legitimacy.

Certainly, there are limits on the extent to which consent gives rise to moral obligations. As Raz points out, there are certain restrictions on how an authority may decide what a subject must do—even if the subject consents. Consider the context of a legal system. A judge, for example, could not legitimately decide a legal dispute on the basis of a coin flip, again, even if the parties consent to the judges doing so. Likewise, mutual consent and reliance are not enough to rescue a bargain if it is sufficiently unfair—either because of the bargain's content or because it was not negotiated at arm's length.

But these are exceptional circumstances and not the general rule with respect to the relation between consent and authority. If the parties are capable of giving effective consent to authority and the consent is secured in a fair manner, then the conditions articulated by NJT are not necessary for consent to authority to create a moral obligation to obey the directives of that authority. If, e.g., one party foreseeably relies to her detriment on the other's consent to abide by the directives of an authority, then the latter is bound by her consent, even if she gave her consent for reasons other than those described in NJT—indeed, even if she gave her consent for imprudent or ill-advised reasons. If this is correct, then NJT is neither a sufficient condition nor a necessary condition for the legitimacy of authority.

Raz would respond that he intends NJT as providing neither necessary nor sufficient conditions for justifying practical authority; his claim is the considerably weaker one that satisfaction of NJT is the "normal" way to justify that "[one] person should be acknowledged as having practical authority over another" (Raz 1994, 214). Accordingly, his response is that such criticisms misrepresent his position with respect to NJT.

There are two problems with this response. To begin, the claim that this is the *normal* way to justify practical authority is an empirical claim that would have to be supported with sociological evidence that has not been provided. Even so, there is little reason to think that people commonly cite epistemic authority as sufficient to justify practical authority, which is what the word "normal" connotes. It is surely true, as an empirical matter, that people commonly justify *A*'s accepting a piece of advice on the ground that its source, *B*, is more likely to be correct about what *A* should do. But practical authority differs from advice in morally salient ways. Unlike advice, for instance, practical authority involves a content-independent moral obligation to do what the authority's directives require, as well as a justified expectation on the part of the practical authority that the subject complies with the authority's directives. If the authority may use coercive enforcement mechanisms to enforce her directives, the problem becomes still more complex, from a moral standpoint. While Raz's model of advice is a useful heuristic device for understanding the motivation for NJT, the morally significant differences between being an advisor and being a practical authority are such as to require much more by way of justifying practical authority than by way of justifying being an advisor. Raz's theory of justifying authority seems problematic, no matter how it is construed—whether as providing necessary and sufficient conditions or as expressing the *normal* way to justify authority.

References

Christiano, T. 2013. Authority. In *The Stanford Encyclopedia of philosophy*, ed. E.N. Zalta. http://plato.stanford.edu/archives/spr2013/entries/authority/.

Darwall, S. 2009. Authority and Second-Personal Reasons for Acting. In *Reasons for Action*, ed. J. Wall, and D. Sobel. Cambridge: Cambridge University Press.

Dworkin, R. 2002. Thirty Years On. *Harvard Law Review* 115: 1655–1687.

Frankena, W. 1966. The Concept of Morality. *Journal of Philosophy* 63: 688–696.

Gert, B. 2012. The definition of morality. In *The Stanford Encyclopedia of philosophy*, ed. E.N. Zalta. http://plato.stanford.edu/archives/fall2012/entries/morality-definition/.

Gilligan, C. 1982. *In a Different Voice*. Cambridge, Mass.: Harvard University Press.

Green, L. 1989. *The authority of the state*. Oxford: Oxford University Press.

Hart, H.L.A. 1982. Commands and authoritative reasons. In *Essays on Bentham*, ed. H.L.A. Hart. Oxford: Oxford University Press.

Hershovitz, S. 2003. Legitimacy, democracy, and Razian Authority. *Legal Theory* 9: 201–220.

Himma, K.E. 2001. Law's Claim to Authority. In *Hart's postscript: essays on the postscript to the concept of law*, ed. Jules L. Coleman. Oxford: Oxford University Press.

Himma, K.E. 2007. Just 'cause you're smarter than me doesn't give you a right to tell me what to do: legitimate authority and the normal justification thesis. *Oxford Journal of Legal Studies* 27: 121–150.

Himma, K.E. 2016. *Can there really be law in a society of angels?* Available at SSRN: https://ssrn.com/abstract=2839942 or http://dx.doi.org/10.2139/ssrn.2839942.

Himma, K.E. 2017. The authorization of coercive enforcement mechanisms as a conceptually necessary feature of law. *Jurisprudence* 7: 593–626.

Hurd, H. 1991. Challenging authority. *Yale Law Journal* 100: 1611–1677.

Hurd, H. 1999. *Mortal Combat*. Cambridge: Cambridge University Press.

Kohlberg, L. 1984. *The Psychology of moral development: The nature and validity of moral stages*, vols. 1 and 2. New York, N.Y.: Harper and Row.

May, T. 1998. *Autonomy, authority and moral responsibility*. Dordrecht: Kluwer Academic Publishers.

Perry, S. 1989. Second order reasons, uncertainty, and legal theory. *Southern California Law Review* 62: 913–994.

Rosen, G. 2014. Abstract Objects. In *The Stanford Encyclopedia of Philosophy*, ed. E.N. Zalta. http://plato.stanford.edu/archives/fall2014/entries/abstract-objects/.

Raz, J. 1975. *Practical reason and norms*. Oxford: Oxford University Press.

Raz, J. 1985. Authority and justification. *Philosophy & Public Affairs* 14: 3–29.

Raz, J. 1994. *Ethics in the public domain*. Oxford: Oxford University Press.

Shapiro, S. 2011. *Legality*. Cambridge, Mass: Harvard University Press.

Simmonds, J.A. 2001. *Justification and legitimacy: Essays on rights and obligations*. Cambridge: Cambridge University Press.

Wolff, R.P. 1970. *Defense of Anarchism*. New York, N.Y.: Harper and Row.

Zagzebski, L.T. 2012. *Epistemic authority: A theory of trust, authority, and autonomy in belief*. Oxford: Oxford University Press.

The Authority of Law

Veronica Rodriguez-Blanco

Law transforms our lives in the most important way: it changes how we act and because of this it gives rise to fundamental questions. One such question concerns legal authority and individual autonomy and asks; if we are autonomous agents how do legislators, judges and officials have legitimate authority to change our actions and indirectly change how we conduct our lives? We conceive ourselves as active agents who determine how and when to act, and we conceive ourselves as the planners of our own lives and the creators of change. Law asks us, however, to perform actions that range from the trivial to the complex. Law requires us, for example, to stop at traffic lights; park our vehicles in specially allocated areas; exercise our professional judgment in a responsible and non-negligent manner; pay our taxes; recycle our rubbish. Law asks us to perform innumerable tasks, almost all of which we perform intentionally and in full awareness. But how is it possible for me to do, in full awareness, as the law asks and, at the same time, be in control of my own destiny? How is my free will affected by the law?

But how is this possible when I am simply trying to conform with what the law says? This means, I am trying to follow what the law says without giving much thought or without engaging my will or intention.

Legal and political philosophers have tended to examine legal authority and autonomy and have consequently put forward the following questions: (a) Can there ever be legitimate authority?; (b) What are the conditions of legitimate authority? and (c) Does the possibility of legitimate authority diminish or assuage the antagonism between authority and autonomy?

I find that posing the problem and the questions in this way is unsatisfactory because it presupposes what we need to explain; i.e. the nature of authority and

The entry relies on material previously published in Rodriguez-Blanco (2014b, 2017).

V. Rodriguez-Blanco (✉)
School of Law, University of Surrey, Guilford, UK
e-mail: v.rodriguez-blanco@surrey.ac.uk

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_9

whether there is a "genuine" antagonism between autonomy and legal authority. Within this framework authority is given, and the starting point of the theorist is the following statement: If there is a legitimate authority then conditions *x*, *y* and *z* need to be fulfilled, but it is not shown how there is or whether there could be something such as legitimate authority. The received view begins by recognizing the phenomenological fact that legal officials and authorities issue commands and directives. It is usually said that if authorities have the right to command and addressees the duty to obey, then the officials have legitimate authority.

Theorists usually argue in favour of a particular political theory, for example liberalism or perfectionism, and engage with a set of key values, for instance expert knowledge or democratic values that provide the grounds for "rights" and "duties" and that enable us to grasp the conditions of legitimate authority. The traditional strategy, therefore, begins top-down from a plausible view on political theory that leads to the framework that justifies authority. There is no doubt that the traditional strategy has provided us with a rich understanding that has advanced our grasp of the normative conditions that make possible legitimate legal authority. However, the traditional strategy fails to provide a microscopic view of the phenomenon of legal authority and falls short of explaining how legal authority truly operates on individual human beings.

By contrast, the strategy of this study is to focus on the agent; i.e. the addressee of the legal command or directive who performs the action requested by the legal official. This strategy is bottom-up, from the level of agency and practical reason to the justificatory framework of authority. It also begins with the naive phenomenological observation that *X* commands *Y* to perform the action *p* (an action *p*-ing to *Y*). Thus, it is intelligible to us that *Y* performs the action *p* as requested by *X*. The key question that this study aims to investigate is how a legal command or directive, just because it is a legal command or directive, effectively changes the agent's course of action. A set of sub-questions arise: Does the command intervene in the practical reasoning of the agent or addressee? If this is the case, how does this intervention operate? Moreover, what are the limits of our phenomenological observations, in other words can I truly observe that you are performing an action because you are complying with a legal directive or command? What happens in the agent that enables her to comply with the legal command or directive? When we perform an action because we are complying with the legal command or directive, are we still active, self-governed autonomous agents? In what sense are we still autonomous agents? The task of this study is to explain what legal authority is and the premise of the study is that this question can only be answered through understanding of how legal authority operates upon the agent: if we recognize that legal commands or directives intervene upon, affect and change the agent's practical reasoning, then we need to understand and explain how this happens.

1 Human Action and Authority: Tracing the Correct Relationship

It is recognized by theorists and laymen that law is a social practice. However, if social practices are constituted by human actions then the following question arises: "What is the sound characterization of human action that enables us to provide a satisfactory explanation of the production of authoritative legal rules and directives?" The key feature of legal rules and directives is that they guide the behaviour of citizens and has a normative force on the addressees of legal rules and directives. It is, however, puzzling how human beings are able through their actions to produce such a complex state of affairs; i.e. a legal rule that is authoritative and intervenes in the reasoning and actions of the addressees of the rule.

Let us suppose that we explain human action as merely an empirical phenomenon, i.e. a set of regular patterns produced by the reason-beliefs or acceptance-beliefs of the participants, which are construed as mental states.[1] Within this framework of explanation, the authoritative character of the legal rules, their guiding role and normative force are utterly mysterious. For example, let us think about the legal rule that demands that citizens stop at red traffic lights and also about citizen "c" who does this numerous times every morning when driving to work. Following the empirical model of human action, the empiricist will say that citizen *c*'s action is explained by the fact that "there is a rule that is grounded on reasons that respond to what everyone does" (Lewis 1969); or, rather, "there is a rule that is grounded on our accepted reason-beliefs towards such a rule or accepted reason-beliefs towards a second-order rule about such a rule" (Hart 2012); or, even more, "there is a rule that is the result of deep conventions, which are the result of social practices, responsive to our social and psychological needs, arbitrary, grounded on a reason-belief to follow them, instantiated in superficial conventions and resistant to codification" (Marmor 2007. For a criticism of this view, see Rodriguez-Blanco 2016).

We feel, however, that there is something fundamentally missing in this purely empirical portrait of human action. It seems to imply that if one day citizen "c" decides not to do what everyone does or accepts, and decides instead not to stop at the red traffic lights, and consequently her vehicle collides with a number of other vehicles and she kills a child, then (following the empiricist explanation of human action) the only mistake she made in her reasoning that led her to the catastrophic action is that she did not accept what everyone accepted, or rather she did not have the appropriate reason-belief as mental state to follow the rule. This is a strange understanding of her reasoning, though it follows logically from an explanation of human action in terms of purely empirical features; i.e. social facts, beliefs or intentions as mental states, and reasons explained in terms of beliefs and therefore mental states.

[1] According to the empirical account of intentional action, the acceptance of legal rules provides reasons for actions in the context of the law. For a full explanation of the empirical account of action in the context of the law and its criticism, see Rodriguez-Blanco (2014b, Chap. 5). I argue that the empirical account of intentional action is parasitic on the "guise of the good" explanation of intentional action.

The explanation of the reasoning of the agent in empirical terms is equally unintelligible in examples where what is at stake is the life, dignity or another fundamental value that we human beings care about. Let us scrutinize the following example. If an official aims to enforce the court decision that has established that citizen “p” has violated the physical integrity of another citizen and therefore should be punished with imprisonment and we ask for an explanation of the official’s coercive action, it would be puzzling to hear the following response: “Citizen “p” has violated a constitutional rule which is grounded on our acceptance-belief or reason-belief which lies behind the constitutional rule.” This value-free or value-neutral response cannot truly explain why citizen “p” has to go to prison according to a court decision. Does it mean that if citizen “p” escapes from the coercion of the official and manages to leave the country, then the only mistake in her reasoning that leads her to flee the country is her disagreement with either the acceptance-belief that there is a valid constitution or secondary rule, or her disagreement with the acceptance-belief towards the constitutional rule and penal code that protects the physical integrity of all citizens? Thus, it is not that she disagrees with the value that is the content of the acceptance-belief or reason-belief, rather she disagrees with the acceptance-belief or reason-belief. The disagreement is just about beliefs, and therefore, according to the empirical account, the parties in disagreement are in different mental states. This is an equally strange and puzzling diagnosis of our disagreements.

When we characterize what legislators, judges, officials and citizens do in terms of actions as empirical phenomena, we seem to miss something fundamental. Worse, the empirical account of action cannot satisfactorily explain the guiding role of the law.[2]

Let us go back to our first example. Citizen “c” is a law-abiding citizen who aims to follow and be guided by the law, and on her journey to work, she knows there is a legal rule that states she ought to stop at red traffic lights. According to the empirical characterization of human action, she stops at red traffic lights because she has the acceptance-belief or reason-belief that there is such a rule and this acceptance-belief or reason-belief causes her to press the brake pedal on each relevant occasion. If she were asked why she presses the brake pedal she will reply, “because there is a red traffic light,” and if she were asked, “why do you stop at the red traffic light?” she would reply, “because there is a secondary rule that is accepted by the majority of the population and this establishes the validity of the rule ‘citizens ought to stop at red traffic lights’.” Alternatively, she might reply, “I stop at the red traffic light because of the rule,” but now the mystery is “why do you act according to the rule?,” to which she might answer, “because rules give me reasons for actions.” The empirical account explains reasons in terms of beliefs/desires as mental states (Davidson 1980), and

[2] Arguably, Raz’s explanation of how legal rules intervene in our reasoning is non-empirical since he has emphasized that a reason for action should not simply be understood as beliefs as mental states (Raz 1979, 1986, 1999). However, in Rodriguez-Blanco (2014b, Chap. 8), I argue that Raz’s explanation of legal authority is a theoretical explanation of our reasoning capacities; i.e. when we explain how legal directives and rules intervene in the citizen’s practical reasoning from the third-person perspective. His explanation ignores the first person or deliberative point of view of the citizen who follows legal rules.

then it seems that it is the mental state that is causing the action. This is a problematic picture because it supposes that for each action I need to "remember" my belief/desire so that I am able to be in the right mental state so that I can stop at the red traffic light. However, we stop at red traffic lights even when we are tired or when we do not "remember" (Wittgenstein 1953, Section 645) that we ought to stop at red traffic lights, and therefore we somehow just "know how to go around" and stop at red traffic lights. Furthermore, the predominant empirical picture of human intentional action cannot explain the diachronic structure of intentional action. That is, we stop at red traffic lights over a prolonged period of time and even though the relevant mental state might be absent, we still continue doing it and it seems that we do it for a "reason" that tracks values or good-making characteristics.

Imagine that there is an emergency. Citizen "c" needs to bring her neighbour to the hospital because he is dying and consequently she decides not to stop at a red traffic light. Does this mean, if we follow the empirical account of human action, that in order to explain her action we need to say that she surely needed to "forget" that she had the relevant belief as mental state of "stopping at red traffic lights," or perhaps she decided "to get rid" of her acceptance-belief concerning the rule "we ought to stop at red traffic lights"? Or, perhaps, she "decided to suspend" her beliefs about the rule "we ought to stop at red traffic lights." In the two latter possibilities, we cannot say that it is not only a "belief" that plays a role in action, but rather the "will" of the rule-follower. She has used the words "get rid of" and "decided to suspend." It seems that there is something else going on. Imagine that we ask her, "why did you not stop at the red traffic light?" The empirical answer, "because I do not have the acceptance-belief or reason-belief towards the rule now for this specific instance," would be an odd one. Citizen "c" is more likely to say, "Don't you see it? My neighbour is dying and I want to save his life." Furthermore, if reasons for actions are belief-acceptance or if they give me a reason for action and this reason is merely a mental state, how can I be guided by rules and principles? If the empirical explanation of action is the sound characterization, then the guidance of rules and principles is effective because I am in the correct mental state. The entire work is done by my mental states as long as I am in the supposedly correct mental state. The deliberation of the legislator or judge and/or my own deliberation plays no role in the execution of my action of rule-following or principle-following. The content of the legal rule is irrelevant as long as the majority of the citizens are in the allegedly correct mental state.

Arguably, we need to resort to values in the form of good-making characteristics that are relevant to the specific form of life that is ours and that reflect what we care about individually and collectively. We need to understand human action in its naïve or fundamental form and this understanding, I argue, sheds light on the kind of things we produce, including human institutions such as law. Thus, if someone asks citizen "c" why she stops at the red traffic light on her way to work, there is a naïve explanation of her action that seems to be more primary than any other explanation. Thus, she might respond, "because I do not wish to collide with other vehicles and kill pedestrians." If we ask her, "why do you not wish to collide with other vehicles and kill pedestrians?," she most is likely to reply, "because I value my property, other people's property, and life" and if we keep asking, "why do you value property and life?," she will respond, "because property and life are goods."

We have learned that a mistaken conception of human action can take us down misleading routes in our understanding of the nature of law and, more specifically, its pervasive, authoritative, normative and guiding role in our lives. The sound explanation of human action will illuminate how human beings produce law and will also shed light on the authoritative and normative features of law. In Sect. 2, I explain and defend a conception of human action that diverges from the standard empirical conception. In Sect. 3, I scrutinize the consequences of this conception of human action for our understanding of the nature of law and its authoritative and normative character.

2 Intentional Action Under the Guise of the Good

We will now concentrate on intentional human action as the paradigmatic[3] example of human action to shed light on the making of law by legislators and judges and the character of legal rule-following.[4]

In her book *Intention* (1957) Elizabeth Anscombe engages with the task of explaining intentional action along the lines of the philosophical tradition of Aristotle and Aquinas and identifies a number of key features that characterize intentional action. These features include:

(a) The former stages of an intentional action are "swallowed up" by later stages

Intentional action is composed of a number of stages or series of actions. For example, if I intend to make a cup of tea, I first put on the kettle in order to boil water; I boil water in order to pour it into a cup of tea. Because my action of making tea is intentional, I impose an order on the chaos of the world and this order is the order of reasons. Thus, I put on the kettle in order to boil water and I boil water in order to pour it into a cup. This is how I understand the sequence of happenings in the world that I, as an agent, produce or make happen. But, arguably, there could be an infinite number of series of actions; there could be a continuous infinite, or ceaseless, seamless web of actions. The question "Why?" can always be prompted: "Why are you making tea?" and the agent might reply, "Because it gives me comfort in the morning." There is, however, an end to the "Why?" series of questions and the end comes when the agent provides a characterization of the end or telos as a good-making characteristic. The action becomes intelligible and there is no need to ask "Why?" again. The end as the last stage of the "Why?" series of questions swallows up the former stages of the action and makes a complete unity of the action. Intentional actions are not fine-grained, they are not divisible into parts. Thus, parts

[3] For a defence of a conception of paradigms as the best methodology to understand social and human concepts see Rodriguez-Blanco (2003).

[4] I am using the term "rule-following" but the same explanation applies to principle-following. See Rodriguez-Blanco (2012, 2014a).

of series of actions are only intelligible because they belong to an order that finds unity in the whole.

(b) Intentional action is something actually done, brought about according to the order conceived or imagined by the agent

Intentional action is not an action that is done in a certain way, mood or style (Anscombe 1957, Section 20). Thus, it is not an action plus "something else"; i.e. a will or desire that is directed towards an action. Intention is not an additional element; e.g. an interior thought or state of mind, it is rather something that is done or brought about according to the order of reasons that has been conceived by the agent. Consequently, if the question "Why?" has application to the action in question, we can assert that the action is intentional. The prompting of the question "Why?" is the mechanism that enables us to identify whether there is an intentional action. Intentional action is neither the mere movements of our body nor the simple result of transformations of the basic materials upon which agency is exercised, e.g. the tea leaves, kettle, boiling water. It is a doing or bringing about that is manifested by the expression of a future state of affairs and the fact that the agent is actually doing something or bringing it about according to the order of reasons as conceived or imagined by the agent (Anscombe 1957, Section 21–22).

(c) Intentional action involves knowledge that is non-observational, but it might be aided by observation

What is the distinction between practical and theoretical knowledge? Let us take a modified version of the example provided by Anscombe (1957, Section 32). A man is asked by his wife to go to the supermarket with a list of products to buy. A detective is following him and makes note of his actions. The man reads in the list "butter," but chooses margarine. The detective writes in his report that the man has bought margarine. The detective gives an account of the man's actions in terms of the evidence he himself has. By contrast, the man gives an account of his actions in terms of the reasons for actions that he himself has. However, the man knows his intentions or reasons for actions not on the basis of evidence that he has of himself. His reasons for actions or intentions are self-intimating or self-verifying. He acts from the deliberative or first-person perspective. There is an action according to reasons or an intention in doing something if there is an answer to the question why. It is in terms of his own description of his action that we can grasp the reasons for the man's actions. In reply to the question "why did you buy margarine instead of butter?," the man might answer that he did so because it is better for his health. This answer, following Aristotle's theory of action[5] and its contemporary interpretations advanced by Anscombe provides a reason for action as a desirability or good-making characteristic. According to Anscombe, the answer is intelligible to us and inquiries as to why the action has been committed stops. However, in the case of the detective when we ask why did you write in the report that the man bought margarine, the answer is that it is the truth about the man's actions. In the case of the detective, the knowledge is theoretical, the detective reports the man's actions in terms of the

[5] Aristotle (1934). Nicomachean Ethics I. i. 2; III. V. 18–21. See also Aquinas, Summa Theologiæ. I–II, q8, a1, Kenny (1979), Pasnau (2002), Finnis (1998, 62–71 and 79–90).

evidence he has of it. In the case of the man, the knowledge is practical. The reasons for action are self-verifying for the agent. He or she does not need to have evidence of his own reasons for actions. This self-intimating or self-verifying understanding of our own actions from the deliberative or practical viewpoint is part of the general condition of access to our own mental states that is called the "transparency condition."[6] It can be formulated as follows:

(TC for reasons for actions) "I can report on my own reasons for actions, not by considering my own mental states or theoretical evidence about them, but by considering the reasons themselves which I am immediately aware of."

The direction of fit in theoretical and practical knowledge is also different. In the former case, my assertions need to fit the world whereas in the latter, the world needs to fit my assertions. The detective needs to give an account of what the world looks like, including human actions in the world. He relies on the observational evidence he has. The detective's description of the action is tested against the tribunal of empirical evidence. If he reports that the man bought butter instead of margarine, then his description is false. The man, by contrast, might say that he intended to buy butter and instead bought margarine. He changed his mind and asserts that margarine is healthier. There is no mistake here.

The idea that we accept from the internal point of view primary or secondary legal rules (Hart 2012) presupposes an inward-looking approach to action as opposed to an outward-looking approach. The latter examines intentional actions as a series of actions that are justified in terms of other actions and in view of the purpose or end of the intentional action as a good-making characteristic, e.g. to put the kettle on in order to boil the water, in order to make tea because it is pleasant to drink tea. The former examines the mental states that rationalize the actions; however, at the ontological level, arguably, it is mental states that cause the actions. The mental states consist of the belief/pro-attitude towards the action. If the "acceptance thesis" is the correct interpretation of Hart's central idea concerning the internal point of view towards legal rules, then criticisms that are levelled against inward-looking approaches of intentional actions also apply to the "acceptance thesis" (Hart 2012). The main criticism that has been raised against the idea that the belief/pro-attitude pairing can explain intentional actions is the view that it cannot explain deviations from the causal chain[7] between mental states and actions. Let us suppose that you intend to kill your enemy by running over him with your vehicle this afternoon when you will meet him at his house. Some hours before you intend to kill your enemy, you drive to the supermarket, you see your enemy walking on the pavement and you suffer a nervous spasm that causes you to suddenly turn the wheel and run over your enemy. In this example, according to the belief/pro-attitude view, there is an intentional action if you desire to kill your enemy and you believe that the action of killing your enemy, under a certain description, has that property. Ontologically, the theory would establish that you had both the desire to kill your enemy and the belief

[6] See Evans (1982), 225 and Edgeley (1969). The most extensive and careful contemporary treatment of the "transparency condition" is in Moran (2001).

[7] The first person to discuss deviant causal chains was Chisholm (Chisholm 1976).

that this action has the property “killing your enemy.” Thus, this mental state has caused the action and there is an intentional action. The problem with this view is that it needs to specify the “appropriate causal route.” Davidson has made much effort to specify the “attitudes that cause the action if they are to rationalize the action” (Davidson 1980, 79). In the following paragraph, Davidson seems to fear that the idea of attitudes causing action might lead to infinite regress:

A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to lose his hold, and yet it might be the case that he never chose to loosen his hold, nor did he do it intentionally. It will not help, I think, to add that the belief and the want must combine to cause him to want to loosen his hold, for there will remain the two questions how the belief and the want caused the second want, and how wanting to loosen his hold caused him to loosen his hold.

Here we see Davidson struggling with his own proposal (for an illuminating discussion of this point see Vogler (2007)). He asks how attitudes must cause actions if they are to rationalize actions. Davidson’s model of intentional action does not help us to determine whether there is an intentional action, it only helps us to determine the conditions that would explain the existence of an intentional action. The intentional action is already given. A similar criticism is applicable to the “acceptance thesis” and to this we now turn.

Let us suppose that I intend to go to the park in my car however I read a sign at the entrance of the park that states “Vehicles are not allowed to park in the park,” I turn the wheel of my vehicle, reverse it and park a few streets away. You ask me why I turned the wheel of my vehicle, reversed and parked a few streets away from the park; I answer that I carried out these actions because there is a rule that states “Vehicles are not allowed to park in the park.” According to the “acceptance thesis,” my desire to follow the pattern of behaviour indicated by the rule and my belief that turning the wheel of my vehicle, reversing it and not parking in the park is the type of action or pattern of behaviour indicated by the rule. However, let us suppose that I desire to avoid parking in the park and have the respective belief. In other words, I accept “not parking in the park.” On my way to the park, however, while following directions to the park, I take a wrong turning and end up parking just outside the park entrance. Even though the two criteria of the “acceptance thesis” have been met, this was not a case of following the legal rule by acceptance since I comply with the rule by accident.

The problem with the “acceptance thesis” is that it does not consider the action from the deliberative point of view; i.e. as it is seen from the point of view of the agent or deliberator. When the agent explains his actions he does not examine his own mental actions, rather he looks outwards to the vehicle, the park, the sign and so on. The reasons for actions, i.e. turning the wheel to reverse the vehicle, then parking outside the park to follow the rule, are self-evident or transparent to him. But then, an objector might advance, what is the good-making characteristic of a rule that, as in the example of the shopper who intends to buy margarine because is healthier, is the goal of the action of avoiding parking in the park. My reply is as follows. When the driver is asked why he or she is turning the wheel and reversing the vehicle, his answer will be "because it is the rule." But this is still not completely intelligible unless we assume or know that the driver is a law-abiding citizen or that he believes in the general fairness of legal rules, etc. We can still ask him, "Why, because of the rule, do you do this?" His answer would need to be in terms of reasons as good-making characteristics for him, in order to make intelligible his intentional action. He will probably reply that he has reasons to follow the legal rule because it is the best way of preserving the peace of the park, or that he has reasons to follow legal rules in general because it is the best way of preserving coordination[8] among the members of a community. In a nutshell, the agent or deliberator needs to provide the reasons for the action in terms of good-making characteristics and the end or reason of the action provides the intelligible form of the action. This explanation of action has also been called a naïve explanation of action as opposed to a more sophisticated explanation of action; i.e. in terms of mental states.

If I am an agent that acts in an intentional way, I know that I am bringing about something and I know this without the need to observe every single step of my series of actions to verify that (effectively) I am acting (Anscombe (1957), Section 28–29). In performing my action I might be aided by observation, but I know what is the order of the series of actions and why. This is the essence of practical knowledge. You do not need a theoretical stance towards yourself, a verification and observation of the movements of your body to know that you are performing an intentional action and bringing about something. Following the previous example, you do not need to observe that "you are making tea" to know that you intend to "make tea" and that you are bringing this about. You put on the kettle and boil the water, you do not ask yourself, "let me see what my body is up to, let me observe what I am doing," and then infer from the movements of your body that you are actually bringing about "making tea." Of course you can be aided by observation, you need your sight to put the kettle in the right position and to pour the boiling water without spilling it. But you do not use your observation and inferences from the observational data to know that you are making tea.

The state of affairs that you intend to bring about is at a distance, it might not be within your sight (Anscombe 1957, Section 29–30). Imagine a painter who intends to make a painting. He has an idea about what the painting will look like, e.g. how the colours will be distributed across the canvas, and what topics and concepts will be at work in the painting. The painting is at a distance and the painter does not need to observe the movements of his body and the motion of the brushes to know what he is painting and why he is painting what he is painting. Certainly, his sight will help him to find the adequate colour at the correct time and to shape the figures at the right angle, but his intentional action is not what he observes; it is not the result of his painting but what he is actually doing. We do what happens.

(d) In acting intentionally, we exercise our practical knowledge. We can understand practical knowledge if we understand the structure of practical reasoning

[8] See Anscombe (1981) for an argument of authority as practical necessity.

Intentional action is not in the mind; it is not primarily a mental state; it is not an internal thought (Anscombe 1957, Sections 21–22, 25, 27–28). Rather, it manifests itself publicly and within the public reasons that we share as creatures with certain constitutions and belonging to a particular time and place. For example, we eat healthy food because it is good to survive; we look after our family because we love them; we avoid harm because we aim to enjoy pleasant things and so on. Similarly, we know that to make a cake you need flour, sugar, eggs and milk. If I see you mixing grass and earth and you tell me that you are making a cake, then I can assert, if I consider that you are in sound mind (your full capacities), that there might be a mistake in your performance or that you do not understand what it is "to make a cake."

According to Anscombe, Aristotle establishes a strong analogy between practical and theoretical syllogism and this has led to misinterpretations about what practical syllogism is (Anscombe 1957, Sections 33, 33–34). Like theoretical syllogism, practical syllogism is often systematized by Aristotelian interpreters as having two premises, i.e. major and minor, and a conclusion. It is said that, as in the case of theoretical syllogism, the practical syllogism is a proof or demonstration. The typical form might be as follows:

Vitamin X is good for all men over 60
Pig's tripe is full of vitamin X
I am a man over 60
Here is pig's tripe

But in this case nothing seems to follow about doing anything. Furthermore, the practical syllogism is sometimes interpreted as having an ethical or moral character and establishing a way to prove what we ought to do. Following the previous example, the conclusion might be, "I should eat pig's tripe." Anscombe rejects this view since Aristotle's examples are not in ethical contexts, i.e. "dried food is healthy," "tasting things that are sweet" that are pleasant. Additionally, the word "should" (*dei*) as it appears in the Aristotelian texts has an unlimited number of applications and does not necessarily refer to the ethical or moral context (Anscombe 1957, Section 35).

Aristotle insists that the starting point of any intentional action is the state of affairs or something that the agent wants and is wanted because it is presented to the agent as having good-making characteristics or as being valuable. For example, the man wants to have vitamin *X* because it is healthy. Furthermore, the practical syllogism is not limited to two premises and a conclusion; there can be many intermediate instances that are part of the syllogism. After a close analysis, the analogy between practical and theoretical syllogism breaks. Unlike theoretical syllogism, practical syllogism is not a proof or demonstration of a true proposition, nor is it a proof or demonstration of what ought to be done or what we ought to do. It is a form of how and why we are bringing something about when we are actually bringing it about.

Anscombe presents us with an alternative analysis to the practical syllogism and a different way to understand practical reasoning. Thus, the series of responses to the question "Why?" manifests or reveals the practical reasoning of the agent and enables us to identify whether the action that the agent is performing is intentional or not. However, she warns us, the why-question methodology is as "artificial" as the Aristotelian methodology of practical syllogism (Anscombe 1957, Section 41–42). When we act intentionally, we are exercising a kind of reasoning which is not theoretical and which is grounded on a desire for that which seems to the agent to be constituted by good-making characteristics. You know the thing or state of affairs that you are bringing about because you desire the thing or state of affairs that you are bringing about, and you are able to desire the thing or state of affairs that you are bringing about because you know practically the state of affairs. Your desire arises because you represent the thing or the state of affairs to be brought about as valuable or good. Volition and knowledge do not fall apart (Anscombe 1957, Section 36). For example, if you are a painter, you know how and why the shapes and colours on the canvas are what they are, it is because you desire and value the painting you will produce that it should be such and such a colour and shape. But it is also true that because you desire and value this and not that arrangement of colours and shapes that you are able to know it practically. Consequently, moral approbation is irrelevant for practical reasoning and for our practical engagement with the world (Anscombe 1957, Section 37–38). This does not mean that there are no instances of objectively justified reasons for actions. On the contrary, we aim at getting it right and finding the genuine good-making characteristics that will provide meaning and intelligibility to the movement of our bodies. Therefore, the possibility of hitting the target of genuine good-making characteristics resides in our good characters and capacities. But to understand the basic structure of practical reason and the different scopes of agency, we do not need to begin from fully justified and objective values.[9]

Whatever strategy we follow to show the structure of intentional action, whether we take the Aristotelian practical syllogism or the Anscombian series of actions revealed by the question "Why?," we are able to grasp the mechanism of practical reasoning in its different manifestations.

In this section, I will argue that if Anscombe is right and both strategies are "artificial" ways of understanding (Anscombe 1957, Section 41–42), then a deeper and more "natural" way of understanding practical reasoning is by grasping the nature of the capacity that is exercised by the agent. In other words, the answers to the "Why?" questions show a capacity that the agent is exercising when acting. In the next section, I will show that the Aristotelian potentiality/actuality distinction sheds light on understanding the exercise and nature of our practical reasoning capacities. Furthermore, the potentiality/actuality distinction illuminates each of the key features of intentional action (*a*, *b*, *c,* and *d*) and their interplay as identified by Anscombe.

[9] In Chapter 9 of Rodriguez-Blanco (2014b), I show that robust value realism is indispensable to making sense of our actions, practices and first-order deliberative phenomenology. See Chap. 3 for a full defence of the "guise of the good model." See also Grisez's interpretation of Aquinas's precepts of natural law in Grisez (Grisez 1969, 368).

2.1 Aristotle's Distinction Between Actuality and Potentiality

Contra Parmenides, who argued that motion is impossible since something cannot come from nothing, Aristotle advances the idea that motion or change is possible if there is an underlying nature or constant feature that does not change. To explain this, Aristotle resorts to the distinction between potentiality and actuality. In *Metaphysics*, *book Θ*, Aristotle uses the analogical method to show that particular instances of the scheme or idea of potentiality and actuality[10] have a pattern.[11] Thus, he begins with the particular instances of capacity/change and matter/form to explain the common patterns that will illuminate the general scheme of potentiality/actuality. However, since our purpose is to elucidate the character of practical reasoning which is a power or capacity, and I have argued that the general scheme of potentiality/actuality will help us to clarify the nature of practical reason, it is circular to resort now to a particular instance of capacity/change to explain potentiality/actuality. I will, therefore amend the Aristotelian argumentative strategy and explain the general scheme of potentiality/actuality. I will then proceed to explain the particular instance of exercising our practical capacities as the actuality of a potentiality.

It is difficult to capture what "motion" is and many definitions of "motion" tend to use terms that presuppose motion (e.g. "a going-out from potency to act which is not sudden," but "going-out" presupposes motion and "sudden" is defined in terms of time which is also defined in terms of motion). Therefore, this kind of definition is discarded by Aristotle for being circular and unhelpful. Nor can we define motion in terms of pure potency, because if we say that "bronze is potentially a statue," we are merely referring to the piece of bronze which has not yet been changed and therefore there is no motion. You can neither refer to motion nor to change as what is actual. For instance, you cannot refer to what has been built or transformed, e.g. a building or statue, because it is not being moved, but has already moved. In the example of a building, the bricks, wood, clay, cement of the building have been already moved; and in the case of a statue, the bronze has already been transformed. Thus, Aristotle defines motion as a kind of actuality which is hard to grasp. In other words, the actuality of what exists potentially, in so far as it exists potentially (Aristotle, *Physics*, III.1.201a9–11). Motion is an actuality that is incomplete. It is hard to grasp and the tendency is to say that motion is the actuality. In the example of the house, it is the house that has been built. The other tendency is to say that motion is the privation of something; i.e. the going from nothing to something, from not being a house to being a house. Finally, the tendency is also to think that motion is what exists before potentiality, e.g. the bricks, steel, wood, cement. Contrary to these tendencies, Aristotle insists that motion is what happens exactly at the midpoint,

[10] I use this term as Kosman and Coope interpret it from Aristotle's *Physics, Books III and IV*. This means, the change that acts upon something else so that this something else becomes *F*; i.e. the fulfilment of a potentiality. For example, the building of a house by a builder so that the house becomes built. See Kosman (1969) and Coope (2009).

[11] I follow the interpretation of Aristotle's *Metaphysics book Θ* advanced by Frede (1994) and Makin (Aristotle 2006, 133). Cf. also Ross (1995).

neither before when nothing has been moved and is mere potentiality, and neither after, when something has been moved. Furthermore, motion is not privation; it is rather constitutive actuality. For example, if the baby has not learned to speak English, we say that the baby is potentially a speaker of English, when a man knows how to speak English and is in silence, he is also potentially a speaker of English, and finally when the man is speaking English, we say that he is actually an English speaker speaking English. However, the potentiality of the baby (*p*1) is different from the potentiality of the man in silence (*p*2), and motion is located in the second potentiality (*p*2), when the man is in silence, but begins to pronounce a sentence to speak English. Motion is midway and is not privative, but rather constitutive. We do not say that the man speaking English went from being a non-speaker of English to a speaker of English, we say that he spoke English from being in silence (he knew how to speak English, but did not exercise his capacities).

The previous example locates us in the domain of the particular instance of capacity and change as exemplified by the potentiality/actuality distinction. Aristotle argues that there are many different types of capacity, i.e. active/passive, non-rational/rational, innate/acquired, acquired by learning/acquired by practice and one-way/two-way capacities. Two-way capacities are connected to rational capacities, whereas one-way capacities are linked to non-rational capacities. For example, bees have a natural capacity to pollinate a foxglove flower in normal circumstances (Makin in Aristotle (2006), 43), ("normal" circumstances might include a healthy bee in an adequate foxglove, and the absence of preventive circumstances). In the case of two-way capacities there ought to be an element of choice or desire to act, and the rational being can exercise her capacity by producing or bringing about "p." Furthermore, she also knows how to produce or bring about "non-p." The paradigmatic example used by Aristotle is medical skill. The doctor knows how to make the patient healthy (*p*) and how to provoke disease or illness (non-*p*). Therefore, the doctor can bring about two opposite effects (Aristotle, *Metaphysics, Book Θ*, 1046b 4–5, 6–7). For Aristotle, to have a rational capacity is to have an intellectual understanding of the form that will be transmitted to the object of change or motion. Thus, the doctor will have an understanding of what it means to be healthy and without illness, but also of what it means to be ill. Let us suppose that a doctor is producing illness in the enemies through prescribed drugs. She needs to understand the order of the series of actions that will result in sickness for the enemies and she needs to possess knowledge about the necessary drugs to make the enemies to collapse. Her action will be directed to produce illness. But the doctor can choose otherwise, e.g. she can choose to make the enemy healthy.

In the exercise of practical reason, we choose to act (Aristotle, *Metaphysics, Book Θ*, 5, 1048 a10–11) and this choosing activates the action and directs the capacity towards the series of actions that will be performed. By contrast, a non-rational capacity is non-self-activating; its acts are necessary. If the bee is in good health and there are no obstacles, it will pollinate the foxglove flower. By contrast, rational agents need to choose or decide to act to produce a result.

When we say that the medical doctor has the rational capacity to change the unwell patient into a healthy human being, we say that she has the "origin of change." She is curing the patient and therefore she is in motion because she actualizes her practical reasoning capacities to bring about the result as she understands it. She has an order of reasons that connects a series of actions and knowledge of how to produce changes.

She is the origin of change because her medical knowhow explains why certain changes occur in situations involving that object, e.g. the patient who suffers chickenpox has fewer spots and less fever. For example, when a teacher intends to teach and starts to say some sentences on the topic of "Jurisprudence" to her pupils, we say that she is teaching. She is the origin of change in the pupils who are the objects of change. Thus, the students begin to understand the topic and have a grasp of the basic concepts.[12] Similarly, when legislators create the law and judges decide cases, they establish rules, directives and principles and these rules, directives and principles can be found in statutes and case reports. Can we say that legislators and judges have reached the end of the process? No, we cannot: statutes and case reports do not represent the end of the process since citizens need to comply with the legal rules and directives and perform the actions as intended by the legislators and judges. We say that legislators and judges are the origin of change because they know how and have an order of reasons that enables citizens to comply with legal rules and directives. The order or reasons as good-making characteristics ground the rules, decisions and legal directives. In parallel to the situation of the teacher, I cannot say that I am teaching unless my pupils begin to understand the topic that I am teaching. Thus, the legislator cannot say that she is legislating and the judge cannot say that she is judging, in paradigmatic cases, unless there is some performance of their actions by the addressees as they intend.

The distinction between potentiality/actuality clarifies the structure of practical reason as a capacity that is actualized when we act intentionally. We can now understand that the features of an intentional action identified by Anscombe can be illuminated by the potentiality/actuality distinction. The idea that the former stages of an intentional action are swallowed up by the later stages is explained by the idea that motion is constitutive and not privative. It is not that when I begin to act I do so as an irrational or a rational being, and that I when finish acting I am a rational being, or that I go from non-intentional to intentional action, but rather that I go from being a rational being and potentially intentional action to being a rational being and actual intentional action. Later stages begin to actualize something that was potentially there. My practical reason was always there potentially and the intentional action actualizes an order of ideas provided by my practical reason. For Anscombe, intentional action is something actually done, brought about according to the order conceived or imagined by the agent. If practical capacity is understood in the light of the general scheme of actuality/potentiality, then intentional action involves knowledge that is non-observational, but it might be aided by observation. In acting intentionally, I am exercising my practical reasoning capacity and this capacity is in motion. This motion is represented at the midpoint; after I potentially have an

[12] Makin argues that the teacher analogy is intended to show that the teleological perspective is equally appropriate for other-directed capacities and self-directed capacity. See Aristotle (2006), 198.

intention to act and before I have reached the result of my intentional action. It is not that the forming of an intention from nothing to something is a magical process. It is rather that I potentially have the power to intend which in appropriate circumstances can be exercised. As being in motion, I am the agent who knows what she is doing and why she is doing what she is doing, but if I observe myself doing the action, then I have stopped the action. There is no action. There is no more motion and no exercise of my capacities. Finally, Anscombe asserts that in acting intentionally, we exercise our practical knowledge. Because we are the kind of creatures that we are, we can choose or decide to bring about a state of affairs in the world and we do this according to our order of reasons. Practical knowledge is potentially in all human beings and when we decide to bring about a situation or do certain things, then we actualize this potentiality. We can direct our actions to produce either of two opposing results, e.g. health or illness, ignorance or knowledge, as opposed to non-rational creatures who can only produce one result under normal circumstances and with no impeding conditions, e.g. the bee pollinating the foxglove. It should be noted that to have an actual capacity, such as practical reasoning and the capacity to act intentionally, does not mean that A can Φ, nor that A will Φ if there are normal conditions and no impending elements. Instead, it means that A will Φ unless she is stopped or prevented. Thus, once our practical reasoning capacity begins to be actualized, it will strive to produce or do what A (she) has conceived. Once A (she) decides or chooses to act, then a certain state of affairs will be produced unless she is prevented or stopped. Intentional action and practical reasoning are not dispositions like being fragile or elastic, nor are they possibilities that something will be done. They are powers.

Now that we have grasped the idea of potentiality/actuality as the general scheme for explaining the structure of practical reason, we can turn to the rule-compliance phenomenon and the creation of legal rules by legislators and judges, which raises a different set of difficulties that will be dealt with in the next section.

3 Law and *Energeia*: How Do Citizens Comply with Legal Rules?

So far we have argued that an intentional action is the bringing about of things or states of affairs in the world. We can argue, too, that there are different kinds of bringing about. Human beings can produce houses, clocks, tables, teacups and so on, but we can also produce rules of etiquette, rules for games, and legal directives, rules, and principles. Legislators create legal rules and directives and judges create decisions according to underlying principles and rules. These legal rules and directives are directed to citizens for them to comply with. They are meant to be used in specific ways. When a legislator creates a rule or a judge reaches a decision that involves rules and principles, she creates them exercising her practical capacities with the intention that the citizens comply with them. But how is this compliance possible? How do legislators and judges create legal rules and directives that have the core purpose of directing others' intentional actions and of enabling them to engage in bringing about things and states of affairs in the world? In other words, how do other-directed capacities operate? This is the question that we aim to explore in this section.

Let us give two examples of authoritative commands to highlight the distinction between different kinds of authoritative rules:

Scenario 1 (REGISTRATION): you are asked by a legal authority to fill in a form that will register you on the electorate roll.

Scenario 2 (ASSISTANCE AT A CAR ACCIDENT): you are asked by an official to assist the paramedics in a car accident, e.g. to help by transporting the injured from the site of the accident to the ambulance, to assist by putting bandages on the victims, to keep the injured calm.

Arguably, the performance required by the addressee is more complex in the latter example than in the former since the latter requires the engagement of the will and the performance of a series of actions over a certain period of time, and it requires that the addressee should circumvent obstacles to achieve the result according to what has been ordered. It requires that the addressee exercises her rational capacity in choosing this way rather than that way of proceeding. While the addressee executes the order she needs to make judgments about how to do this or that. Successful performance as intended entails knowledge about how to proceed at each step in order to perform the series of actions that are constitutive of what has been commanded. This cannot be done unless our practical reasoning and intentional action are involved in the performance. In other words, the successful execution of the order requires the engagement of practical reasoning and therefore of our intentions. Furthermore, it requires an understanding of the telos or end as a good-making characteristic of what has been commanded. In the case of ASSISTANCE AT A CAR ACCIDENT, it requires engagement with the health and well-being of the victims of the accident. Thus, the addressee needs to know that the bandage ought to be applied in this way and not that way in order to stop the bleeding, and she knows that she needs to stop the bleeding in order for the victim to have the right volume of blood in his body. The victim needs a certain volume of blood in his body in order to be healthy and being "healthy" is something good and to be secured.

Because our practical reasoning capacity is a two-way capacity the agent needs to decide or choose to actualize this capacity which, prior to actuality, is mere potentiality. As in our example in Sect. 2.1, the speaker needs to decide or choose to speak in order to actualize their potentiality of speaking English. Then the exercise of their capacity to speak actualizes according to a certain underlying practical knowledge, e.g. the order of the sentences, grammar, style. It is not the case that as a bee pollinates a foxglove without any decision or choice by the bee, the agent will speak English and actualize their potential capacity to speak. In the case of legal rules, the question that emerges is how a legislator or judge can produce or bring about something that will engage the citizens' intentions so that they comply with legal rules or directives that are constituted by a complex series of actions. The core argument is that legislators and judges intend that citizens comply with legal directives and rules, and this intention is not merely a mental state that represents accepted reasons or reason-beliefs. On the contrary, for the legislators' and judges' intentions (i.e. to engage the citizens' practical reasoning) to be successful, they need to exercise their own practical reason. It is not that they interpret or construct the citizens' mental states and interior thoughts so that their values and desires can constitute the ground that enables legislators, judges and officials to construct the best possible rules, directives or legal decisions according to the citizens' values as represented in their beliefs. On the contrary, they will look outward to what is of value and why certain states of affairs and doings are valuable (see the discussion on practical knowledge as non-observational Sect. 2, c). Reasons for actions as values and goods that are the grounds of legal rules and directives will engage others' practical reason. Therefore, the citizens' practical reasoning power or capacity become an actuality. If, as I have argued, our intentional actions become actuality by an order of reasons in actions and for actions that are ultimately grounded on good-making characteristics, then legislators and judges need to conceive the order of reasons as good-making characteristics that will ground their legal rules, legal directives and decisions. Judges and legislators would hence take the first-person deliberative stance as the privileged position of practical reasoning to disentangle what good is required and why it is required. In other words, if as judge or legislator you intend that your legal rule or directive is to be followed by the addressees and, arguendo, because these legal rules and directives are grounded on an order of reasons, then you cannot bring about this state of affairs, i.e. rule-compliance, without thinking and representing to yourself the underlying order of reasons. Let me give a simple example. You are writing an instruction manual on how to operate a coffee machine. You need to represent to yourself a series of actions and the underlying order of reasons to guide the manual's users. If you are a person of certain expertise, e.g. a manufacturer of coffee machines, then the practical knowledge that entails the underlying order of reasons is actualized without much learning and thinking. The required operating instructions are actualized as a native English speaker speaks English, after being in silence. By contrast, if you have only just learned to write instruction manuals for coffee machines, then you need to ask yourself "Why do it this way"? at each required action to make the machine to function. This process guarantees understanding of the know-how to operate the machine, and the success of the manual is measured by the fact that future buyers of the coffee machine are able to operate it. When legislators and judges create legal directives and legal rules they operate like the writers of instruction manuals, though at a more complex level. They need to ensure that the addressees will decide or choose to act intentionally to comply with the legal rules or directives and thereby bring about the intended state of affairs. But they also need to ensure that the order of reasons is the correct one so that the intended state of affairs will be brought about by the addressees. We have learned that the early stages of an intentional action are "swallowed up" by the later stages and ultimately by the reason as a good-making characteristic that unifies the series of actions. Thus, for addressees with certain rational capacities and in paradigmatic cases, understanding the grounding reasons as good-making characteristics of the legal rules and legal directives will enable them to decide or choose to comply with the rule and will guide them through the different series of actions that are required for compliance with the rules and directives.

Legal rules and directives do not exist like houses, chairs, tables or cups of tea. We need to follow them for them to exist. But we create legal rules and directives as we create houses, chairs, tables. We bring these things about by exercising our practical capacity and we are responsive to an order of reasons as good-making characteristics that we, as creators, formulate and understand. Thus, builders create houses that are either majestic or simple, elegant or practical, affordable or luxurious. To achieve the intended features of a house, builders need to select specific materials and designs, hire skilled workers, and so on. Similarly, legislators, officials and judges create legal directives and rules to pursue a variety of goods, e.g. to achieve safety, justice, the protection of rights. Legislators, officials and judges actualize their practical reasoning by creating an order of reasons in actions that will ground rules so that we are able to comply with them because we actualize our practical reasoning. Like builders, legislators, officials and judges need to choose values, goods and rights that will be fostered or protected by their rules or directives. Likewise, they need to formulate legal rules and directives that will have appropriate sanctions and are clearly phrased and followed procedures for their publicity. Arguably, what is at stake is not the mere publicity of a rule, but the publicity of the values that are embedded in the set of legal rules and principles. In this way, judges make the addressee of a directive choose or decide to actualize their potential practical reasoning capacity to comply with legal rules and directives. The addressees of a legal directive or rule are not like bees, who without decision and, given normal conditions and the absence of impediments, will pollinate the foxglove. As addressees of legal directive and legal rules, we need to choose or decide to bring about a state of affairs or things which are intended by the legislator, official or judge. This can be summarized as the idea that legal authority operates under the guise of an ethical-political account since it needs to present legal rules and directives as grounded on reasons for action as good-making characteristics.

As rational creatures, we are responsive to reasons as grounded in good-making characteristics, but if this is truly the case, how do mere expressions of doing as brute facts, such as "because I said so," or beliefs, intentions or reasons construed as mere mental states make possible the actuality of our practical reason? In fact, this is only possible if "because I said so" involves reasons in action that are grounded in good-making characteristics, e.g. "I am the authority and compliance with the authority has good-making characteristics." For example, compliance with authority is a secure way that some goods—apparent or genuine—will be achieved. The potentiality/actuality and capacity/change discussion show that as intellectual and rational beings, we need to apprehend the "form" that underlies the brute fact "because I said so," so that we are able to comply with legal directives and rules. As theoreticians, we now understand the limits of the empirical explanation of action, i.e. it has no "form" that makes intelligible the actuality of our practical reason and explains the dynamic reality of our intentional actions. Of course, we can decide that there is no such a thing as practical reason and that it is perfectly reducible to theoretical reason,[13] but

[13] See Enoch (2011) for a recent defence of the reductive approach. See Rodriguez-Blanco (2012) for a criticism of his position.

then the price we pay for this simple approach is too high: it leaves a set of human actions and the phenomenology of our first order or deliberative stance in the mists of mystery.

The "form" takes the shape of goods and values that are intended to be achieved by legislators, officials and judges. If it were a matter of mental or social facts, and we were able to apprehend the brute fact "because I said so" by our senses, or access legislators' and judges' reasons and values via our mental states only, without directly engaging with values and reasons, then how could we control and direct the doings and bringing about that are intended by legislators and judges? Some stages of the action will seem this and other stages will seem that. There is no way to bring about this and not that. Let us take the example of ASSISTANCE AT THE CAR ACCIDENT. I assist the official at the car accident because he has said so. I have no reason to assist him at the car accident; my action is only caused by my fear of sanction; i.e. a psychological impulse in me. But now as I am merely guided by my senses, it seems to me that I need to put the bandage on in this way rather than that way, but my sight alone cannot guide me on this. Since I am guided by my eyes and other senses, I do not know why I should apply the bandage or how I should apply the bandage. Furthermore, how can we attribute responsibility as we cannot be blamed for not "seeing" or "hearing" appropriately? By analogy, mere scribbles on the board by the teacher cannot make the pupil understand the topic that the teacher is teaching. The teacher needs to make transparent the premises and conclusions of her arguments so that the pupils can "grasp" the form of the argument and can themselves infer its conclusion.

Let us return to our initial example. Citizen "c" stops at the red traffic lights on her way to work. If we ask her "why are you stopping at the red traffic lights?" and we are satisfied with the empirical explanation which is, "because there is a secondary rule that is accepted by the majority of the population and this establishes the validity of the rule 'citizens ought to stop at red traffic light'," then how can we attribute responsibility to citizen "c," who just happen to have certain mental states? How can citizen "c" produce the required action just by remembering her mental state? By contrast, within the framework of the notion of practical reason that we have defended in this article, she will naïvely reply, "because the legal rules say so," and to reach intelligibility we could continue by asking, "why do you follow what the legal rules say?." She could then naïvely reply, "because I do not wish to damage my vehicle or other vehicles and I do not wish to kill other people." We can try to reach yet further intelligibility of her actions and ask, "why do you not wish to damage other people's vehicles or kill people?," and her reply will be, "because property and life are valuable."

We are now in a position to understand that citizen "c's" answers have a structure which is the structure of practical reason, where reasons are connected to other reasons, whose chain has a finality. The finality is provided by the agent from the first person or deliberative perspective when she advances a value or good-making characteristic that swallows the earlier stages of the action and provides intelligibility to the movements of "c's" body. This explanation seems primary and more fundamental than the explanation in terms of acceptance-beliefs, reason-beliefs as a mental state of either primary or secondary rules of the legal system or exclusionary reasons.[14]

If citizen "c" decides not to stop at the red traffic light because she is driving her neighbour to the hospital, who is dying, then to the question "why are you not stopping at the red traffic light?," she might reply, "don't you see it? My neighbour is dying and I need to get to the hospital as soon as possible." And to the question, "why do you need to bring him to the hospital as soon as possible?" she might reply, "because I want to save his life"; in response to the question "why do you want to save his life?," the answer will be, "because life is valuable." This set of answers will give intelligibility to her actions, which includes the movements of her body and what she produces, i.e. a vehicle moving in the direction of the hospital, and will also explain why she did not stop at red traffic lights. Thus, she went through the red traffic light not because of her belief that on this occasion there was no valid legal rule, nor because of her belief that the rule of "stopping at red traffic lights" does not protect or ensure values such as property or life. Her mistake lies, arguably, in not "perceiving" that the life of her neighbour is as valuable as the lives of pedestrians and the drivers of other vehicles. Her mistake lies in her understanding of the goods or values at conflict in the particular situation.

The classical model of practical reasoning and intentional action laid out the view that for an action to be controlled and guided by the agent the reasons need to be in the action and therefore transparent to the agent (see Sect. 2, c). The answers to the question "Why"? provide the order of reasons that guarantees successful compliance with the legal rules and directives by the agent. They are the reasons in action that the agent has together with the values or good-making characteristics that the legislator and or judges aim to promote and want the citizens to "grasp" as the grounding of their actions. The transparency condition of practical reason warrants that the citizen is able to engage with the good-making characteristics that ground legal rules. But if the order of reasons is opaque, how can there be an action as intended by the legislator or judge as an order of reasons that has as a finality a value or good-making characteristics? If the reasons are opaque and you do something "because someone says so" you do not know "why" you are performing the action and therefore the action is not intentional. Furthermore, one might assert, the legislator, judge or official is not the origin of change and the origins of change are in external empirical factors, e.g. the fear mechanism that acts within the agent, psychological processes in the agent, mental states such as beliefs, acceptance-belief or reasons-belief.

References

Anscombe, E. 1957. *Intention*. Cambridge, Mass: Harvard University Press.

Anscombe, E. 1981. On the source of authority of the state. In *Ethics, religion and politics: collected philosophical papers*, ed. E. Anscombe. Hoboken, N.J.: Wiley-Blackwell.

[14] Raz's exclusionary reasons account (Raz 1999) privileges the theoretical point of view. See also note 2 above.

Aquinas. 2006. *Summa theologiæ*, trans. T. Gilby. Cambridge: Cambridge University Press.

Aristotle. 1934. *Nicomachean ethics,* trans. H. Rackham. Cambridge, Mass: Harvard University Press.

Aristotle. 1983. *Physics books III and IV*, trans. E. Hussey. Oxford Clarendon Press.

Aristotle. 2006. *Metaphysics book Θ,* trans. S. Makin Introduction and commentary. Oxford: Oxford University Press.

Chisholm, R. 1976. Freedom and action. In *Freedom and determinism*, ed. K. Lehrer. New York, N.Y.: Random House.

Coope, U. 2009. Change and its relation to actuality and potentiality. In *A companion to Aristotle*, ed. G. Anagnostopoulos. Hoboken, N. J.: Wiley-Blackwell.

Davidson, D. 1980. *Essays on actions and events*. Oxford: Oxford University Press.

Enoch, D. 2011. *Taking morality serious: A defense of robust realism*. Oxford: Oxford University Press.

Edgeley, R. 1969. *Reason in theory and practice*. London: Cornerstone-Hutchinson.

Evans, G. 1982. *The varieties of reference*. Oxford: Oxford University Press.

Finnis, J. 1998. *Aquinas*. Oxford: Oxford University Press.

Frede, M. 1994. Aristotle's notion of potentiality in metaphysics Θ. In *Unity, identity and explanation in Aristotle's metaphysics*, ed. T. Scaltsas, D. Charles, and M. Gill. Oxford: Clarendon Press.

Grisez, G. 1969. The first principle of practical reason, a commentary on the summa theologiae, 1–2, Question 94, article 2. In *Aquinas. A collection of critical essays*, ed. A. Kenny. London: MacMillan.

Hart, H.L.A. 2012. *The concept of law.* Oxford: Clarendon Press (1st ed. 1961).

Kenny, A. 1979. *Aristotle's theory of the will*. London: Duckworth.

Kosman, L.A. 1969. Aristotle's definition of motion. *Phronesis* 14: 40–62.

Lewis, D. 1969. *Convention*. Cambridge, Mass.: Harvard University Press.

Marmor, A. 2007. Deep conventions. *Philosophy and Phenomenological Research* 74: 586–610.

Moran, R. 2001. *Authority and estrangement*. Princeton, N.J.: Princeton University Press.

Pasnau, R. 2002. *Thomas aquinas on human nature*. Cambridge: Cambridge University Press.

Raz, J. 1979. *The authority of law*. Oxford: Oxford University Press.

Raz, J. 1986. *The morality of freedom*. Oxford: Oxford University Press.

Raz, J. 1999. *Practical reason and norms*. Oxford: Oxford University Press.

Rodriguez-Blanco, V. 2003. Is finnis wrong? *Legal Theory* 13: 257–283.

Rodriguez-Blanco, V. 2012. If you cannot help being committed to it, then it exists: A defence of robust realism in law. *Oxford Journal of Legal Studies* 32: 823–841.

Rodriguez-Blanco, V. 2014a. Does practical reason need interpretation? *Ragion Practica* 317–340.

Rodriguez-Blanco, V. 2014b. *Law and authority under the guise of the good*. Oxford: Hart-Bloomsbury.

Rodriguez-Blanco, V. 2016. Re-examining deep conventions: practical reason and forward-looking agency. In *Metaphilosophy of law*, ed. T. Gizbert-Studnicki. Oxford: Hart-Bloomsbury.

Rodriguez-Blanco, V. 2017. Practical reason in the context of law: What kind of mistake does a citizen make when she does violate legal rules? In *Cambridge companion to natural law jurisprudence*, ed. R. George, and G. Duke. Cambridge: Cambridge University Press.

Ross, W.D. 1995. *Aristotle's physics: a revised text with introduction and commentary*. Oxford: Oxford University Press.

Vogler, C. 2007. Modern moral philosophy again: isolating the promulgation problem. *Proceedings of the Aristotelian Society* 106: 345–362.

Wittgenstein, L. 1953. *Philosophical investigations,* trans. E. Anscombe. Oxford: Blackwell.

Part II
Kinds of Reasoning and the Law

Deductive and Deontic Reasoning

Antonino Rotolo and Giovanni Sartor

1 Introduction

This chapter offers a concise and elementary introduction to fundamental concepts in deductive and deontic reasoning.[1]

Reasoning is based on arguments, and arguments can be simply viewed as ways to state that, whenever one asserts some claims $P_1, \ldots, P_n$, one engages in a commitment to other claims $C_1, \ldots, C_m$ that follow from $P_1, \ldots, P_n$[2]: in this sense, $P_1, \ldots, P_n$ are the premisses of $C_1, \ldots, C_m$, which are in fact the conclusions of $P_1, \ldots, P_n$. To keep things simple—and unless differently specified—we will assume without loss of generality to work with arguments with one or more premisses $P_1, \ldots, P_n$ and one conclusion C.[3] Arguments can thus be graphically represented as

$$\begin{array}{c} P_1 \\ \vdots \\ P_n \\ \hline C \end{array}$$

or as

$$\langle\{P_1, \ldots, P_n\}, C\rangle.$$

[1] The reader is not required to have any advanced technical background. Some basic knowledge on general philosophy, elementary set theory and propositional logic can be useful but not necessary.

[2] More on the concept of argument in Walton, chapter 3, part I, this volume, on "Legal Reasoning and Argumentation."

[3] Indeed, if there is an argument from $P_1, \ldots, P_n$ into $C_1, \ldots, C_m$, it is usually the case that there are arguments for each conclusion in $C_1, \ldots, C_m$ from $P_1, \ldots, P_n$, and vice versa.

A. Rotolo (✉) · G. Sartor
Dipartimento di Scienze giuridiche, Università di Bologna, Bologna, Italy
e-mail: antonino.rotolo@unibo.it

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_10

Under these assumptions, *deductive arguments* are intuitively the ones that, *if valid*, guarantee their conclusions *with no exception*, i.e. in all possible cases. A standard example of deductively valid argument is the following:

$$\frac{\begin{array}{l}\text{All humans are mortal}\\ \text{Plato is a human}\end{array}}{\text{Plato is mortal}} \tag{1}$$

It is intuitive to read (1) as an argument where the statements above the line work as the premisses for the *indisputable* conclusion that Plato is mortal (given such premisses).

Orthogonally to the idea of deductive reasoning, *deontic reasoning* consists of arguments where premisses or conclusions are deontic claims—i.e. statements about concepts such as obligations, permissions and prohibitions—and where the fact that conclusions *validly* follow from such premisses *essentially* depends on deontic nature of the claims involved in the argument. Here below is an example:

$$\frac{\text{It is obligatory for citizens to pay taxes}}{\text{It is permitted for citizens to pay taxes}} \tag{2}$$

There are good reasons to consider (2) as a valid deductive argument, since it may sound strange to assume that, in some cases, something that is obligatory is not permitted. However, (2) is essentially different from (1), because the validity of the former depends on the very nature of obligations and permissions.

The choice of treating deontic reasoning together with deduction can be somehow seen as arbitrary or partial, since the former type of reasoning does not necessarily fall within the latter one. However, for decades the greatest part of research in deontic logic has characterised deontic concepts especially in the context of classical logic, i.e. using a reference standard for deductive reasoning. In the remainder, I will briefly mention the main reasons in deontics that suggest in some cases to go beyond deductive reasoning, and I will refer to other chapters of this handbook for a clarification of those alternative forms of reasoning.

2 Deductive Reasoning

As we have just mentioned, a *valid deductive argument* $\mathscr{A} = \langle \Gamma, C \rangle$ is intuitively one where C follows *with no exception* from the premisses Γ. Hence, a fundamental issue is to clarify what we mean by saying "validity with no exception." Philosophers and logicians can often explain this intuition by resorting to the idea of *necessity*. The following is a common definition grasping this approach:

Definition 1 An *argument* $\mathscr{A} = \langle \Gamma, C \rangle$ is *deductively valid* iff it is *impossible* for C to be false while the elements of Γ are true.

The above definition is standard, but let us leave for a moment aside the concept of truth, and reframe it into the following:

Definition 2 An *argument* $\mathscr{A} = \langle \Gamma, C \rangle$ is *deductively valid* iff it is *impossible* for C not to be the case while the elements of Γ are.

Definition 2 means that $\mathscr{A}$ is deductively valid if, and only if, C is *necessarily* the case whenever Γ are so. This definition relies, as we said, on the idea of necessity: that C follows *with no exception* from the premisses Γ is a matter of *necessity*.

However, the concept of necessity is far from obvious and univocal, since one may consider different types of necessity concerning diverse domains (see Fine 2002; Kment 2017), such as the epistemic or metaphysical ones. Indeed, if a variety of modalities allows for grounding deduction on a solid philosophical basis and, e.g. for reconstructing, within a sound framework of deductive reasoning, inferences from knowledge of essential properties of reality or to justified beliefs, it clearly raises the open question of clarifying to which domain Definition 2 correctly applies.

A way to clarify and somehow mitigate the problem is to work with the strongest idea of necessity, i.e. *logical necessity*. In this way, deductively valid arguments are nothing but *valid logical arguments*. Under this last assumption, the two options in the following definition should be considered (cf. Blanchette 2001):

Definition 3 Let $\mathscr{A} = \langle \Gamma, C \rangle$ be an argument.

1. $\mathscr{A}$ is *deductively valid* iff C is a *logical consequence* of Γ;
2. C is a *logical consequence* of Γ iff Γ *entails* C.

The concepts of logical consequence and of entailment are the cornerstones of any logical theory of deduction. The next two sections will introduce the reader to both options.

2.1 Deductive Reasoning—Logical Consequence as Deducibility

If deductive reasoning is conceived in terms of a logical consequence, one can say that $\mathscr{A} = \langle \Gamma, C \rangle$ is deductively valid means that C is *logically deducible* from Γ (cf. Etchemendy 1990), where

- the elements of Γ and C are expressed in a certain formal language $\mathscr{L}$, i.e. they are well-formed formulae in $\mathscr{L}$, and
- C is obtained from Γ by using certain principles for reasoning about statements expressed in that language.

Definition 4 (*Formal Language*) A formal language $\mathscr{L}$ is defined solely in terms of its syntax, by stipulating two components:

- *an alphabet*, i.e. a collection of symbols from which all expressions are built;

- *formation rules*, i.e. a grammar stating how strings of symbols can be obtained from the alphabet; such strings are the well-formed expressions (well-formed formulae) of the formal language.

A formal language is thus a class of sentences described by a formal grammar.

One of the simplest formal languages is the one for propositional logic (cf. Mendelson 1987). As is well-known, this propositional formalisation captures the basic structure of complex sentences by identifying their propositional units and their formal connections.

Example 1 (*Formal Language for Propositional Logics*) The language $\mathscr{L}$ of propositional logic consists of

- *alphabet*:
  - a set of small letters $p, q, \ldots$ to denote atomic propositions;
  - a set of capital letters $A, B, \ldots$ to denote arbitrary propositions;
  - brackets to secure unambiguous readability of well-formed formulae;
  - logical connectives $\neg$ (not), $\wedge$ (and), $\vee$ (inclusive or), $\rightarrow$ (if …then), $\equiv$ (if and only if) to build formulae from propositions;
- *formation rules*:
  - every atomic proposition is a well-formed formula;
  - If A and B are well-formed formulae, then $(\neg A)$, $(A \wedge B)$, $(A \vee B)$, $(A \rightarrow B)$ and $(A \equiv B)$ are well-formed formulae;
  - nothing except those defined in the steps above are well-formed formulae.

Once the language $\mathscr{L}$ is defined, we can introduce a deductive system $\mathscr{D}$ working on expressions of $\mathscr{L}$, i.e. a set of reasoning principles that guarantee in $\mathscr{D}$ to correctly derive conclusions from any Γ and to build *proofs* of these formulae, given Γ. A well-known reasoning rule in most propositional logics called Modus Ponens, is the following:

$$\frac{\begin{array}{l} A \rightarrow B \\ A \end{array}}{B} \tag{3}$$

Here is an example:

$$\frac{\begin{array}{l} \text{If All humans are mortal then no human in an angel} \\ \text{All humans are mortal} \end{array}}{\text{No human is an angel}} \tag{4}$$

Roughly, a proof for a formula C consists in a finite number of steps, ending with C, which satisfy the principles in $\mathscr{D}$. In this way, if we write $\langle \Gamma, C \rangle_{\mathscr{D}}$ we can say that C is *deducible* or *provable* in $\mathscr{D}$ from Γ. In a nutshell:

Definition 5 (*Logical consequence as deducibility*) Any well-formed formula C is a *logical consequence* in $\mathscr{D}$ of a set Γ of well-formed formulae, i.e. $\Gamma \vdash_{\mathscr{D}}$, iff C is *deducible* in $\mathscr{D}$ from Γ, i.e. iff there is a proof from Γ to C that satisfy principles in $\mathscr{D}$.

This is a *deductive- or proof-theoretic definition* of a logical consequence, which corresponds to a syntactic relation between well-formed formulae.

Before providing some more details on the concept of proof, let us briefly mention some properties of deductive consequence relation, which have recalled, for example, by Blanchette (2001). Indeed, we have already discussed the property of necessity which correlates premisses and conclusions, but other features can be added. Notice that such properties are widely discussed in the philosophical literature and are not indisputable (cf. Etchemendy 1990).

Topic-neutrality

A property that a deductive consequence relation can enjoy is the so-called *topic-neutrality*. This means that a validly deductive argument does not need to be "'about' the same subject-matter as are its premises" but that "the fundamental principles of logic must hold independently of any particular subject-matter" (Blanchette 2001, p. 116). In other words, deduction can be claimed to work on the basis of purely formal considerations.

A rather strong case is offered in classical propositional logic $\mathscr{P}$, where, e.g. a disjunction holds whenever at least one of the disjuncts holds, and so we indisputably have that $A \vdash_{\mathscr{P}} A \vee B$, independently of whether B has to do with A. A concrete example is as follows:

$$\frac{\text{All humans are mortal}}{\text{(All humans are mortal)} \vee \text{(The dinner is ready)}} \tag{5}$$

The general idea of topic-neutrality has been challenged, for example, by Relevance Logics (Anderson and Belnap 1975).

Epistemological Inertness

If deduction works on purely formal basis, a further assumption is that no deductive argument can change the epistemological nature of its premisses into something different in the conclusion: "the logical consequences of things knowable a priori, or knowable non-empirically or without the aid of intuition, are themselves, respectively, knowable a priori, non-empirically, without the aid of intuition [...]. There is a rough sense, then, in which the logical consequences of a claim have no 'new content' over and above that had by the original claim" (Blanchette 2001, p. 124).

This means that deduction can never license the inference of claims whose nature is like "All crows are black" from claims whose nature is like "All crows are birds."

Formal Conditions on Deductive Consequence

Other properties have a more formal import and concern how conclusions are in general obtained, i.e. how the relation $\vdash_{\mathscr{D}}$ formally behaves for any single system $\mathscr{D}$

or, more interesting, for classes of systems. Notably, Alfred Tarski (1936) was one of the first pointing out that, for any deductive system $\mathscr{D}$, the corresponding consequence relation must satisfy some minimal and intuitive properties[4] (cf. Gabbay 1994):

Reflexivity: If $C \in \Gamma$, then $\Gamma \vdash_{\mathscr{D}} C$;
Monotonicity: If $\Gamma \vdash_{\mathscr{D}} C$, then $\Gamma, \Gamma' \vdash_{\mathscr{D}} C$;
Transitivity (cut): If $\Gamma \vdash_{\mathscr{D}} A$ and $\Gamma, A \vdash_{\mathscr{D}} C$, then $\Gamma \vdash_{\mathscr{D}} C$.

Reflexivity looks intuitive, as any set of premisses Γ is supposed to prove each element of it. Monotonicity captures the fact that the accumulation of data does not affect conclusions, while "transitivity is nothing but lemma generation," i.e. if $\Gamma \vdash_{\mathscr{D}} A$, then A can be used as a lemma to derive C in $\mathscr{D}$ from Γ (Gabbay 2006, p. 743).

It should be noted that these properties can be questioned, but this usually means going beyond deduction in the strict sense (for a discussion in regard to normative reasoning, see Parent 2001; Rotolo 2017).

Let us confine ourselves to Monotonicity. In fact, there is a well-known link between challenging Monotonicity and modelling defeasible reasoning. Deduction "is monotonic: as long as we accept all premises of a deductive inference, we must continue to accept its conclusion. Therefore, we also say that deductive inference is conclusive: as long as we maintain the premises, any additional information will not affect the conclusion."[5] Defeasible reasoning is non-monotonic, i.e. it does not enjoy Monotonicity. However, Gabbay (1985) and Kraus et al. (1990) pointed out that a restricted version of Monotonicity can work for a large class of non-monotonic (and defeasible) systems without abandoning Reflexivity and Cut:

Cumulativity (Cautious Monotonicity): If $\Gamma \vdash_{\mathscr{D}} A$ and $\Gamma \vdash_{\mathscr{D}} B$, then $\Gamma, A \vdash_{\mathscr{D}} B$.

A variety of systems satisfy this new property (see Kraus et al. 1990). Finally, notice that the consequence relation corresponding to any *defeasible* system $\mathscr{D}$ can incorporate a *deductive* system $\mathscr{P}$, typically by concluding any B that is always deductively implied (via classical implication) by any A, which is a defensible conclusion as well:

Right Weakening (Kraus et al. 1990): If $\vdash_{\mathscr{P}} (A \rightarrow B)$ and $\Gamma \vdash_{\mathscr{D}} A$, then $\Gamma \vdash_{\mathscr{D}} B$.

More specific properties are rather the following (see Shapiro 2013):

Double-negation Elimination: If $\Gamma \vdash_{\mathscr{D}} \neg\neg C$, then $\Gamma \vdash_{\mathscr{D}} C$.

Double-negation Elimination holds for all systems of classical logic (see Shapiro 2013, Sect. 3), such as standard propositional logic—which adopts the formal language presented in Example 1—but it is rejected by those who do not claim that each sentence is either true or not true. Typically, Intuitionistic Logic does not accept it (see Dummett 2000).

[4] As usual, Γ, Γ' and Γ, A are abbreviations, respectively, for $\Gamma \cup \Gamma'$ and $\Gamma \cup \{A\}$.

[5] Sartor, Chapter 3, Part II, this volume, on "Defeasibility in Law."

Ex Falso Quodlibet: If $\Gamma \vdash_{\mathscr{D}} C$ and $\Delta \vdash_{\mathscr{D}} \neg C$, then $\Gamma, \Delta \vdash_{\mathscr{D}} A$.

Notice that A can be any formula. Some philosophers and logicians reject Ex Falso Quodlibet, since it is meaningless to conclude whatever A, which could be irrelevant with respect to any elements in Γ or Δ. Deductive systems that do not enjoy Ex Falso Quodlibet are called *paraconsistent* (Priest 2006). As expected, many relevance logics are also paraconsistent (Anderson and Belnap 1975).

Denying Double-negation Elimination and Ex Falso Quodlibet leads to non-classical systems, but we are still in the domain of deductive reasoning, even though we may argue that some properties of it, such as topic-neutrality, are significantly challenged.

2.1.1 Reasoning Styles for Deduction

Since a deductive system consists of a formal language to build formulae and of some reasoning principles to prove formulae from formulae, one can intuitively think of $\mathscr{D}$ as associated with a (possibly infinite) set of formulae, which are thus closed under deduction, i.e. under the reasoning principles of $\mathscr{D}$. Given the same formal language, it is quite possible that more deductive systems equivalently characterise the same set of formulae—i.e. that, for any sets of formulae, different systems prove exactly the same sets of conclusions: if so, those systems are different ways to express the same *logic*, i.e. the same set of formulae. Accordingly, different, but equivalent systems express diverse reasoning styles for one logic. Put it differently, we can say that the same general reasoning principles can be formulated into different ways, thus leading to different (but equivalent) deductive methods that allow for deriving the same set of formulae.

Sundholm (1983) argues that there are basically three main types of deductive systems: Hilbert-style systems, natural deduction systems and sequent systems. Let us consider here the first type of systems, which are also known as the axiomatic ones (cf. van Benthem et al. 2001, Chap. 2):

Definition 6 (*Axiomatic systems*) An axiomatic system consists of a set of *axioms*, which are schemata that are indisputable and independent one from another, and a set of *inference rules*, i.e. ways for drawing conclusions.

A *proof* is a finite sequence of formulae, where each formula is either an axiom, or follows from previous formulae in the proof by an inference rule.

A formula is a *theorem* if it occurs in a proof, typically as the last formula in the sequence.

A set of axioms and rules defines an axiomatisation for a given logic **L**.

Classical propositional logic **CPL** is a logic that reads in a certain way the connectives mentioned in Example 1.[6] A possible axiomatisation for **CPL** is the following (Mendelson 1987):

[6] See the next section for more details.

- *Axiom schemata*:

$$\begin{array}{ll}
\textbf{A1} & A \rightarrow (B \rightarrow A) \\
\textbf{A2} & ((A \rightarrow (B \rightarrow C)) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C)) \\
\textbf{A3} & (\neg B \rightarrow \neg A) \rightarrow ((\neg B \rightarrow A) \rightarrow B)
\end{array}$$

- *Inference rules*:

$$\frac{A \rightarrow B \qquad A}{B} \qquad (\textbf{Modus Ponens})$$

One may contend that other systems, like natural deduction, are much easier to understand, at least with **CPL**:

> The formalisation of logical deduction especially as it has been developed by Frege, Russell, and Hilbert, is rather far removed from the forms of deduction used in practice in mathematical proofs [...] In contrast, I intended first to set up a formal system which comes as close as possible to actual reasoning. The result was a "calculus of natural deduction." (Gentzen 1934, p. 68)

Nevertheless, we stick in this chapter to axiomatics, because it offers a compact and intuitive representation of deontic reasoning.

2.2 Deductive Reasoning—Logical Consequence as Entailment

Despite the interesting intuitions behind a syntactic characterisation of deduction, Patricia Blanchette (2001) has recalled that "the logical relations do not obtain between bare syntactic items, but only between items which make some determinate claim on the world; which are, in brief, meaningful."[7] Minimally, the meaning of any formula in any logic corresponds to its truth conditions, which requires to have a *semantics* for that logic. A semantics is in turn a set of *models* which assign an *interpretation* for the formulae, i.e. a truth value to them.

Truth preservation

If we approach deduction semantically, we are able to focus on another very important property of it, i.e. the fact that deduction is *truth-preserving*: an argument is deductively valid if there is no interpretation (in the semantics) in which its premisses are all true and its conclusion false, something that we have already mentioned in Definition 1.

It was again Alfred Tarski (1936) among the first ones who argued that a *model-theoretic* definition of consequence relation is fundamental. This intuition can be now formalised as follows:

[7] But compare the interesting counterarguments offered by Alchourron and Martino (1990).

Definition 7 (*Logical consequence as entailment*) Any well-formed formula C is a *logical consequence* in $\mathscr{D}$ of a set Γ of well-formed formulae, i.e. $\Gamma \models_{\mathscr{D}}$, iff every model for $\mathscr{D}$ that makes all elements of Γ *true* also makes C *true*.

Formal semantics vary from logic to logic, since they are meant to provide an interpretation for expressions and operators such as the logical connectives in Example 1. The following definition provides the standard semantics for Classical Propositional Logic (**CPL**):

Definition 8 (*Semantics for* **CPL**) Let $\mathscr{L}$ be the language of **CPL** (as defined in Example 1).

A model for $\mathscr{L}$ is an assignment v of exactly one truth value (true **T** , or false **F**) to every letter of $\mathscr{L}$. The truth conditions for the logical connectives are as follows:

- $v(\neg A) = \mathbf{T}$ iff $v(A) = \mathbf{F}$;
- $v(A \wedge B) = \mathbf{T}$ iff $v(A) = \mathbf{T}$ and $v(B) = \mathbf{T}$;
- $v(A \vee B) = \mathbf{T}$ iff at least A (or B) is such that $v(A) = \mathbf{T}$ (or $v(B) = \mathbf{T}$);
- $v(A \rightarrow B) = \mathbf{T}$ iff at either $v(A) = \mathbf{F}$ or $v(B) = \mathbf{T}$;
- $v(A \equiv B) = \mathbf{T}$ iff either both $v(A) = \mathbf{T}$ and $v(B) = \mathbf{T}$, or both $v(A) = \mathbf{F}$ and $v(B) = \mathbf{F}$.

A final, but fundamental question is whether proof-theoretic and model-theoretic definitions of deduction are equivalent. While this result is very much desirable, it is not in general guaranteed, and it requires to prove two separate, important results:

Theorem 1 *(*Soundness of $\mathscr{D}$*) For every set Γ of formulae and every formula C of $\mathscr{D}$, if $\Gamma \vdash_{\mathscr{D}} C$, then $\Gamma \models_{\mathscr{D}} C$.*

Theorem 2 *(*Completeness of $\mathscr{D}$*) For every set Γ of formulae and every formula C of $\mathscr{D}$, if $\Gamma \models_{\mathscr{D}} C$, then $\Gamma \vdash_{\mathscr{D}} C$.*

Soundness results guarantee that we can never derive false conclusions from true premisses, using the reasoning principles of $\mathscr{D}$. Failing to prove soundness means that $\mathscr{D}$ is wrong and of no use. Completeness is desirable as well, but it is not always guaranteed, nor easy to prove: it says that, whenever there is a model for Γ and C, then there a proof for C from Γ.

3 Deontic Reasoning

3.1 Introduction

In the last 60 years, there has been a large amount of research on formal models of normative concepts. This is an interdisciplinary domain, where logicians (like von Wright 1951), legal theorists (like Alchourrón 1969) and computer scientists (like McCarty 1986) have merged their efforts.

Any logical investigation of the main normative legal concepts requires a formal account of deontic notions, such as obligation (duty) and permission. These ideas have been characterised by using different logical tools, most frequently related to the possible-worlds semantics of modal logic (for an overview, see Åqvist 2001). In this section[8], we shall attempt to provide a gentle introduction to basic modal deontic logic. In particular, we shall present the so-called Standard Deontic Logic, which for long time has been considered a reference for deontic logicians. We will discuss standard semantics for deontic logics. We will then show how to extend obligations and permissions in order to capture the idea of directed deontic modalities. The most comprehensive introduction to deontic reasoning can be found in (Gabbay et al. 2013).

3.2 Obligations and Permissions: Basics

We can represent obligations through formulae having the following structure:

$$\begin{array}{l} \boldsymbol{Obl}\ A \\ \text{(it is obligatory that) } A \end{array} \tag{6}$$

where A denotes any action or state of affairs, and ***Obl*** is the deontic operator for obligation, to be read as "it is obligatory that."

For instance, the following formula states that John has the obligation to take security measures (to protect the personal data that John is processing):

$$\boldsymbol{Obl}\ \text{[John takes security measures.]} \tag{7}$$

As intuitively expected, when one (e.g. John) is obliged not to perform a certain action we can say that one is forbidden from doing that action. Here is an example:

$$\begin{array}{l} \boldsymbol{Forb}\ \text{[John downloads copyrighted work]} \\ \text{(it is forbidden that John downloads copyrighted work)} \end{array} \tag{8}$$

Finally, we need to express permissions which we do through the operator ***Perm***. For example, to indicate that Tony is permitted to run the program MySoft we write:

$$\begin{array}{l} \boldsymbol{Perm}\ \text{[Tony runs MySoft]} \\ \text{(it is permitted that Tony runs MySoft)} \end{array} \tag{9}$$

Usually, deontic logicians define the concept of permission as the "dual" of the one of obligation.[9] Indeed, it seems intuitively obvious that

[8] Sections 3.2 and 3.3 freely elaborate on parts of (Sartor 2006).

[9] This is somehow a simplification. More on permissions in Section 4.1.1.

Table 1 Complete deontic qualifications

Country	Wearing the veil (V)	Not wearing the veil ($\neg V$)
France	$\boldsymbol{Forb}\ V$	$\boldsymbol{Obl}\ \neg V$
Iran	$\boldsymbol{Obl}\ V$	$\boldsymbol{Forb}\ \neg V$
UK	$\boldsymbol{Perm}\ V$	$\boldsymbol{Perm}\ \neg V$

$$\begin{array}{l}\boldsymbol{Forb}\ A \equiv \boldsymbol{Obl}\ \neg A \\ \text{(Being prohibited to perform an action means being obliged not to do it)}\end{array} \tag{10}$$

For instance, that Tom is forbidden from smoking means that he is obliged not to smoke. Hence, since it is likewise intuitive that some A is permitted means that A is not forbidden

$$\boldsymbol{Perm}\ A \equiv \neg \boldsymbol{Forb}\ A \tag{11}$$

then

$$\boldsymbol{Perm}\ A \equiv \neg \boldsymbol{Obl}\ \neg A. \tag{12}$$

3.3 Facultativeness

The deontic qualifications "obligatory" and "forbidden" are complete, in the sense that they determine the deontic status of both the action or state of affairs A they are concerned with and the complement of A: to say that action or state of affairs A is obligatory is equivalent to saying that $\neg A$ is forbidden, and to say that A is forbidden is equivalent to saying that $\neg A$ is obligatory. For instance, to say that it is obligatory to wear a tie means that it is forbidden not to wear it, and to say that it is forbidden to smoke means that it is obligatory not to smoke.

On the contrary, when we only know that an action or state of affairs is permitted, we do not know the status of its complement. In particular, when a positive action or state of affairs is permitted (namely the action is not forbidden), then its negation can be likewise permitted or it can be forbidden (this will be the case when the action, besides being permitted, is also obligatory).

Consider for example,[10] a girl wearing a veil at school, which we abbreviate as V (V = "the girl wears a veil"):

Let us assume that V is permitted in the UK, obligatory in Iran, and forbidden in France. Consider now the omission of the veil, namely $\neg V$. $\neg V$ is permitted as well in the UK, but it is forbidden in Iran, and is obligatory in France.

From Table 1 it appears that to express the normative qualification of wearing a veil by a girl at school in Iran or in France, it is sufficient to say that in Iran wearing the veil is obligatory while in France it is forbidden. In fact, the deontic propositions

[10] The example is due to Sartor 2005, p. 453.

⌈***Obl*** V in Iran⌉ and ⌈***Forb*** V in France⌉ entail respectively ⌈***Forb*** $\neg V$ in Iran⌉ and ⌈***Obl*** $\neg V$ in France⌉.

On the contrary, saying that V is permitted (namely not prohibited) in the UK is not sufficient to fully specify V's normative status in that country: the permission to wear a veil (***Perm*** V) is consistent both with the permission not to wear it (***Perm*** $\neg V$) and with the prohibition not to wear it (with ***Forb*** $\neg V$), that is, with the obligation to wear it (with ***Obl*** V).

Thus, to provide a complete deontic specification, we need to specify whether not wearing the veil is forbidden or permitted. In the UK, wearing a veil is permitted (as in Iran, and contrary to what is the case in France), but not wearing the veil is permitted too (as in France, and contrary to what is the case in Iran).

In conclusion, besides an action or state of affairs being obligatory (and its negation being forbidden), and besides its being forbidden (and the negation being obligatory), there is a third way for the deontic status of an action or state of affairs to be fully specified: this consists in both the action or state of affairs being permitted and its negation being permitted.

In common parlance, when one says "permitted," one usually refers to this third option.[11] We prefer to use the specific term *facultative*—abbreviated as ***Facult***—to express this notion.

Definition 9 (Facultative) Any A is *facultative*[12] when both A and $\neg A$ are permitted:

$$\textbf{\textit{Facult}}\ A \equiv (\textbf{\textit{Perm}}\ A \wedge \textbf{\textit{Perm}}\ \neg A) \tag{13}$$

(⌈it is facultative that A⌉ is equivalent to ⌈it is permitted that A and it is permitted that $\neg A$⌉)

For example, saying that ⌈in the UK, for a girl going to school, it is facultative to wear the veil⌉ amounts to saying that ⌈she is permitted both to wear the veil and to not wear it⌉. Note that something being facultative does not entail that others are forbidden to prevent it (or that other are forbidden to prevent its negation). In this sense, facultativeness is a weak notion of freedom. This is because in general, we need to distinguish the permission that one agent j does some A from the prohibition that another (or all others) prevents j from doing this A: it is possible (and very common indeed) that one is permitted to do actions that others are permitted to prevent.

This distinction is significant since even mere permissions (i.e. permissions which are not coupled with a prohibition to prevent the permitted action) are not useless: the very fact that an action is permitted is often sufficient to ensure a possibility of performing that action. This is because there are general prohibitions upon others that—by limiting in general their action—proscribe certain ways of interfering with the holder of a mere permission, and thus indirectly provide a certain legal protection for the possibility of doing the permitted action. As (Hart 1982, 171) puts it:

[11] If one knew that the action or state of affairs was not only permitted but obligatory, one would use the latter qualification, according to the Gricean principle of quantity, which requires that we provide all the relevant information we have (see Grice 1989).

[12] In deontic logic, when it is said that any A is facultative, this is equivalent to saying that A is indifferent.

> at least the cruder forms of interference, such as those involving physical assault or trespass, will be criminal or civil offences or both, and the duties or obligations not to engage in such modes of interference constitute a protective perimeter behind which liberties exist and may be exercised.

For instance, I have the faculty of smoking inside my house (I am neither obliged nor forbidden to smoke there). Such a faculty is indirectly protected by various legal provisions, like those entailing on the one hand the prohibition of assaulting me and on the other hand the obligation to respect my property right over my cigarettes. Notwithstanding such protection, others are permitted to use many means in order to prevent me from smoking inside my house. For instance, they may buy all cigarette boxes available at the tobacco shop and destroy them, they may refuse to lend me their lighter, they may threaten to leave the room if I smoke, and so on.

However, there are also cases where one's permission (and, in particular, one's faculty, in the sense above specified, namely as a permission to do and to not do) is coupled with another's prohibition from preventing exactly the permitted action, or at least with the prohibition from preventing it on purpose. Alexy (1985, 208ff) refers to such faculties by speaking of *directly protected freedoms*, as opposed to *unprotected freedoms*, which are not accompanied by the prohibition on interference. Among such directly protected freedoms are the negative liberties one has towards the State in liberal countries (for instance, freedom of speech, of religion, and so on). Thus we would express a protected freedom, with regard to action A as:

$$\boldsymbol{Facult}\ A = \boldsymbol{Forb}\ [\text{prevent } A] \wedge \boldsymbol{Forb}\,[\text{prevent } \neg A] \tag{14}$$

When a freedom is unprotected, the possibility of exercising that freedom (of performing the facultative action) depends upon the arbitrary choice of others, who could, if they wished, interfere, even if, as a matter of fact, they kindly abstain from doing so. On the contrary, when a liberty is protected, arbitrary inferences are prohibited (this may be linked to the republican idea of liberty as freedom from arbitrary interference; see Pettit 1997).

An even stronger notion of one's liberty to do A is obtained when the others' prohibition from interfering with A is coupled with the obligation (upon others or upon the government) to provide some means for performing A, namely the obligation to ensure that the concerned agent has the effective capability of doing what he or she is permitted to do (in this way, the so-called negative freedom becomes as well a positive or substantive freedom; see Sen 1999).

One fact should be clear in concluding this section: to comprehensively capture the idea of facultativeness we have to enrich the basic language of deontic logic we have presented so far. In particular, we would need a formalism able to represent, and reason about actions. Indeed,

- in many cases the permission that one agent j does some A does not necessarily imply that there is a prohibition for another agent to prevent j from doing this A;
- consider formula (14): even this case requires to refer to agents' actions, which is something that in the formula is captured by introducing 'prevent'.

How to extend deontic logic with a logic of agency is out of the scope of this Chapter (see Sergot 2013).

3.4 Deontic Logic: Axiomatics and Semantics

In the previous sections, we have informally introduced the concepts of obligation, prohibition, permission and facultative. However, we still have to formally account for them: so far we only have some uninterpreted symbols denoting deontic notions and we know how to reduce all of them, for instance, into ***Obl*** by using the logical negation ¬. A comprehensive and rigorous approach to deontic logic can consist in identifying some axiom schemata that are suppose to hold for ***Obl*** and show how they can be characterised in a semantic setting.

3.4.1 Some Deontic Schemata

A deontic logic can be easily obtained by extending the language of Classical Propositional Logic with one deontic operator ***Obl***: in fact, recall that ***Forb*** is equivalent to ***Obl*** ¬ and ***Perm*** is equivalent to ¬***Obl*** ¬. Thus, we would strictly need to use just one operator. However, we will use all operators since, if this will keep things conceptually clear.

With this said, we can build an axiomatic deontic system, namely a logical system that, first of all, takes as valid some axiom schemata that state some non-trivial properties of ***Obl*** and ***Perm***.

Let us consider the following schemata:

$$\boldsymbol{Obl}\,(A \rightarrow B) \rightarrow (\boldsymbol{Obl}\,A \rightarrow \boldsymbol{Obl}\,B) \tag{15}$$

$$\boldsymbol{Obl}\ A \rightarrow \boldsymbol{Perm}\ A \tag{16}$$

$$\boldsymbol{Obl}\,A \rightarrow A \tag{17}$$

Axiom schema (15) looks intuitively acceptable: if it is obligatory that, if you buy a car, then you pay, then, if it is obligatory that you buy a car, then it is obligatory to pay.

Axiom schema (16) seems to be reasonable: indeed, if it is obligatory to compensate damages, then it is permitted to do so.

The schema (17) is on the contrary problematic: can we say that the fact that it is obligatory to compensate damages necessarily implies that we do it? If this schema holds, then we would assume that all obligations are fulfilled.

Hence, it seems that schemata such as (15) and (16) can be adopted, while a (17) should be abandoned.

3.4.2 A Semantic Reading of Deontic Operators

What do we mean when we say that ***Obl*** A is true? Several philosophers argued that this question is meaningless because norms and obligations are not susceptible of any truth evaluation. However, let us ignore here this objection (which is debatable, too) and follow a major part of the literature on deontic reasoning, trying to check when sentences like ***Obl*** A are true. For instance, suppose that it is obligatory to pay taxes. If this is true, this means that in all (e.g. legally) ideal situations we in fact pay taxes.

How can we formally express this intuition? We can use the concept of *possible world*: any possible world is a sort of description of how things are in the current situation (the actual world) or how they could be (worlds alternative to the actual one). Worlds can thus be analysed in terms of possible truth assignments to all the atomic propositional letters describing how things are in a given situation: in our world legal philosophers are smart, while we can conceive an alternative situations where they are not smart at all.

Notice that not all worlds are legally or morally ideal: we can imagine situations where many individuals massively commit atrocities. However, we can isolate a subset of possible worlds that are inherently good, where we always pay taxes, compensate damages, and do not commit any atrocity. At this point, it should clear what we mean by saying that ***Obl*** A is true: it means that A is true in all (legally, morally, etc.) ideal worlds.

A formal method for checking the truth of obligations thus needs a way to identify the set of ideal worlds with respect to each other world w: such a set includes those worlds that are ideal from the perspective of w. This can be done using the following formal structures from modal logic (called Kripke frames and Kripke models), (see Kripke 1959, 1963; Blackburn et al. 2001):

Definition 10 (*Kripke frames and models*) A *Kripke frame* $\mathscr{F}$ is a structure

$$\langle W, R \rangle$$

where

- W is the set of all possible worlds;
- R is a binary relation over W that determines the ideal worlds in W for each world in W.

A *Kripke model* $\mathscr{M}$ based on the frame $\mathscr{F}$ is a structure

$$\langle W, R, v \rangle$$

where

- $\mathscr{F} = \langle W, R \rangle$;
- v assigns the truth values **T** or **F** to any atomic sentence in any given world (i.e. it states what atomic sentences are true or false in each world).

First, we should note that logical sentences are now evaluated with respect to worlds: a formula A (e.g. "we pay taxes") can be true at world w and false at world v. Second, since the relation R selects for each world w the those worlds that are ideal with respect to w, it works as follows. Suppose $W = \{w_1, w_2\}$, then R selects some pairs of worlds in the following set

$$\{\langle w_1, w_1 \rangle, \langle w_1, w_2 \rangle, \langle w_2, w_2 \rangle, \langle w_2, w_1 \rangle\}.$$

For example,

$$R = \{\langle w_1, w_2 \rangle\}.$$

where such pairs say what worlds are ideal with respect to which world: specifically, we know that w_2 is ideal with respect to w_1, whereas there no ideal world for w_2.

Formulae of classical propositional logic are evaluated as usual, but with respect to worlds. As usual in the literature, given any Kripke model $\mathscr{M}$ and any world w in it, we write $\models^w_{\mathscr{M}} A$ to say that A is true at w in $\mathscr{M}$. Hence, if we consider for instance the propositional connectives $\neg$ and $\rightarrow$, the procedure to check the truth value of formulae is as follows: given any Kripke model $\mathscr{M}$ and any world w in it

- if A is an atomic formula, $\models^w_{\mathscr{M}} A$ iff $v(A, w) = \mathbf{T}$;
- if $A = \neg B$ then $\models^w_{\mathscr{M}} A$ iff $\not\models^w_{\mathscr{M}} B$;
- if $A = B \rightarrow C$ then $\models^w_{\mathscr{M}} A$ iff $\not\models^w_{\mathscr{M}} B$ or $\mathscr{M} \models^w_{\mathscr{M}} C$.

What we have informally said before on the semantic meaning of ***Obl*** should make now clear how to evaluate any sentence of the form ***Obl*** A: given any Kripke model $\mathscr{M}$ and any world w in it

- $\models^w_{\mathscr{M}}$ ***Obl*** A iff, for each world w', if w' is ideal with respect to w according to R, then $\models^{w'}_{\mathscr{M}} \models^w_{\mathscr{M}} A$.

How can we evaluate permissions such as ***Perm*** A? Since ***Perm*** $= \neg$***Obl*** $\neg$, then ***Perm*** A states nothing but the following:

- $\models^w_{\mathscr{M}}$ ***Perm*** A iff there exists at least one world w' such that w' is ideal with respect to w according to R and $\mathscr{M} \models^{w'}_{\mathscr{M}} A$.

If it is permitted to pay taxes, then there should be at least one ideal world where we pay taxes.

As usual in any modal logic, notice that this semantics requires to define different perspectives where a formula can be evaluated:

Definition 11 A formula A is *true* in a world w of a model $\mathscr{M}$ iff $\models^w_{\mathscr{M}} A$. A formula A is *true* in a model $\mathscr{M}$ $\models_{\mathscr{M}} A$, iff for all w, $\models^w_{\mathscr{M}} A$. A formula A is *valid* on a frame $\mathscr{F} = \langle W, R \rangle$ iff for any model $\mathscr{M} = \langle W, R, v \rangle$ based on $\mathscr{F}$ we have $\models_{\mathscr{M}} A$.

Given a class of frames X, a formula A is X-*valid*, $\models_X A$, iff for any frame $\mathscr{F} \in X$, $\models_{\mathscr{F}} A$.

Let us conclude this section with some final semantic remarks on the axiom schemata (15), (16) and (17).

Schema (15) is always valid, namely there is no way to falsify it in the semantics we have discussed above (i.e. there is no model based on any Kripke frame that makes (15) false).

The other schemata instead are not necessarily valid, but need some additional constraints on the relation R.

For example, consider (17), i.e. $\textbf{\textit{Obl}}\ A \rightarrow A$, which means that if A is obligatory at any world w, then A is true at w. This schema assumes that the relation R be reflexive, namely that each world is ideal with respect to itself. Technically, this means that (17) is valid in the class of Kripke frames where R is reflexive. To understand why, try to falsify (17) at the world w in the structure of Fig. 1.

If (17) is false at w, then its antecedent $\textbf{\textit{Obl}}\ A$ must be true at w while the consequent A must be false there. We have two ideal worlds for w: the world z and w itself. Thus, if $\textbf{\textit{Obl}}\ A$ is true in w, A must be true in w and z, which is the case. However, if (17) is false, then A must be false in w, which is not the case. To successfully falsify the schema, we should exclude that w is ideal with respect to itself and state that A is false in w, thus obtaining the structure of Fig. 2.

Since the axiom schema (17) should not be valid in deontic logic, then we should exclude that the relation is reflexive.

Consider now the axiom schema (16), $\textbf{\textit{Obl}}\ A \rightarrow \textbf{\textit{Perm}}\ A$, and try to falsify it. As we have said, $\textbf{\textit{Perm}}\ A$ is true at any world w if there is at least one ideal world where A is true. So, to falsify (16) in any world w we should assume that $\textbf{\textit{Obl}}\ A$ is true at w and $\textbf{\textit{Perm}}\ A$ false there. This is guaranteed in the structure depicted in Fig. 3: no world is ideal with respect to w, which vacuously makes $\textbf{\textit{Obl}}\ A$ true in w (remember that the truth of $\textbf{\textit{Obl}}\ A$ requires that IF any w' is ideal with respect to w according to R, THEN $\models_{\mathscr{M}}^{w'} A$), while $\textbf{\textit{Perm}}\ A$ requires that w should have at least one ideal world, which is not the case.

So we should assume that for every world there should exist at least one ideal world, otherwise we cannot guarantee that all obligations imply the corresponding permissions: in other words, the schema (16) is valid in the class of Kripke frames where, for each world w, there is at least a world which is ideal with respect to w. This property on Kripke frames is called *seriality*.

Fig. 1 Model 1

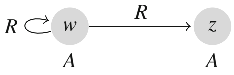


Fig. 2 Model 2

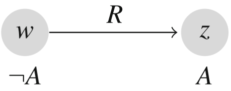


Fig. 3 Model 3

3.4.3 Standard Deontic Logic

The first modern deontic system has been introduced by von Wright (1951). Some later revisions of von Wright (1951)'s system transformed it into what is usually called "the standard system of deontic logic" (**SDL**) (Føllesdal and Hilpinen 1971; Hansson 1971), which is certainly the most cited and investigated system of deontic logic (Chellas 1980). **SDL** is the deontic system consisting any axiomatisation of Classical Propositional Logic, the axiom schema (15)

$$\boldsymbol{Obl}\,(A \rightarrow B) \rightarrow (\boldsymbol{Obl}\,A \rightarrow \boldsymbol{Obl}\,B)$$

the axiom schema (16)

$$\boldsymbol{Obl}\,A \rightarrow \boldsymbol{Perm}\,A$$

and the following inference rule

$$\frac{\vdash A}{\vdash \boldsymbol{Obl}\,A} \tag{18}$$

This system is equivalent, for example, to accepting (16) and the following inference rule (Chellas 1980):

$$\frac{\vdash A_1 \wedge \cdots \wedge A_n \rightarrow B}{\vdash \boldsymbol{Obl}\,A_1 \wedge \cdots \wedge \boldsymbol{Obl}\,A_n \rightarrow \boldsymbol{Obl}\,B} \tag{19}$$

Obviously, (19) implies (when $n = 1$)

$$\frac{\vdash A \rightarrow B}{\vdash \boldsymbol{Obl}\,A \rightarrow \boldsymbol{Obl}\,B} \tag{20}$$

Similarly, (19) also implies (18) (when $n = 0$).

An intuitive instance, for example, of (20) is the following:

$$\frac{\vdash (\text{bombing} \wedge \text{killing}) \rightarrow (\text{killing})}{\vdash \boldsymbol{Obl}\,(\text{bombing} \wedge \text{killing}) \rightarrow (\boldsymbol{Obl}\,\text{killing})} \tag{21}$$

Clearly, since $A \equiv B$ is nothing but $(A \rightarrow B) \wedge (B \rightarrow A)$, then, we also have

$$\frac{\vdash A \equiv B}{\vdash \boldsymbol{Obl}\,A \equiv \boldsymbol{Obl}\,B} \tag{22}$$

which can be intuitively instantiated into the following example:

$$\frac{\vdash (\text{murdering} \rightarrow \text{killing}) \equiv (\neg\text{murdering} \vee \text{killing})}{\vdash \textbf{\textit{Obl}}\,(\text{murdering} \rightarrow \text{killing})) \equiv (\textbf{\textit{Obl}}\,(\neg\text{murdering} \vee \text{killing})} \tag{23}$$

An intuitive derived inference rule in **SDL** is also the following:

$$\frac{\textbf{\textit{Obl}}\,(A \rightarrow B) \qquad \textbf{\textit{Obl}}\,A}{\textbf{\textit{Obl}}\,B} \tag{24}$$

An example of it can be

$$\frac{\textbf{\textit{Obl}}\,(\text{buy} \rightarrow \text{pay}) \qquad \textbf{\textit{Obl}}\,\text{buy}}{\textbf{\textit{Obl}}\,\text{pay}} \tag{25}$$

Finally, since Modus Ponens is imported from propositional logic, we have

$$\frac{A \rightarrow \textbf{\textit{Obl}}\,B \qquad A}{\textbf{\textit{Obl}}\,B} \tag{26}$$

An example of it can be

$$\frac{\text{murder} \rightarrow \textbf{\textit{Obl}}\,\text{punishment} \qquad \text{murder}}{\textbf{\textit{Obl}}\,\text{punishment}} \tag{27}$$

One important result regarding **SDL** is that proof-theoretic and model-theoretic definitions of deduction are equivalent:

Theorem 3 (SDL: Soundness and Completeness) *Let $\mathscr{S}$ the class of serial Kripke frames. For every set Γ of formulae and every formula C of* **SDL***, $\Gamma \models_{\mathscr{S}} C$ iff $\Gamma \vdash_{\mathbf{SDL}} C$.*

3.5 Directed Obligations

More articulate normative notions and, in particular, the idea of a right, cannot be built on the basis of obligations and permissions alone. Such notions embed a *teleological* perspective, namely a focus on purposes or interests (final or intermediate values, ends, objectives) which a normative proposition[13] is meant to serve: only when a such a proposition is concerned with the interests of certain individuals, can we view it as conferring rights upon these individuals.

The purpose of a normative proposition should not be mistaken for the aim (possibly a self-interested one, or also an illegal one) that is pursued by the individual members of the legislative body when voting for that proposition.[14] To establish what

[13] We use the expression *normative proposition* to mean any possible legal content: a rule, a principle, the connection between a factor and the outcome it favours, and so on.

[14] This issue has been famously addressed by von Jhering (1924), III, 35, who distinguished the purpose of a duty (the interest it is intended to serve, according to the point of view of the legal

interests are served by a proposition of law, besides considering the (legitimate) objectives of the historical legislator, we often need to engage in an exercise in rationalisation, aimed at establishing which ones, among the effects of the adoption of that proposition (i.e. among the consequences ensuing from its use as a standard for acting and adjudicating), may represent valuable reasons for its communal adoption and its persistent endorsement.

Often the purpose for (the adoption of) a normative proposition is to protect the interests of certain individuals, though these interests may take different contents (within certain ranges), according to the choices of the individuals concerned. The individuals whose interest is protected or promoted by a normative proposition are called beneficiaries. In the simples case, where normative propositions just correspond to obligations, we may need to determine the beneficiaries of certain obligations.

These considerations lead us to introduce the notion of an *directed obligation*.[15]

Definition 12 (*Directed obligation*) It is *obligatory, towards k*, that A

$$\boldsymbol{Obl}^{k} A$$

iff $\boldsymbol{Obl}\ A$ advances the interest of k.

For example, the notion enables us to express normative propositions like the one stating that it is obligatory, towards Mary, that Tony pays $1,000 to Mary.[16]

By denying directed prohibitions, we get other-directed permissions. Thus the directed permission, towards k, A is the case only means that it is not obligatory, towards k, that $\neg A$ holds: there is no such obligation for the benefit of k (though A may be obligatory towards other people). For example, suppose that Mary and her neighbour Tom make an agreement according to which she is permitted to erect a building up to 15 m high. We may then say that in this case it is permitted, towards Tom, that Mary erects a building up to 15 m high. The fact that Mary is permitted towards Tom to erect the building does not exclude that she still is forbidden towards Ann, another of Mary's neighbours, who has not consented to the construction.

It is often assumed that $\boldsymbol{Obl}$ and $\boldsymbol{Obl}^{k}$ (for any agent k) obey the same basic logical principle. If so, for any given set Ag of agents, we can simply reframe the axiom schemata and inference rules of **SDL** for $\boldsymbol{Obl}$:

system, or of the legal community) from its various side effects (reflex-effects, *Reflexwirkungen*). Jhering considers, for instance, the case of a law prohibiting the import of certain goods, which was enacted by politicians having the aim of favouring a particular domestic producer (who had lobbied for this result). He argues that the fact the individual lawmakers had this aim in mind does not imply that the law confers a right on that manufacturer: from a legal perspective the manufacturer's advantage is rather to be viewed as a side effect of that law.

[15] On the idea of a directed obligation, see the seminal contribution by Krogh and Herrestad (1996), though their formalisation does not fully coincide with the one here described. More details in Sartor 2005; Sergot 2013.

[16] Notice that, as we have argued at the end of Sect. 3.3 in regard to facultativeness, also here a better way for representing directed obligations would need to extend the language with actions.

$$\boldsymbol{Obl}^k(A \to B) \to (\boldsymbol{Obl}^k A \to \boldsymbol{Obl}^k B) \qquad (k \in Ag) \tag{28}$$

$$\boldsymbol{Obl}^k A \to \boldsymbol{Perm}^k A \qquad (k \in Ag) \tag{29}$$

$$\frac{\vdash A}{\vdash \boldsymbol{Obl}^k A} \tag{30}$$

On the contrary, it is not immediately obvious how to semantically render directed obligations in Kripke semantics. The simplest solution is the one proposed for any multi-modal logics (Kurucz et al. 2003): indeed, given two expressions $\boldsymbol{Obl}^k A$ and $\boldsymbol{Obl}^j A$ we could say that A is obligatory according to two different deontic operators. This idea can be easily implemented by extending Kripke structures as follows:

$$\langle W, \{R_k\}_{\forall k \in Ag}, v\rangle$$

where

- W is the set of all possible worlds;
- $\{R_k\}_{\forall k \in Ag}$ is a set of binary relations over W: each of them uniquely corresponds to an agent k in Ag and it determines the ideal worlds for k in W for each world in W;
- v as usual assigns the truth values **T** or **F** to any sentence in a given world.

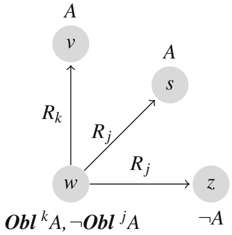
Hence, the evaluation clause for any expression like $\boldsymbol{Obl}^k A$ is as follows: given any model $\mathscr{M}$ and any world w in it

- $\models^w_{\mathscr{M}} \boldsymbol{Obl}^k A$ iff, for each world w', if w' is ideal with respect to w according to R_k, then $\models^{w'}_{\mathscr{M}} A$.

As we have said, it may be of course the case that A is obligatory in the interest of an agent k but not of another agent j. In other words, we may have, for example, that $\boldsymbol{Obl}^k A$ is true in some world w of a model $\mathscr{M}$ while $\boldsymbol{Obl}^j A$ is false in it. In other words

- A is true in all ideal worlds with respect to w for the agent k;
- there is a world z, which is ideal for j with respect to w, where A is false.

For example, consider the following example:



Here, given the world w, all ideals worlds (the world v) for k make A true, while there is an ideal world for j (world z) where A is false, thus making false in w the formula $\boldsymbol{Obl}^{j} A$.

4 Deontic Reasoning—Some Glimpses Beyond

This section briefly mentions some interesting developments and advanced issues in deontic reasoning.[17]

4.1 Normative Systems

Another influential formal account of deontic notions, complementary to the (modal) logic-based approaches we discussed in the previous section, is the one sparked by Alchourrón and Bulygin (1971). The key feature of this approach is to study norms—viewed as dyadic constructs connecting a fact to a deontic consequence—not as formulae in some logical language, but rather as primitive ordered pairs $\langle condition, consequence\rangle$. A large number of such pairs would constitute an interconnected system called a *normative system*.

Viewed as parts of a bigger system, norms are therefore considered to be uninterpretable if taken in isolation—unlike in logical semantics—and they acquire meaning only by relating to other norms in the system. The focus falls then on the problem of normative reasoning and its most characteristic features, such as: defeasibility, to which Chap. 3, in Part II of this volume, is devoted; the validity of closure principles (e.g. *nullum crimen sine lege*[18]); the problem of handling legal gaps.

The basic idea behind normative systems goes hand in hand with the thesis according to which norms do not bear truth values, and hence that deontic logics do not actually deal with norms, but rather with normative propositions, i.e. statements to the effect that certain norms exist. For instance, in this view, $\boldsymbol{Obl}\ A$ would actually mean something like "there exists a norm commanding A."[19]

In what follows we sketch, very briefly, the basic ideas behind two of the approaches that in recent years have taken up and developed the normative systems approach to the analysis of norms.

[17] This section elaborates on parts of (Grossi and Rotolo 2011).

[18] No crime without law, that is, everything that is not explicitly prohibited should be considered as permitted.

[19] The problem of whether norms bear or not truth values is an old one in philosophy and was put forth in modern times by (Jørgensen 1937). The significance of the problem has recently been reemphasised in (Hansen et al. 2007), and a new approach to the problem emerged from the view of norms as "dynamic" operators—speech acts—modifying ideality orders.

4.1.1 Concepts of Permission

The concept of permission plays an important role in many normative domains in that it may be crucial in characterising notions such as those of authorisation and derogation (Boella and van der Torre 2005; Sartor 2005; Stolpe 2010c). For example, consider when we subscribe to an online sale agreement accepting to enter our personal data on the condition that this information is only used for shipping, and other necessary purposes to communicate with us or deliver the products to us. Here, the permission to use our personal data is an exception to a general prohibition.

Despite this fact, the concept of permission is still elusive in deontic logic and has not been extensively investigated in this field as the notion of obligation. For a long time, deontic logicians mostly viewed permission as the dual of obligation: $\boldsymbol{Perm}\ A \equiv \neg \boldsymbol{Obl}\, \neg A$ (see, e.g. Sect. 3.2). This view is unsatisfactory, as it hardly allows us to grasp the meaning of examples like the one previously mentioned. This is one of the reasons why the attempt to reduce permissions to duals of obligations has been criticised (see Alchourrón and Bulygin 1984; Alchourrón 1993).

One important distinction that has traditionally contributed to a richer account of this concept is the one between *weak* (or *negative*) and *strong* (or *positive*) permission (von Wright 1963). The former corresponds to saying that some A is permitted if A is not provable as mandatory. In other words, something is allowed by a code only when it is not prohibited by that code. At least when dealing with unconditional obligations, the notion of weak permission is trivially equivalent to the dual of obligation (Makinson and van der Torre 2003).

The concept of strong permission is more complicated, as it amounts to saying that some A is permitted by a code iff such a code explicitly states that A is permitted. It follows that a strong permission is not derived from the absence of a prohibition, but is explicitly formulated in a permissive norm. The complexities of this concept depend on the fact that, besides "the items that a code explicitly pronounces to be permitted, there are others that in some sense follow from the explicit ones." The problem is hence "to clarify the inference from one to the other" (Makinson and van der Torre 2003, p. 391–392). For example, if some B logically follows from A, which is strongly permitted, is B strongly permitted as well?

Features such as the distinction between strong and weak permission show the multifaceted nature of permission and permissive norms, which has been overlooked by most logicians for a long time, even though a new interest has emerged in the last few years (Makinson and van der Torre 2003; Boella and van der Torre 2003a, b; Brown 2000; Stolpe 2010c, b; Governatori et al. 2013).

4.1.2 Contrary-to-Duty Reasoning

One of the main research themes in deontic logic is about reasoning with contrary-to-duty (CTD) obligations (Carmo and Jones 2002). These are obligations that are triggered by the violation of other obligations. For example, "you ought not to kill, but if you kill you ought to do it in self-defence." Roughly, contrary-to-duty obligations have to do with sub-ideal, or reparatory obligations. That this notion is impossible to capture in **SDL** was made manifest in the literature by a number of scenarios—often called, with a stretch, paradoxes.

One paradigmatic example is the so-called "gentle murder" paradox:

> Let us suppose a legal system which forbids all kinds of murder, but which considers murdering violently to be a worse crime than murdering gently. [...] The system then captures its views about murder by means of a number of rules, including these two: 1) It is obligatory under the law that Smith not murder Jones. 2) It is obligatory that, if Smith murders Jones, Smith murders Jones gently. (Forrester 1984, p. 194)

The first obligation can be clearly formalised as $\boldsymbol{Obl}\,\neg murder$. For the second, we can have two options:

$$murder \rightarrow \boldsymbol{Obl}\,(murder \wedge gentle) \tag{31}$$

and

$$\boldsymbol{Obl}\,(murder \rightarrow (murder \wedge gentle)). \tag{32}$$

Assume that *murder* is true. By (31) we would conclude that $\boldsymbol{Obl}\,(murder \wedge gentle)$, i.e. that it is obligatory to murder gently in the first place, and hence that $\boldsymbol{Obl}\,murder$ (via the inference rule (20): see Sect. 3.4.3), thereby reaching a contradiction. By (32) we could also reach a contradiction by deriving that $\boldsymbol{Perm}\,(\neg murder \vee (murder \wedge gentle))$ and hence $\boldsymbol{Perm}\,(murder \wedge gentle)$ from which $\boldsymbol{Perm}\,murder$. For a more extensive overview of similarly problematic scenarios for **SDL**, we refer the reader to the aforementioned Hilpinen (2001) and Åqvist (2001).

It is widely acknowledged that the crisis of Standard Deontic Logic is historically and technically related to the formulation of some notorious paradoxes, such as the one above, centring around the regulation of the violation of obligations.

The deontic logic literature on CTD reasoning is immense. However, two fundamental mainstreams have emerged as particularly interesting.

A first line of inquiry is mainly semantic-based. Moving from well-known studies on dyadic obligations, CTD reasoning is interpreted in settings with ideality or preference orderings on possible worlds or states (Hansson 1969). The idea is to substitute the serial ideality relation by a total preorder $\succeq$, i.e. a reflexive, transitive and total binary relation, with the following intuitive reading: $s \succeq s'$ means that state s is at least as good/ideal as s'. Now the most ideal states are the maximal of such an order, and sub-ideality can easily be represented by considering the maximals of some subset of states. On this basis, dyadic obligations of the type "it is obligatory that A under condition B" are interpreted as follows:

$$\models^{w}_{\mathscr{M}} \boldsymbol{Obl}\,(A \mid B) \text{ IFF } Max_{\succeq}(||B||_{\mathscr{M}}) \subseteq ||A||_{\mathscr{M}}$$

where $||.||_{\mathscr{M}}$ denotes the truth-set function of $\mathscr{M} = \langle W, \succeq, v\rangle$ and $Max_{\succeq}$ the function extracting the maximals of a given set. A contrary-to-duty obligation will then be represented by taking condition B to be the violation of some other obligation.

The value of this approach is that the semantic structures involved are quite flexible: depending on the properties of the preference or ideality relation, different deontic logics can be obtained. This semantic approach has been fruitfully renewed in the '90 for example by Prakken and Sergot (1996), van der Torre (1997), and most recently by Hansen (2005) and van Benthem et al. (2013), which have confirmed the vitality of this line of inquiry.

The second mainstream is mostly proof-theoretic. Examples, among others, are various systems springing from Input/Output Logic (Makinson and van der Torre 2000, 2001) (see the next section) and the Gentzen system proposed by Governatori and Rotolo (2006). Both perspectives clearly distinguish in the language and in the logic structures representing norms from those representing obligations, i.e. the consequences generated by norms. These systems follow the slogan "no logic of norms without attention to the normative systems in which they occur" (Makinson, 1999), which draws inspiration from the pioneering works by Stenius (1963) and Alchourrón (1993). While Input/Output approach mainly works by imposing some constraints on the manipulation of conditional norms, Governatori and Rotolo (2006)'s approach is first of all based on the introduction of the new non-classical operator $\otimes$: the reading of an expression like $A \otimes B \otimes C$ is that A is primarily obligatory, but if this obligation is violated, the secondary obligation is B, and, if the secondary (CTD) obligation B is violated as well, then C is obligatory. The intuition behind this construction is that CTD obligations are a special kind of exception. Following the approach by Governatori and Rotolo (2006), let $\vdash$ be a non-classical consequence relation used characterise normative conditionals generating obligations. Hence, an expression like

$$\mathit{Invoice} \vdash \mathit{PayBy7days} \otimes \mathit{Pay5\%Interest} \otimes \mathit{Pay10\%Interest}$$

can be intuitively viewed as a norm meaning the following:

1. if *Invoice* is the case, then *PayBy7days* is obligatory, but,
2. if *PayBy7days* is obligatory and ¬*PayBy7days* is the case, then *Pay5%Interest* is obligatory, but
3. if *Pay5%Interest* is obligatory and ¬*Pay5%Interest* is the case, then *Pay10%Interest* is obligatory.

4.1.3 Input/Output Logic

Input/Output Logics (henceforth IOL) are a formalism introduced in Makinson and van der Torre (2000) that has been applied to the study of normative systems in a long series of papers (e.g. Boella and van der Torre 2004) by viewing them as rule-based process of manipulation of inputs (factual premisses) into outputs (normative conclusions).

The key idea behind the application of IOL to the analysis normative systems consists in representing conditional norms simply as ordered pairs (a, b) where a

represents the antecedent of the rule, and b its consequent: “if a then b” where a has factual content and b normative content, viz. an obligation or a permission. Typically, both a and b are taken to be formulae from propositional logic. Each set of such ordered pairs can be seen as an inferential mechanism which, given an input, determines an output based on those connections.

Various definitions can be given of how to produce the output on the basis of a set of pairs, and all consist in ways of closing the given set of pairs by adding new pairs in accordance to some principles, of which we give two very simple examples:

$$SI : \frac{(a, b)}{(a \land c, b)} \quad CT : \frac{(a, b), (a \land b, c)}{(a, c)} \tag{33}$$

where SI stands for strengthening of the input—essentially an antecedent strengthening property—and CT stands for cumulative transitivity. Formally, given a set $NORM$ of pairs, a closure operation C defined in terms of some of the above principles, and a set of facts A, the output of $NORM$ given C and a set of input formulae I is:

$$out_C(NORM, A) = \{b \mid (a, b) \in C(NORM) \text{ and } s \in A\} \tag{34}$$

Intuitively, $NORM$ represent the norms of a normative system and C the principles according to which the system makes the norms interact with one another. As the reader might have already noticed, this represents a very high-level abstraction of the workings of a normative system. Depending on the (many) ways the output operation is defined, IOL can be used to capture very different principles for reasoning with norms (among which defeasibility). This modelling freedom brought IOL to be applied not so much to the study and analysis of normative reasoning in *actual* normative systems, but rather to the specification of *artificial* normative systems in the field of artificial intelligence (see the aforementioned Boella and van der Torre 2004).

4.1.4 Algebras of Normative Systems

Lindahl and Odelstad (2000) advocate an algebraic analysis of normative systems. The approach is very close in spirit to the one, discussed above, of IOL. However, the formal machinery deployed is not based on logic and hinges on several algebraic and order-theoretical notions. In this section, we provide just a brief sketch of the basic technical ideas underpinning the framework.

According to this approach, norms can be seen—exactly as in IOL—as simple pairs $\langle a, b\rangle$ connecting (factual) conditions to (normative) consequences. Both conditions a and consequences b are taken to be elements of a set X upon which a Boolean algebra $\langle X, \sqcap, -, \bot\rangle$ is defined. Within such a structure, the normative relation between condition a and consequence b is given by extending the preorder yielded by the algebra.[20] The idea is that while the preorder—let us call it $\preceq$—represents some form of logical implication, normative systems add on the top of it the possibility of drawing more conclusions by some form of "legal" implication—let us call it ρ. In other words, each normative system introduces, by stipulation, a consequence relation which is stronger than the logical one: $\preceq \subseteq \rho$. The intuition is that, for instance, the fact that being obliged to pay taxes follows from having a paid job is not a matter of logic, but a matter of stipulation.[21]

Therefore, in Lindahl and Odelstad's view normative systems can be studied as Boolean algebras supplemented by a binary relation ρ. This is, in a nutshell, the key idea behind the approach. Space limitation prevents us to provide more details. It should be mentioned, however, that Lindahl and Odelstad (2000) was followed by a number of papers developing an extensive theory of normative systems on the ground of the simple intuition we have sketched above.[22]

4.2 Normative Dynamics

One peculiar feature of many normative systems, such as the law (Kelsen 1991; Hart 1994), is that it necessarily takes the form of a dynamic normative system. Despite the importance of norm-change mechanisms, the logical investigation of legal dynamics is still underdeveloped. However, recent contributions exist and this section is devoted to a brief sketch of this rapidly evolving literature.

In the Eighties, a promising research effort was devoted by C. E. Alchourrón, P. Gärdenfors and D. Makinson to develop a logical model (AGM) for modelling norm change. As is well-known, the AGM framework distinguishes three types of change operation over theories. Contraction is an operation that removes a specified sentence ϕ from a given theory Γ (a logically closed set of sentences) in such a way as Γ is set aside in favour of another theory Γ^-_ϕ which is a subset of Γ not containing ϕ. Expansion operation adds a given sentence ϕ to Γ so that the resulting theory Γ^+_ϕ is the smallest logically closed set that contains both Γ and ϕ. Revision operation adds ϕ to Γ but it is ensured that the resulting theory Γ^*_ϕ be consistent, i.e. that no contradiction arise (Alchourrón et al. 1985). Alchourrón, Gärdenfors and Makinson argued that, when Γ is a code of legal norms, contraction corresponds to norm derogation (norm removal) and revision to norm amendment.

AGM framework has the advantage of being very abstract but works with theories consisting of simple logical assertions. For this reason, it is perhaps suitable to capture

[20] A preorder can always be associated with a given Boolean algebra in the following way:

$$a \preceq b \text{ IFF } a \sqcap b = a. \tag{35}$$

[21] As is well-known, the idea that legal effects do not follow from norms by logic but, rather, by stipulation was notably defended in legal theory by Kelsen (1991).

[22] An interesting contribution is, for instance, offered by Lindahl and Odelstad (2008).

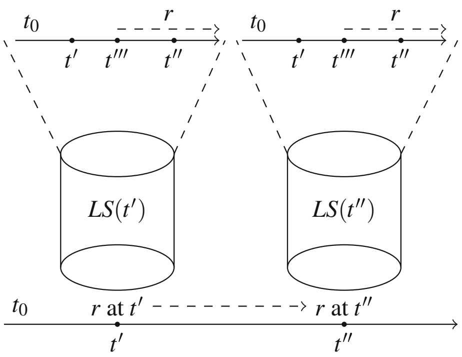


Fig. 4 Legal System at t' and t''

the dynamics of obligations and permissions, not of norms. In fact, it is essential to distinguish norms from obligations and permissions (Boella et al. 2009; Governatori and Rotolo 2010): the latter ones are just possible effects of the application of norms, and their dynamics do not necessarily require to remove or revise norms, but correspond in most cases to instances of the notion of *norm defeasibility* (Governatori and Rotolo 2010).

Some research has been carried out to reframe AGM ideas within rule-based logical systems, which take this distinction into account (cf. Stolpe 2010a; Rotolo 2010). However, also these attempts suffer from some drawbacks especially when applied to the legal domain, as they fail to handle the following aspects of legal norm change:

1. the law usually regulates its own changes by setting specific norms whose peculiar objective is to change the system by stating what and how other existing norms should be modified;
2. since legal modifications are derived from these peculiar norms, they can be in conflict and so are defeasible;
3. legal norms are qualified by temporal properties, such as the time when the norm comes into existence and belongs to the legal system, the time when the norm is in force, the time when the norm produces legal effects, and the time when the normative effects hold.

Hence, normative dynamics can be hardly modelled without considering temporal reasoning. Some works (see, e.g. Governatori and Rotolo 2010) have attempted to address these research issues. All norms are qualified by the above-mentioned different temporal parameters, and the modifying norms are represented as meta-rules, i.e. rules where the conclusions are temporalised rules.

If $t_0, t_1, \ldots, t_j$ are points in time, the dynamics of a legal system LS are captured by a time-series $LS(t_0), LS(t_1), \ldots, LS(t_j)$ of its versions. Each version of LS is called a *norm repository*. The passage from one repository to another is effected by legal modifications or simply by temporal persistence. This model is suitable for modelling complex modifications such as retroactive changes, i.e. changes that affect the legal system with respect to legal effects which were also obtained before the legal change was done. The dynamics of norm change and retroactivity need to introduce another timeline within each version of LS (the timeline placed on top of each repository in Fig. 4). Clearly, retroactivity does not imply that we can really change the past: this is "physically" impossible. Rather, we need to set a mechanism through which we are able to reason on the legal system from the viewpoint of its current version but *as if* it were revised in the past: when we change some $LS(i)$ retroactively, this does not mean that we modify some $LS(k), k < i$, but that we move back from the perspective of $LS(i)$. Hence, we can "travel" to the past along this inner timeline, i.e. from the viewpoint of the current version of LS where we modify norms. Figure 4 shows a case where the legal system LS and its norm r persist from time t' to time t''; however, such a norm r is in force in LS (it can potentially have effects) from time t''' (which is between t' and t'') onwards.

References

Alchourrón, C., P. Gärdenfors, and D. Makinson. 1985. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic* 50: 510–530.

Alchourrón, C.E. 1969. Logic of norms and logic of normative propositions. *Logique et Analyse* 12: 242–268.

Alchourrón, C.E. 1993. Philosophical foundations of deontic logic and the logic of defeasible conditionals. In *Deontic logic in computer science: Normative system specification*, ed. J.J. Meyer, and R.J. Wieringa. London: Wiley.

Alchourrón, C.E., and Bulygin, E. 1971. *Normative systems*. Vienna: Springer.

Alchourrón, C.E., and Bulygin, E. 1984. Permission and permissive norms. In *Theorie der Normen* ed. W.K. et al. Berlin: Duncker & Humblot.

Alchourron, C.E., and A.A. Martino. 1990. Logic without truth. *Ratio Juris* 3 (1): 46–67.

Alexy, R. 1985. *Theorie der Grundrechte*. Frankfurt am Main: Suhrkamp.

Anderson, A., and N. Belnap. 1975. *Entailment: The logic of relevance and necessity I*. Princeton: Princeton University Press.

Åqvist, L. 2001. Deontic logic. In *Handbook of philosophical logic*, ed. D. Gabbay, and F. Guenthner, 2nd ed. Dordrecht: Kluwer.

Blackburn, P., M. de Rijke, and Y. Venema. 2001. *Modal logic*. Cambridge: Cambridge University Press.

Blanchette, P. 2001. Logical consequence. In *The Blackwell guide to philosophical logic*, ed. L. Goble, 2001–2115. Oxford: Blackwell.

Boella, G., Pigozzi, G., and van der Torre, L. 2009. A normative framework for norm change. In *Proceedings of AAMAS 2009*. New York: ACM.

Boella, G. and van der Torre, L. 2003a. Permissions and obligations in hierarchical normative systems. In *ICAIL'03*, 109–118. New York: ACM.

Boella, G. and van der Torre, L. 2003b. Permissions and undercutters. In *NRAC'03*, 51–57, Acapulco.

Boella, G. and van der Torre, L. 2004. Regulative and constitutive norms in normative multiagent systems. In *Proceedings of KR2004*, ed. D. Dubois, A. Christopher, C.A. Welty, and M. Williams, 255–266. Menlo park: AAAI press.

Boella, G. and van der Torre, L. 2005. Permission and authorization in normative multiagent systems. In *ICAIL'05*, 236–237. New York: ACM.

Brown, M. 2000. Conditional obligation and positive permission for agents in time. *Nordic Journal of Philosophical Logic* 5 (2): 83–111.

Carmo, J., and A. Jones. 2002. Deontic logic and contrary to duties. In *Handbook of philosophical logic*, ed. D. Gabbay, and F. Guenther, 2nd ed. Dordrecht: Kluwer.

Chellas, B.F. 1980. *Modal logic*. Cambridge: Cambridge University Press.

Dummett, M. 2000. *Elements of intuitionism*. Oxford: Oxford University Press.

Etchemendy, J. 1990. *The concept of logical consequence*. Cambridge, MA: Harvard University Press.

Fine, K. 2002. Varieties of necessity. In *Conceivability and possibility*, ed. T.S. Gendler, and J. Hawthorne, 253–281. Oxford: Oxford University Press.

Føllesdal, D., and R. Hilpinen. 1971. *Deontic logic: An introduction*, 1–35. Netherlands, Dordrecht: Springer.

Forrester, J. 1984. Gentle murder, or the adverbial samaritan. *Journal of Philosophy* 81: 193–197.

Gabbay, D. 1985. Theoretical foundations for non-monotonic reasoning in expert systems. In *Logics and models of concurrent systems—NATO ASI series*, ed. K. Apt. Berlin: Springer.

Gabbay, D. 2006. Sampling labeled deductive systems. In *A companion to philosophical logic*, ed. D. Jacquette. Oxford: Blackwell.

Gabbay, D., J. Horty, and X. Parent (eds.). 2013. *Handbook of deontic logic and normative systems*. London: College Publications.

Gabbay, D.M. 1994. What is a logical system? In *What is a logical system?*, ed. D.M. Gabbay, 179–216. Oxford: Oxford University Press.

Gentzen, G. 1969 [1934]. Investigations into logical deduction. In *The collected papers of Gerhard Gentzen*, ed. M.E. Szabo, 68–213. Amsterdam: North-Holland.

Governatori, G., F. Olivieri, A. Rotolo, and S. Scannapieco. 2013. Computing strong and weak permissions in defeasible logic. *Journal of Philosophical Logic* 42 (6): 799–829.

Governatori, G., and A. Rotolo. 2006. Logic of violations: A Gentzen system for reasoning with contrary-to-duty obligations. *Australasian Journal of Logic* 4: 193–215.

Governatori, G., and A. Rotolo. 2010. Changing legal systems: Legal abrogations and annulments in defeasible logic. *The Logic Journal of IGPL* 18 (1): 157–194.

Grice, P. 1989. *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Grossi, D. and Rotolo, A. 2011. Logic in law: A concise overview. In *Logic and philosophy today—Volume 2* ed. D.M. Gabbay, 251–274. London: College Publications.

Hansen, J. 2005. Conflicting imperatives and dyadic deontic logic. *Journal of Applied Logic* 3 (3–4): 484–511.

Hansen, J., G. Pigozzi, and L. van der Torre. 2007. Ten philosophical problems in deontic logic. In *Normative multi-agent systems, number 07122 in DROPS proceedings*, ed. G. Boella, L. van der Torre, and H. Verhagen. Germany: Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl.

Hansson, B. 1969. An analysis of some deontic logics. *Nous* 3: 373–398.

Hansson, B. 1971. *An analysis of some deontic logics*, 121–147. Netherlands, Dordrecht: Springer.

Hart, H. 1994. *The concept of law*. Oxford: Clarendon.

Hart, H.L.A. 1982. *Essays on Bentham*. Oxford: Clarendon.

Hilpinen, R. 2001. Deontic logic. In *The Blackwell guide to philosophical logic*, ed. L. Goble. Oxford: Blackwell.

Jørgensen, J. 1937. Imperatives and logic. *Erkentniss*, 288–296.

Kelsen, H. 1991. *General theory of norms*. Oxford: Clarendon.

Kment, B. 2017. Varieties of modality. In *The Stanford encyclopedia of philosophy*, ed. E.N. Zalta, Spring 2017 edition. Metaphysics Research Lab, Stanford University.

Kraus, S., D. Lehmann, and M. Magidor. 1990. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence* 44: 167–207.

Kripke, S. 1959. A completeness theorem in modal logic. *Journal of Symbolic Logic* 24: 1–14.

Kripke, S. 1963. Semantical considerations on modal logic. *Acta Philosophica Fennica* 16: 83–94.

Krogh, C., and H. Herrestad. 1996. Getting personal: Some notes on the relationship between personal and impersonal obligation. In *Deontic logic, agency and normative systems*, ed. M. Brown, and J. Carmo, 134–153. Berlin: Springer.

Kurucz, A., Wolter, F., Zakharyaschev, M., and Gabbay, D.M. 2003. Many-dimensional modal logics: Theory and applications, In *Studies in logic and the foundations of mathematics*, Vol. 148. Amsterdam: North Holland.

Lindahl, L., and J. Odelstad. 2000. An algebraic analysis of normative systems. *Ratio Juris* 13: 261–278.

Lindahl, L., and J. Odelstad. 2008. Intermediaries and intervenients in normative systems. *Journal of Applied Logic* 6 (2): 229–258.

Makinson, D. 1999. On a fundamental problem of deontic logic. In *Norms, logics and information systems. New studies in deontic logic and computer science*, ed. P. McNamara, and H. Prakken, 29–54. Amsterdam: IOS Press.

Makinson, D., and L. van der Torre. 2000. Input-output logics. *Journal of Philosophical Logic* 29 (4): 383–408.

Makinson, D., and L. van der Torre. 2001. Constraints for input/output logics. *Journal of Philosophical Logic* 30 (2): 155–185.

Makinson, D., and L. van der Torre. 2003. Permission from an input/output perspective. *Journal of Philosophical Logic* 32 (4): 391–416.

McCarty, L.T. 1986. Permissions and obligations: An informal introduction. In *Automated analysis of legal texts*, ed. A.A. Martino, and F. Socci, 307–337. Amsterdam: North Holland.

Mendelson, E. 1987. *Introduction to mathematical logic*. Belmont, CA: Wadsworth & Brooks.

Parent, X. 2001. Cumulativity, identity and time in deontic logic. *Fundamenta informaticae* 48 (2–3): 237–252.

Pettit, P. 1997. *Republicanism: A theory of freedom and government*. Oxford: Oxford University Press.

Prakken, H., and M.J. Sergot. 1996. Contrary-to-duty obligations. *Studia Logica* 57 (1): 91–115.

Priest, G. 2006. *In contradiction, a study of the transconsistent*. Oxford: Clarendon Press.

Rotolo, A. 2010. Retroactive legal changes and revision theory in defeasible logic. In *Proceedings of the 10th international conference on deontic logic in computer science (DEON 2010)*, ed. G. Governatori, and G. Sartor, vol. 6181 of *LNAI*, 116–131. Berlin: Springer.

Rotolo, A. 2017. Logics for normative supervenience. In *Supervenience and normativity*, ed. B. Brozek, A. Rotolo, and J. Stelmach. Dordrecht: Springer.

Sartor, G. 2005. *Legal reasoning: A cognitive approach to the law*. Dordrecht: Springer.

Sartor, G. 2006. Fundamental legal concepts: A formal and technological characterisation. *Artificial Intelligence and Law* 14(1–2): 101–142.

Sen, A. 1999. *Development as freedom*. New York, NY: Random House.

Sergot, M. 2013. Normative positions. In *Handbook of deontic logic and normative systems*, ed. D. Gabbzy et al., 353–406. London: College Publications.

Shapiro, S. 2013. Classical logic. In *The Stanford encyclopedia of philosophy*, ed. E.N. Zalta, Winter 2013 edition. Metaphysics Research Lab, Stanford University.

Stenius, E. 1963. Principles of a logic of normative systems. *Acta Philosophica Fennica* 16: 247–260.

Stolpe, A. 2010a. Norm-system revision: Theory and application. *Artificial Intelligence and Law* 18 (3): 247–283.

Stolpe, A. 2010b. Relevance, derogation and permission. In *DEON 2010*, 98–115. Berlin: Springer.

Stolpe, A. 2010c. A theory of permission based on the notion of derogation. *Journal of Applied Logic* 8 (1): 97–113.

Sundholm, G. 1983. Systems of deduction. In *Handbook of philosophical logic: Volume I: Elements of classical logic*, ed. D. Gabbay, and F. Guenthner, 133–188. Dordrecht: Reidel.

Tarski, A. 1983 [1936]. On the concept of logical consequence. In *Logic, semantics, metamathematics*, 409–420. Indianapolis: Hackett.

van Benthem, J., P. Dekker, J. van Eijck, M. de Rijke, and Y. Venema. 2001. *Logic in action*. Amsterdam: ILLC, University of Amsterdam.

van Benthem, J., Grossi, D., and Liu, F. 2013. Priority structures in deontic logic. *Theoria*. 80(2): 116–147.

van der Torre, L. 1997. Reasoning about obligations: Defeasibility in preference-based deontic logic. PhD thesis, Erasmus University Rotterdam.

von Jhering, R. 1924. *Geist des römischen Rechts auf den verschiedenen Stufen seiner Entwicklung*, 1st ed. 1852–1865. Leipzig: Breitköpf.

von Wright, G. 1963. *Norm and action: A logical inquiry*. Routledge and Kegan Paul.

von Wright, G.H. 1951. Deontic logic. *Mind* 60: 1–15.

Inductive, Abductive and Probabilistic Reasoning

Burkhard Schafer and Colin Aitken

1 Introduction

Inductive reasoning plays a central role for the way in which we learn about our world as children and adults (Tenenbaum et al. 2006; Osherson et al. 1990), how we reason about it (Heit 2000), structure our experience of it (Halford et al. 1998) and generally navigate more or less rationally our way through our environment (Einhorn and Hogarth 1981). It plays a role in the most mundane of inferences about our chances of catching the bus which is just *always late* at this time of the day (e.g. Nisbett et al. 1983) to the most sophisticated scientific theories. Indeed, in the early days of the scientific revolution, induction and scientific rationality were treated as almost synonymous (so, e.g., famously by Bacon (1887) in 1620 and then Mill (1843) and Whewell (1840); see generally Milton (1987)). Indeed, we find even earlier precursors such as the eleventh century Persian scientist Ibn al-Haytham who berated Aristotle for his neglect of this method (Plott 2000, 462), or his more famous contemporary and compatriot Avicenna, who comes to an equally critical evaluation of Aristotle's work on induction (McGinnis 2003). While induction as the main contender for scientific reasoning had temporarily fallen out of favour under the influence of Popperian falsificationism, it regained more recently the interest of epistemologists in the form of Bayesian confirmation theory. Confirmation theorists see induction as central to the understanding of the scientific reasoning process (see, e.g., Maher 1993; Jaynes 2003), even though its role for epistemic justification has been lost. The centrality of induction to human reasoning has been recognised by cognitive scientists and philosophers alike (see Holland et al. 1998) for an

B. Schafer (✉)
Law School, The University of Edinburgh, Edinburgh, UK
e-mail: B.Schafer@ed.ac.uk

C. Aitken
School of Mathematics and Maxwell Institute, The University of Edinburgh, Edinburgh, UK
e-mail: C.G.G.Aitken@ed.ac.uk

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_11

interdisciplinary analysis and unsurprisingly also informed attempts to replicate human intelligence computationally, in theories of machine learning (Michalski 1983; Goldberg and Holland 1988; Langley and Simon 1995), pattern recognition (Michalski 1980) or information retrieval and data mining (Cooper and Herskovits 1992; Mozer 1984).

Despite this prominence of inductive reasoning as a topic of investigation in a variety of disciplines, as Hunter (1998) noted, induction has received scant attention in the jurisprudential literature on legal reasoning. The same applies with some notable exceptions to research in computer science that aims to model legal reasoning tasks (e.g. Zeleznikow 2004; Rissland and Friedmann 1995). The first reason for this neglect is more terminological than substantive. Analogical and case-based reasoning have of course played a prominent role in theories of legal reasoning, so prominent that they have indeed individual chapters in this book. Many writers on analogical reasoning simply did not explore its relation to induction, while some explicitly juxtaposed and contrasted them (e.g. Ashley 1992; more nuanced Hunter 1998, 391). We will discuss this in more detail later, but note here that there are strong traditions in the theory of science that treat analogical and case-based reasoning as a form of induction rather than an alternative form of reasoning. The second, more substantive reason is historical in nature. An influential school of American legal formalism of the nineteenth century, most notably the legal education reformer Christopher Columbus Langdell and his follower William Keener strongly promoted inductive reasoning as a necessary component of what they perceived to be a proper "science of law" (see, e.g., the preface in Langdell 1871; Keener 1894). In this view, a scientific approach to law meant in particular legal certainty, which in turn was to be achieved by mirroring the natural sciences. In the same way in which the law of gravity is a general, certain, precise and unchanging law, discovered through induction from the multiple observations of singular objects falling to earth, so general, precise, certain and unchanging legal principles can be discovered inductively from analysing cases. The lasting legacy of this approach was the case-based method as a standard of legal education in the USA. For the concept of induction however, the association with a very specific conception of legal science (and indeed science in general) proved problematic. It identified natural sciences with induction at a time when inductivism and its main philosophical proponent, neo-positivism, were already on the wane, and while Popper's claim to have "murdered" neo-positivism in the 1930s proved premature by two decades or so, credible alternatives began to emerge. So even those who considered an orientation of law towards the methodology of the natural sciences as a valid project in principle abandoned the focus on induction that had characterised the movement at least in the USA. The criticism of the very ideal of legal certainty through abstract rules and principles by the American legal realist and related schools of thought, such as critical legal studies or law and economics, raised even more problems. By rejecting the ideal of legal certainty, these intellectual traditions also attacked the method of induction that according to their opponents was its guarantor (see, e.g., Landman 1927). Ironically, their own emphasis on socio-legal studies as foundational for law in juxtaposition to formalist accounts could have provided, as we will see below, a much stronger case for inductive inference in law.

By contrast, formalist approaches in the civilian tradition of continental Europe proved to be more durable, yet these approaches faced the opposite problem. As we will discuss in more detail below, inductive inference is by its very nature "inference under uncertainty," its results open to revision in principle. In the empirical realm, to establish inductively a universal statement like "all ravens are black" involves an inference from the finite number of ravens that we have observed to all ravens, past, present and future. Since we cannot observe them all, a residue of uncertainty always remains. This is at the heart of Hume's influential "problem of induction"—how can we justify our belief in universal laws when our basis will always be insufficient (see, e.g., Howson 2000)? In law, the situation seems to be profoundly different—to establish that murder is always prohibited, we need not observe a finite number of "murder events" and note the reaction of the legal system. Rather, a single data point, the relevant penal statute, is necessary and sufficient to establish if the rule holds in a given legal system. Deduction, arguments from general premisses (statutes) to specific conclusions, rather than induction, arguments from specific observations to general conclusions, consequently becomes the focus of attention. As a result, the "legal syllogism," a form of deductive reasoning that can be traced back to Aristotle, dominates the debate on legal reasoning in all approaches that are influenced by this continental European way of thinking, for proponents and critics of logic in law alike.

A typical early example, taken from the common law tradition, is Baldwin who wrote in 1903:

> Any commendation, therefore, of the inductive method as the only scientific manner of investigating natural phenomena and physical problems is irrelevant to the question of applying the inductive method in legal education. That method is all important in deriving certainties from uncertainties, the knowable from the unknown. But it is worthless, except as a mode of mental discipline, when applied to deriving known principles of law from recorded opinions of certain judges, of which these principles are, or are intended to be, the foundation, and in which they are generally named and stated with more or less of formal precision. When thus used, it becomes more properly a deductive method, proceeding from analysis to synthesis. (Baldwin 1903, 7)

Somewhat paradoxically, induction in law thus became for some legal thinkers too much connected to the notion of legal certainty, for others not enough.

It is only recently that an interest in inductive reasoning in law was rekindled. Two developments facilitated this renewed interest. One development was the revival of the Bayesian approach to probabilities. The term *Bayesian* refers to Thomas Bayes, a nonconformist minister, student of the University of Edinburgh and Fellow of the Royal Society, who in the eighteenth century proved a version of the theorem that is now called after him (Bayes 1764; see Stigler 1982). A general version of Bayes specific case was introduced by Pierre-Simon Laplace, who showed its potential by applying it to topics as diverse as celestial mechanics and medical statistics (Laplace 1814; see Dale 1982). Moreover, he was also the first to notice its relevance for jurisprudence and legal reasoning (Daston 1981). Later in this chapter, the likelihood ratio is discussed as a measure of the value of evidence. It is part of the "odds version" of Bayes' theorem, and in a legal context its use is recorded in the infamous Dreyfus case. There, the question was what odds we should assign to the event that a forged document would display certain characteristics.

> […] since it is absolutely impossible for us (the experts) to know the a priori probability, we cannot say: this coincidence proves that the ratio of the forgery's probability to the inverse probability is a real value. We can only say: following the observation of this coincidence, this ratio becomes *X* times greater than before the observation. (Darboux et al. 1908, 504)

The likelihood ratio and its logarithm, known as the *weight of evidence*, were used at Bletchley Park in World War II for cryptanalysis. The term *weight of* evidence for the logarithm of a Bayes factor was first used by Charles Saunders Peirce (1956) in a paper on the probability of induction. Despite this promising early combination of Bayesian statistics and legal reasoning, the Bayesian approach faded in prominence especially when in 1920 a new paradigm in statistical analysis, frequentist statistics, was developed and promoted by the hugely influential Ronald A. Fisher, Jerzy Neyman and Egon Pearson (on frequentist probability theory, see Neyman 1977).

To understand the difference, we need to explain a bit more about Bayes' ideas first. Thomas Bayes brought two ideas to the consideration of uncertainty. The first idea was that probability, a number between 0 and 1, could be used as a subjective measure of belief in the outcome of an event. The event could be the outcome of a sporting event such as a football match, for example: what do you personally think the odds are that Manchester United will win a game against Barcelona, what would you be willing to bet, for whatever reason, on a win? The second idea is that this measure of belief could, and should if the evidence suggests it, be amended in the light of new evidence. Thus, the measure of belief in the outcome of the football match could be amended if one team scored a goal, and amended again every time a goal was scored. These probabilities could also be combined with so-called objective probabilities of events, such as the outcome of the toss of the coin if this is chosen to decide the result when the game ended in a draw after extra time and penalties, where all are agreed that if the coin is fair, the probability of a head equals the probability of a tail and both equal one half.

The frequentist approach is so-called because probabilities are derived from a thought experiment in which the probability is considered as the long-run relative frequency of the outcomes of repeated experiments made under identical conditions. This idea is easily applicable to tosses of a coin which it is thought may not be fair. The coin can be imagined to be tossed an infinite number of times and the ratio of heads to tails observed from which the probability of a head and of a tail may be derived. In practice, the coin may be tossed a large number (but finite) times and the relative frequency determined. In the frequentist paradigm, it is not possible to consider the probability of the outcome of a unique sporting event, such as a Manchester United versus Barcelona match, as this would require first to have both teams play each other numerous times under identical conditions.

It is easy to see that the Bayesian paradigm fits well with the legal context. The concept of a probability of guilt can be translated into the Bayesian paradigm, while it is not possible, or at least not straightforward to conceive of it in the frequentist paradigm. A frequentist can consider the probability of evidence in the form of measurements, such as the chemical composition of gunshot residue. The frequentist can then consider the probability of these measurements, given the residue came from a specific gun. They can also consider the probability of these measurements given the residue came from some other gun. The Bayesian too can consider these conditional probabilities. The Bayesian, however, can then combine the ratio of these probabilities with the prior odds in favour of the residue having come from a specific gun, as compared with originating from a different gun to obtain posterior odds in favour of the residue having come from a specific gun, as compared with originating from a different gun. This the frequentist cannot do, but for lawyers, interested in reconstructing individual historical events, it is all important.

Frequentist analysis made a major contribution to empirical, observation-based science, in particular when as in evolutionary biology or sociology, large populations are studied. By the 1960s however, dissatisfaction with the limitations of the frequentist approach resulted in a resurgence of interest in Bayes' theorem, and from then onwards the term "Bayesian" can be found as a label for an entire approach to probabilistic reasoning (Fienberg 2006, 5), greatly helped by the advent of ever-increasing computer power. By then, two different schools of Bayesian reasoning had evolved. One, the objectivist school shared with the frequentist approach a focus on statistical analysis that depends entirely on the analysed data and eschews any element of subjective decision. For legal reasoning however, where the necessary data sets for entire populations are often not available, the emergence of a "subjectivist" school turned out to be of pivotal importance. The idea that "probability" should be interpreted as "subjective degree of belief in a proposition" was proposed by amongst others John Maynard Keynes and taken further by Bruno de Finetti and Frank Ramsey (Gillies 2000, 1–50). This understanding of probability theory as a theory of rational belief revision turned out to be of particular relevance for legal reasoning: Jurors for instance will enter the jury room with a rich background knowledge and assumptions, for instance "the chances that police officers sometimes lie" or that "alcohol leads often to violence" that cannot be reduced easily to objective statistical statements. Regardless however of what their starting point is, we can model how their ideas about the defendant's guilt or innocence should change as new evidence is adduced (see, e.g., Froeb and Kobayashi 1996). This leads us to the second development in the renewed interest in inductive and probabilistic reasoning in law, the new evidence scholarship movement and a new focus on reasoning about facts in law (see, e.g., Lempert 1977; Twining 1984; Jackson 1996). We noted above that inductive inferences are central to scientific reasoning. With the increasing importance of forensic science in the trial process, a more pressing need arose to understand this previously often neglected part of legal reasoning, and with it also a new interest in inductive inferences (see, e.g., Aitken and Taroni 2004).

This chapter focuses on inductive reasoning, but we will also touch briefly on a related form of reasoning under uncertainty, "abductive" or "retrograde" inferences. We begin by illustrating the differences between these types of reasoning through a number of short examples that will follow us through the chapter. At the end of this introductory part, we briefly discuss where in the legal system we can encounter these forms of argument and try to match them with key legal concepts and ongoing legal or jurisprudential debates. In the remainder of the chapter, we discuss both of the inference forms in turn, linking them to wider philosophical discussions and problems in argumentation theory and jurisprudence. One important aspect will be to establish what distinguishes a valid from an invalid, a convincing from a weak instantiation of each of these argument forms. As we will see, even consideration of this aspect raises complex questions for epistemology, logic and theory of argumentation.

2 Tales of Woe and Reasoning

Consider the following situation: two police officers, one very experienced, the other a novice, are called to a residential address. The neighbours are worried, they have not seen the elderly resident for a week, and full milk bottles are accumulating at his doorsteps. When the neighbours knock on the door, there is no answer.

The experienced police officer reasons:

Argument (1)

- Last time, I was called out at this address, he had gone to visit his son and forgotten to tell the neighbours and the milkman.
- The time before, he had gone on holiday and forgotten to tell the neighbours and the milkman.
- The time before that, the housing association had temporarily assigned him a new place to stay, so that they could repair the plumbing in his flat, and nobody had told the neighbours or the milkman.
- Therefore, he is (probably) just gone again somewhere without telling anybody.

He therefore assigns the investigation a low priority and does not intend to ask a crime scene specialist to accompany them to the address.

His younger colleague however, fresh from a training course, remembers a bit of information about the estate in question: the complex of flats are owned by the council, which uses them to house elderly and vulnerable people. There had been a spate of burglaries over recent weeks. Only two weeks ago, they installed a concierge in the entry hall to increase security. Whenever a resident now leaves, he signs out with the concierge. In the absence of the resident, visitors for him will be denied entry, including delivery and salespeople.
He therefore reasons:

Argument (2)

- Whenever a resident leaves the complex, the concierge signs him out.
- Whenever the concierge signs someone out, visitors, including those making deliveries, are denied entry.
- Full milk bottles at his flat door, within the complex, mean that those making deliveries were not denied entry to the complex.
- He therefore did not leave the complex.

He manages to convince his colleague with his reasoning that the issue might be more urgent, and they immediately go to the flat, bringing a doctor with them. Upon entering the flat, they immediately notice a terrible smell. In the bedroom, they find a partly decomposed body of a male on the bed. The body is dressed in pyjamas, and despite the decomposition it is visible from the exposed arms that the veins have been cut several times, resulting in severe blood loss visible on the bed sheets. A kitchen knife is held in one hand.

The younger officer, distraught from the find and not looking too closely, reasons:

Argument (3a)

- There is a dead body with wounds on one arm.
- Suicide with a knife regularly causes wounds on the arms.
- Therefore, suicide is a good explanation of what we found, he must have killed himself.

His older colleague however, hardened by a lifetime experience with dead bodies, looks a bit more closely. He notices that the deceased had deformed fingers typical of severe arthritis. The officer doubts that the deceased could have handled a knife. He also notices that several cupboards have been opened and their content left untidy, as if searched, and that there is an empty space next to a TV cable entry point where discoloration of the surface indicates that an object the size of a TV had been sitting there not long ago.
He therefore reasons:

Argument (3b)

- There is a dead body with wounds on the arm.
- However, his hands are too deformed to hold a knife.
- Murderers who torture people to reveal where they hide money regularly cause wounds on the arms of their victims.
- An attempt to dress up a murder as suicide also regularly causes wounds on the arms of the victim.
- Therefore, murder is a good explanation of what we found.
- The killer might have tortured him and then tried to make it look like a suicide.

He therefore asks the doctor and the crime scene specialists to check:

- If it would have been possible for the deceased to hold the knife, giving his medical condition.
- If there are fingerprints other than those of the deceased on the knife.
- If his hands show defensive wounds, and if there are traces of extraneous material under his fingernails.
- If there is a suicide note in the flat.

Subsequently, the forensic analysis reveals indeed that the deceased's arthritis had been so advanced that he could only use specially modified cutlery, unlike the knife that was found on him. An as-yet unidentified fingerprint is on the blade of the knife, consistent with someone holding it there to press the handle into someone else's hand. Furthermore, under his fingernails a number of red hairs, with attached follicles, are found, suggestive of a fight. The follicles allow a DNA expert to obtain a DNA profile. Questioning of the neighbours reveals that shortly after the deceased was last seen, a red-haired male person, acting suspiciously, had been seen by a neighbour near the deceased's flat. The neighbour had approached him. When asked what he was doing there he had replied with a very strong Scottish highland accent that he had tried to deliver a leaflet, but nobody had been home. He then left.

Shortly afterwards, a known burglar from Inverness is arrested. The witness identifies him as the person to whom he had spoken. A forensic expert analyses the burglar's DNA with that found on the hairs under the fingernails of the deceased and finds that they "match." He is subsequently charged with murder and at trial the prosecutor argues:

Argument (4)

- Members of the jury, we are faced with two possibilities: the accused standing before you could be the person whose hair was found under the fingernails of the victim, or that the hair could come from a third party, as the defence wants you to believe.
- The evidence of the corresponding DNA profiles is a billion times more likely if the hair under the fingernails and the hair from the defendant come from the same person than if they come from different persons.
- There is therefore extremely strong support for the proposition that the hair under the fingernails comes from the defendant.

On this basis, the prosecution could develop a further, abductive argument that hypothesises a possible explanation, or cause, for the match. The most plausible explanation why it was found there could be that the deceased, trying to defend himself, pulled some hairs from the head of his attacker.

2.1 Putting Induction and Abduction into Context

In these short stories, we have encountered several different types of arguments. In the next step, we want to use them to develop a more abstract account of them, ultimately aiming at a theory of inductive reasoning. A first step for the development of any theory is to group, classify and label the data that the theories aim to explain. The most famous historical example for this is the Linnaean system of classification, but tree-like structures like that system can be traced back at least to Aristotle and his *Categories.* In law, we encounter this in the subdivisions of law into private, public and criminal law, or of the legal systems of the world into civilian, common law, non-European and mixed legal systems. Argumentation theory too tries to group and classify different types (and subtypes) of argumentation. However, as so often with attempts to classify and label phenomena, not everybody necessarily agrees with a given proposed subdivision, and argumentation theory is no different.

Different writers will label the examples we introduced above slightly differently, and disagree on the best way to group them together. We will accommodate this by committing ourselves to vocabulary and definitions that are widely used and highlight some of the more influential alternative proposals to direct the reader to further literature that focuses on these debates. The reader should bear in mind though that these classifications and definitions are chosen for their usefulness within a specific theoretical framework, they are not "facts" about the world and ultimately neither right nor wrong, but only useful in varying degrees.

With this in mind, we can have a look at the four arguments listed above. Of these, one is very clearly different from the others. In Argument 2, the truth of the premisses guarantees the truth of the conclusion; that is, *if* the premisses are true, the conclusion, with necessity, is true as well. In all the other examples, it is possible that the premisses are true, and the conclusion is nonetheless false. One common way to distinguish a *deductively* valid argument from an inductive argument is to say that in a deductively valid argument, the truth of the premisses *guarantees* the truth of the conclusion, that it is *not possible*, not even in a purely imagined or hypothetical world, that the premisses can be true and the conclusion false. An inductive argument by contrast gives us merely good reasons to accept the conclusion *for the time being* (they make the conclusion "more likely"), but we must be ready to revise the conclusion in the light of new data. This is very obvious in the case of Argument 1: even though it is true that the deceased had left his flat on at least three previous occasions without telling anybody, the assumed conclusion, that he did the same thing this time round, turned out to be wrong. Similarly with Argument 3a, while it is true that suicides can leave knife wounds on the arms, the conclusion, that *these* wounds were self-inflicted, turned out to be wrong. The issue is, however, not just that that these two conclusions were factually wrong. Even in the case of Argument 3b that correctly identified the reason of the knife wounds, the correctness or truth of this conclusion is not *guaranteed* by the fact that the premisses were true as well. For things *could have* easily been otherwise: the police *might* have found after a more thorough search a suicide note for instance where he explains that he had sold the TV to pay for additional (illegal) painkillers, but now, having searched for a hoped-for last remaining dose everywhere (hence the disordered cupboards), he could not cope with the pain any longer and opened his veins by leaning on the knife wedged between his hand and the bed (thus explaining why he was able to hold it after all).

Even in the last example, the conclusion, that the hairs in the victim's hand came from the attacker, was not *guaranteed* by establishing a DNA match. Even though the match probability was very low indeed, the DNA *might* still have come from the accused's twin brother, or someone else, typically a blood relative, with very similar DNA. This means that the reasoning displayed in arguments 1, 3 and 4 is typical for situations where we do not have all the facts (yet), where we have to "reason under uncertainty." As time progresses and we find out more things, a previously held conclusion may suddenly become open to revision. These last two examples give us also another, equivalent way to explain the differences between the types of arguments: in the case of arguments 1, 3a/b and 4, we have to be ready to revise our conclusion if new evidence is added to our premisses—and that even when none of the older premisses has to be abandoned, and they all remain valid and well supported. So in Argument 1, we can simply add the information (e.g. established after further interviews) that there is now a checkout system, and the original inference is no longer plausible. In Argument 3b, we can add the information that there is a suicide note, and the inference is no longer plausible. This is different in Argument 2. Of course, also this argument could lead to a factually wrong conclusion, and we need to keep an open mind during an investigation that this might happen. Crucially though, *if* our conclusion turns out to be wrong, then at least one of the premisses must have been wrong as well. This means that as long as we simply add new information, the inference will remain valid. Only if we modify, or discard one of the original premisses can we reach a different conclusion.

This property of some forms of argument is also called "monotonicity." It means that as we add premisses to the argument, the number of conclusions we draw can remain the same, or increase, but never fall—we can't ever "unlearn" or "forget" a conclusion that was derived from a monotonic argument. Logical deductions, of which Argument 2 is an example (a chain syllogism in the form of *modus tollens*, to be precise) are monotonic. Here, the truth of the premisses guarantees the truth of the conclusion, so the only way to revise the conclusion is also to revise and discard at least one premise. Deductive arguments are discussed elsewhere in this book, and we have introduced one example only because the arguments or inferences that interest us are typically defined by contrasting them to deductively valid arguments. Inductive arguments by contrast are non-monotonic. As we have seen when we introduced new variations to Argument 3b, with these non-deductive arguments, we can add new premisses that force us to revise the conclusion even if all the older premisses remain true—they are non-monotonic. Non-monotonic logics have over the past few decades taken centre stage in a variety of disciplines that analyse our reasoning capacity, especially in computer science and artificial intelligence research (see, e.g., Brewka et al. 1997; Lukaszewicz 1990; McDermott and Doyle 1980). While non-monotonic logic can model the evolution and structure of scientific theories (Aliseda 2004), much of our everyday "common sense" reasoning is also of this form (Brewka 1991), and it is also linked to our ability to learn and to revise old beliefs in the light of new information (Boutilier 1996; Kakas and Riguzzi 2000; Aliseda 2006; see also above, the comments about the idea introduced by Bayes that probabilities of beliefs can be updated in the light of new information).

Legal reasoning too is often best understood as non-monotonic. In particular, non-monotonicity accounts for the adversarial, dialogical nature of legal disputes (Prakken and Sartor 1997; Gordon 1988). In a legal dispute, both sides will present arguments that typically lead to contradictory conclusions. However, it is rarely the case that one side committed a straightforward logical mistake, or stated a falsehood, especially if they are competently represented by professional advocates. Rather, both sides make arguments that are in varying degrees plausible or convincing. This in turn means that even if we can agree with everything one party has said, it might still be rational to consider the arguments of the other side "more convincing." Sometimes, but not necessarily always, this can be due to the second party having information that the first party has not yet uncovered. In the legal doctrine of most jurisdictions, this finds its expression also in the appeal court structure, especially post-conviction scrutiny in criminal cases. This allows reopening a criminal conviction on appeal when new evidence comes to light—without having necessarily to show a mistake (other than not considering the new evidence) in the trial of first instance. For instance, a new scientific test could be discovered that allows the analysis of smaller traces of DNA than possible at the time of the first trial. Using this new test, it can be shown that the accused could not possibly be the culprit. Even if the trial at first instance had very strong eyewitness evidence, we might now have to abandon the original conclusion. If legal reasoning were always strictly deductive and hence monotonic, this would be difficult to explain.

One consequence of the defeasible nature of conclusions that we reach through inductive or abductive reasoning is that it is important to *test* them. We can see this in particular in the case of abductive reasoning—our arguments 3a and 3b. Here, our two police officers derive abductively two competing possible explanations for the same observation they both have made: the dead body with cuts on the arms. Both explanations have a degree of plausibility, so how can we determine which one is true, or at least better? One way is to assume the explanation hypothetically, and then ask what else we should expect to find, assuming it is true, and what we must not find, assuming it is true. In the case of 3a, if it had been suicide, we should expect to find an authentic suicide note, and we must not find that the deceased had been too ill to hold a knife with which to inflict the wounds.

In our example, the police did not find a suicide note, but found that he was indeed incapable of killing himself in the way hypothesised. The latter falsifies the theory—it cannot be the right explanation (on falsification in general, see Lakatos 1970; on the role of falsification in evidentiary legal reasoning, see Kaye 2004). The logical form of falsification is the *modus tollens* in deductive logic:

- If *A* were true, *B* must not be true as well
- *B* is true
- Therefore, *A* is not true.

In this case, *A* is a possible explanation for some fact *F*, in our example the idea that he committed suicide. It is derived abductively, and then tested through a deductive process, resulting in a combination of abductive hypothesis formation and deductive testing that is known as the "hypothetico-deductive model of scientific inquiry" (see, for a legal context, e.g., Keppens and Schafer 2006). The interplay between different modes of reasoning, and the interdependence of inductive, abductive and deductive argument forms in particular, will also concern us later in this chapter.

When the experienced officer tested the "suicide hypothesis," he relied again on certain general statements about the world, derived from his experience. He noted in particular that knife wounds on arms can occur both in murder and suicide cases, having in the past observed both scenarios. However, he also observed that in cases of suicide, there was almost always also a suicide note, something missing in cases of murder. While the observation "the victim has knife wounds" does therefore fail to discriminate between the two hypotheses, the observation of a suicide note by contrast allows us to rule out one hypothesis in favour of the other. This type of reasoning too is a form of induction, but unlike the *enumerative induction* of (3a) and (3b), this type of induction eliminates a hypothesis by varying a parameter ("suicide note found") and is therefore often called "eliminative" or "varying" induction (Hunter 1998, 370). Its most important application outside the law is the double-blind trial that tests the efficacy of medical treatments. John Stuart Mill called this inference the "method of difference":

> If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance save one in common, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or cause, or a necessary part of the cause, of the phenomenon. (Mill 1843, 455)

In our example, the circumstances that murder and suicide have in common are knife wounds. However, a suicide note only appears in case of suicide and not of murder (admittedly, an oversimplification in many ways), and hence, we can infer that in this case, the decision to commit suicide caused the suicide note that was found on the crime scene. Standard definitions of induction that describe it merely as "reasoning from the particular to the general" do not capture this type of induction, rather, in this case we abandon one particular explanation in favour of another. Yet, as Hunter convincingly argues, it is this type of inference that will typically play at least an equally important role in legal reasoning (Hunter 1998, 372). He gives as an example a situation where a lawyer advises a father about his chances of winning a custody battle. The lawyer knows from experience that in all cases he ever attended, the presiding Judge X always awarded custody to the mother. This allows for the enumerative inductive inference that he will probably also award custody to the mother in his client's case. However, there may be a "confounding factor" in the analysis—maybe in all the cases the lawyer attended, the father was in addition to being male also an alcoholic. That would mean that his teetotal client is not necessarily disadvantaged by his gender. If the lawyer had observed the judge in the previous case with an alcoholic mother, the result would allow an eliminative inference: if the judge still decided for the mother, the inference to the general rule "Judge X always favours mothers" would be confirmed, if not, the alternative theory, that he always decides against the alcoholic, gains strength.

Mill offered in his analysis a number of similar inference forms that remain relevant to this day. Next to the method of difference is the method of agreement:

> If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree, is the cause (or effect) of the given phenomenon. (Mill 1843, 454).

This type of inference is crucial for our ability to recognise patterns. Consider a hospital ward. Whenever patients die from apparent (and otherwise unpredicted) cardiac arrest, nurse X is on shift. The patients differ considerably in age, health condition, illness and prognosis. Further investigation discovers that similar occurrences in which patients died from apparent (and otherwise unpredicted) cardiac arrests happened at the nurse's previous place of employment, so that we can rule out things like unhygienic kitchen or asbestos in the walls. At this point, we may infer a hypothesis that nurse X is the cause of the deaths. However, there are dangers with probabilistic reasoning when not applied carefully and without proper analysis of the data. A recent example of these dangers is the case of Lucia de Berk in an infamous miscarriage of justice (Schneps and Colmez 2013, Chap. 7). Here, the prosecution was not careful enough with the elimination of other circumstances that the various deaths had in common.

Both method by agreement and method by difference can be combined into more reliable inferences, what Mill called the "joint method." In the case of the suspected nurse, we can, e.g., observe that whenever she is on shift, and never when she is not on shift the mortality rate increases—induction "by agreement." If we further observe that in one week, she fell unexpectedly ill, and immediately the mortality rate fell, we have further evidence that she is a causal factor in the death of the patients, thus combining inductive "agreement" and "difference."

A slightly different method suggested by Mill is the method of residue;

> Subduct from any phenomenon such part as is known by previous inductions to be the effect of certain antecedents, and the residue of the phenomenon is the effect of the remaining antecedents. (Mill 1843, 465).

Consider in our example the following variation: our murder suspect is found with £500 in his flat. We know that every payday, he takes £200 in cash and leaves the rest on his bank account. In this case, we can "subduct" these £200, assuming he was arrested on his payday, which leaves £300 unaccounted for. For this sum of money, we have not yet established through "previous induction" a general rule that could explain the £300. Abductively, we can now reason that this is money he took from the victim.

Finally, Mill discusses the method of concomitant variations

> Whatever phenomenon varies in any manner whenever another phenomenon varies in some particular manner, is either a cause or an effect of that phenomenon, or is connected with it through some fact of causation. (Mill 1843, 470)

Consider an environmental tort case: due to industrial spillage, a river has been polluted, resulting in high levels of toxicity and the death of all fish. During the trial, it transpires that water has been contaminated with the effluent of two companies, one containing lead, the other cochineal-based dye that coloured the water red. The question is which of the two caused the death of the fish and is liable for damages. If tests show that whenever we increase lead in water, more fish die, and if we reduce the amount, more survive, and by contrast, increase or decrease of the dye has no effect on the number of dying and healthy fish, we can infer that the lead and not the dye caused the damage.

Due to the importance of non-monotonic reasoning for law, where arguments can often be revised ("defeated") in the light of new information, "defeasibility" will be discussed in more detail in a separate chapter. In this chapter, we look at two specific examples of defeasible, non-monotonic reasoning: induction and abduction. Some philosophers, most notably Rudolf Carnap (1952), used the term "induction" to mean all forms of non-deductive, non-monotonic logic. While it has remained difficult to define exactly what sets induction apart from other forms of non-monotonicity, most authors have found it helpful to subdivide the field further. We follow the influential treatise by Charles Saunders Peirce (1903) and distinguish in particular non-monotonic inferences similar to our Argument 1 from those in arguments 3a and 3b. The term *induction* is typically reserved for arguments for the first type and its variants. Arguments of the second type, where we infer a possible explanation for some observed facts, are called *abduction*. Or in Peirce's own words: "Deduction proves that something *must* be; Induction shows that something *actually is* operative; Abduction merely suggests that something *may be*" (Peirce 1903, 173). What about Argument 4, the match probability of two DNA samples that was used to argue for the guilt or innocence of the accused? This too is *prima facie* a non-monotonic argument—we might find new information that renders the inference invalid, even though all premisses remain valid and undisputed. The reason that a so-called match of two DNA profiles has a probability assigned to it—"The probability of two unrelated individuals sharing the same forensic DNA profile is around one in a billion"—is because we never compare the entire DNA profile of the suspect and the profile of the sample from the crime scene, this would be prohibitively expensive and time-consuming, while telling little of value. Rather, selected samples from the two DNA specimens are taken, typically in the teens of "markers," known as short tandem repeats (STRs), and a sex marker. STRs are short sequences of DNA that are repeated in tandem several times, and the number of repeats for each marker varies between individuals. It is possible that after having compared ten such markers, for example, and found that they match, an eleventh marker would show a difference, falsifying the hypothesis that the specimen came from the same individual (a simplification for the sake of the argument, given recent developments in DNA profiling with low copy number technology). As in the first argument, new information, here an additional eleventh test, can defeat an inference, even though the other pieces of evidence, the previously established ten matches, are all correct. The logical form of this inference is therefore not so dissimilar from our first argument after all:

Argument (4b)

- The first marker of the DNA profile obtained from the crime scene specimen corresponded to the same marker from the DNA profile of the suspect.
- The second marker of the DNA profile obtained from the crime scene specimen corresponded to the same marker from the DNA profile of the suspect.
- …
- The tenth marker of the DNA profile obtained from the crime scene specimen corresponded to the same marker from the DNA profile of the suspect.
- Even though there is a match in ten markers, it is still possible that the eleventh markers from the two samples do not match.

2.2 Induction and Abduction in Law

So far, we have gained a very broad idea of two types of non-deductive reasoning. Before looking at them in more detail, we discuss briefly where, in the legal system, we encounter them. Inductive and abductive reasoning are "reasoning under uncertainty." It is therefore commonly found where we have to make a decision, often under time constraints, before we can collect all pertinent facts, or where it is in principle impossible to ascertain all the facts. Within the trial process, this applies typically to reasoning about the facts of the case. In a trial, we have to reconstruct a unique, non-repeatable event from the past, from whatever traces the event has left behind. Just as we often have to revise historical accounts in the light of new data—e.g., the new discovery of a previously lost document, or the discovery of a ruin during an archaeological dig—so our beliefs of the fact of a case may have to be revised if new witnesses come forward, new scientific testing methods are discovered, or another person confesses to a crime for which someone else has been convicted. In law, we recognise this inherent uncertainty through the appeal process, or concepts such as the burden of proof, which never demands total certainty, but at most a belief beyond reasonable doubt, and often only a finding for a plaintiff or defendant on the balance of probabilities. Given all we know now, the account of the plaintiff is more convincing than that of the defendant, but as we learn more, this may change. It is far less obvious how we can have the same type of uncertainty when reasoning about the law. Laws, in order to be valid, need to be promulgated—so in theory, it should always be possible to ascertain with certainty if a putative norm is or is not valid within a legal system. We find an expression of this idea in the legal maxim that "ignorance of the law is no excuse." In reality, this is of course considerably more complicated by the need to interpret statutes, also in the light of precedent cases. But from an idealised epistemological perspective, every citizen is "deemed to know" the law, because the law, in this sense, is essentially knowable. Facts about the physical world or about past events by contrast are never fully knowable; they can only be approximated by better and better theories. Reasoning about facts and evidence in law is such a central concern that a dedicated chapter in this book will deal with this topic more fully. While most of the examples in this chapter will come from reasoning about facts, it would be a mistake to think of induction and abduction as limited to legal fact-finding.

As indicated at the beginning of this chapter, one of the most important examples of inductive reasoning in law is analogical reasoning with cases and precedents. Again, due to its importance, a separate chapter will deal with these issues in detail. Two definitions of "induction," case-based reasoning and inductive reasoning, while sharing some similarities, are ultimately different (see Ashley 1992). Inductive reasoning is in this case reserved for inferences that, in a classical definition of induction, "reason from the particular to the universal." Case-based reasoning by contrast reasons "from the particular to other particulars." We can again use our examples to illustrate the difference. Above, we described the reasoning of the senior police officer in Argument (1) as

- Last time, I was called out at this address, he had gone to visit his son and forgotten to tell the neighbours and the milkman.
- The time before, he had gone on holiday and forgotten to tell the neighbours and the milkman.
- The time before that, the housing association had temporarily assigned him a new place to stay, so that they could repair the plumbing in his flat, and nobody had told the neighbours or the milkman.
- Therefore, he is (probably) just gone again somewhere without telling anybody.

In this case, not only are all the premisses statements about particular events, not general statements in the form of rules, but also the conclusion pertains only to one specific event. However, we could have reconstructed the argument also in a different form, as

Argument (1b)

- Last time, I was called out at this address, he had gone to visit his son and forgotten to tell the neighbours and the milkman.
- The time before, he had gone on holiday and forgotten to tell the neighbours and the milkman.
- The time before that, the housing association had temporarily assigned him a new place to stay, so that they could repair the plumbing in his flat, and nobody had told the neighbours or the milkman.
- Therefore, *it is always the case that when he is not answering the door, he just left without telling people.*

The particular premisses are used to derive a universal conclusion in the form of a general rule: it is *always* the case that X. To reach the desired conclusion, a decision in the specific case, another argument is necessary, now in the form of a deductive inference:

Argument (1c)

- It is always the case that when he is not answering the door, he just left without telling people.
- He is not answering the door.
- He just left without telling people.

For some authors, only arguments of the form 1b are inductive arguments properly so-called. However, Rudolf Carnap (1952) and his extensive definition of induction and John Stuart Mill, whose treatise from 1843 on "*System of Logic, Ratiocinative and Inductive: ...* " can be seen as the first rigorous and systematic analysis of induction, treat inferences from one set of particulars to another particular not just as one particular variety of induction but its paradigmatic form.

To complicate things further, we can also rewrite Argument (1) as

Argument (1d)

- Last time I was called out at this address, he had gone to visit his son and forgotten to tell the neighbours and the milkman.

- The time before, he had gone on holiday and forgotten to tell the neighbours and the milkman.
- The time before that, the housing association had temporarily assigned him a new place to stay, to repair the plumbing in his flat, and nobody had told the neighbours or the milkman.
- If someone left his flat without telling his neighbours at least on three occasions in the past, then he probably has just left without telling his neighbours the next time nobody answers the door.
- Therefore, he probably has just left without telling his neighbours the next time nobody answers the door.

Now the inference is a valid deductive argument, and all doubt or uncertainty has been "smuggled into" the last of the premisses in the form of a general rule. It is indeed always possible to "reconstruct" a prima facie inductive or abductive argument as a deductively valid inference. Which of the three arguments is the "true" account of his reasoning? That is probably the wrong question to ask—it depends on the context, and on what goal we try to achieve with our analysis. Argument (1) is probably closest to the psychological reality, the process by which the officer reached a conclusion over time, at the point of decision-making. It is unlikely that at this point, general rules were in the forefront of his mind. Argument (1b) captures in a more visible form what Johnson-Laird (1993, 60) identified as the distinct psychological or cognitive characteristic of induction as "any process of thought yielding a conclusion that increases the semantic information in its initial observations or premisses." The universal conclusion "it is *always the case that when he is not answering the door, he just left without telling people*" goes in a very intuitive sense "beyond" the particular premisses that only contain knowledge about a limited range of events.

A particular radical form of "increasing the information content" happens when we derive a universal conclusion on the basis of a single observation. In reasoning about the world, it is obvious that this use of induction is fraught with dangers. In legal reasoning however, we find it as a particularly important form of reasoning, reasoning with precedents—to call it "inductive" though is controversial. In common law jurisdictions in particular, a core skill for judges is to identify the "*ratio decidendi*," "the reason" or "the rationale for the decision." This rationale then is universalised and becomes a new legal rule (McCormick 1987). The *ratio decidendi* is thus the universal principle which the case establishes. In this understanding of reasoning with precedents, the judge typically does not simply compare a past case with the one at hand, but follows the same slightly more circuitous route that was indicated in Argument (1b) and Argument (1c)—carrying out first an inductive inference that establishes a general rule, and then applying the rule to the case at hand. However, as Hunter (1998, 377) argues, to see the derivation of a universal rule from a single precedent as a form of induction is potentially misleading. Following Murray (1982), he argues that an inductivist view of analogy (which he ascribes to Richard Posner) requires that a precedent case uniquely determines the ratio that it establishes. Using some famous precedents, including the famous *Donoghue v Stevenson* [1932] UKHL 10, Hunter shows convincingly that it is possible to derive from this precedent several distinct ratios, each of them capable of supporting the actual result. While he succeeds in this part of his critique, it is far less obvious why in an inductive argument, the premisses should uniquely determine a conclusion. If I observe many black ravens, I can infer that all ravens are black. But I can equally infer that all birds are black, or even that blackness enables flight. And yet, this is the sort of enumerative induction that Hunter considers a clear example of inductive reasoning.

Argument (1d) finally could be an answer our officer gives after the event and on reflection, for instance if he has to justify his actions in court. They "fit" to the context of justification, which is commonly distinguished from the "context of discovery" (on the distinction, see, e.g., Hoyningen-Huene 2006; for an application to legal theory Amaya 2007, 440). The former describes the heuristic devices used to reach a new conclusion and the latter focuses our attention on the soundness of the reasoning process. Induction is closely linked to heuristics and discovery, deduction to justification in this view. The idea that inductive and deductive reasoning require different cognitive strategies has recently been given support by results from neuroimaging. Two groups of test subjects performed different reasoning tasks, one group solving a problem through induction and the other through deduction. It seems that even though they reached similar conclusions, they used different parts of the brain for this indicating if nothing else, the reality of the psychological difference between them (Goel et al. 1997). Being able to reformulate the inductive inference as a deductive argument also shows that we can *simulate* inductive reasoning using deductive logic, a notion that is of interest in particular to researchers in artificial intelligence, since deductive logic has particularly well understood computational properties.

Even if we treat reasoning with precedents of the form

Argument (5)

- In a case involving a man being bitten by an Alsatian, the Court of Appeal ruled in
- 1982 that the owner was liable despite the fact that the victim had been warned not to pat the dog.
- Therefore in the case before us involving a mastiff, the owner should be held liable as well even though he warned the victim.

as a separate category of legal inference, *argument by analogy* or *case-based* reasoning, there are other examples from legal argumentation where we very clearly reason inductively from cases to a universal conclusion. An example from German law is the term "ständige Rechtsprechung," or "established case law." In this case, the court has to determine if a specific and often controversial interpretation of a law has by now been "generally accepted" by the courts and is "settled law," or whether the issue is still controversial (see, e.g., Oberhofer 1992). In this case, the court will conclude inductively from the fact that many courts have followed in the past a certain interpretation that this interpretation is now generally, or universally accepted. A very similar process can be observed historically when countries unified and systematised their “common law,” understood here not as a set of binding precedents but merely the actual practice of citizens and courts alike. This involved typically a process of “pattern recognition” where from the multitude of individual decisions and practices, a general rule was synthesised (Cairns 1984). Unlike reasoning with a small number of “landmark decisions” that characterises the use of precedents in modern legal systems, common law or civilian, the interest in these historical codification attempts was in general patterns that could be distilled from the observation of larger numbers of more mundane decisions, a type of reasoning that more unequivocally follows the pattern of inductive inferences. This insight can be used to analyse the change in legal attitudes over time, by drawing inductive inferences from large case collections, not with a view to establish any specific legal rules, but to understand better how legal systems as a whole evolve (Rissland and Friedman1995).

Hunter (1998, 377) too emphasises the difference between applying a single precedent to solve a specific case under consideration from the task to establish what, at any given point in time, the law is. While the type of analogous reasoning that characterises the first reasoning task is for him indeed categorically different from induction, the inferences through which judges “consolidate” an often heterogeneous set of past decisions in order to determine what the law is fall clearly into the field of inductive rule generalisation.

Hunter writes:

> The important distinction to make is between inductive inference used to understand the current state of the law and inductive inference used to decide a single case. [...] Induction is involved only to the point where a judge or lawyer derives general legal rules or principles from available precedents. These general rules will give a basic outline of what the law is, but they cannot determine the outcome in any given case.

Interestingly, for Hunter abduction plays the role of a necessary complement to induction, thus understood. We have seen in arguments (3a) and (b) above that abduction is closely linked to the discovery of explanations. In abductive reasoning, we make an observation, and then ask what general rule we should expect to hold that explains this observation. “Giving reasons” is also at the core of adjudication, and legal systems typically impose a legal duty on the judge to justify their decision. Hunter argues that while enumerative induction can tell us what the law at any given point in time *is*, this alone is insufficient to discard the burden to justify the decision. A mere enumeration of past cases is on its own not sufficient to discard this burden. Neither is a simple generalisation that transforms these individual instances inductively into a general rule. Rather, we need also a reason *why* the cases that we use for our inductive basis were decided the way they were, the rationale or principle that underlies them all. This rationale or principle is what an abductive inference allows us to hypothesise. A court judgement then goes through several stages of reasoning: *inductively*, the judge determines what the law (in general) is, what the governing rules are. Abductively, a rationale for this rule is then derived, and rule and rational together are then applied to the case under decision in a deductive manner.

Reasoning under uncertainty in law happens not only when, during a trial in court, a unique past event is reconstructed from whatever traces of the event remain.

It happens on a much wider scale in legal practice outside the formal court setting, for example when lawyers advise their clients on the possible outcome of their case, drawing on their own experience with the specific court. The uncertainty in the advice to a client comes from planning ahead for an open future. A client of a contract lawyer might learn not just what the law says, but from their experience that whenever a certain clause was put into a contract, this later resulted in protruded and expensive litigation, and hence the insertion of this clause is inadvisable. Someone accused of burglary might be told by their lawyer: "Of course we can plead that the police planted the evidence on you, but the last three clients of mine who tried this strategy were all convicted. The judge who will preside over your case has never sided with the defendant against the police in the past. It is unlikely they will do it again. This is probably because of their conservative upbringing and because his relatives serve in the police force." In this case, the lawyer uses his experience to make an inductive generalisation: this judge has never behaved in a certain way in the past, and therefore, there will be no difference in your case. Here, we find the same assumption in the continuity of nature that for Hume was the underlying (and problematic) justification for inductive inferences (Boyle 2012). We also see here another example of abductive reasoning. Contrary to the argument above, the mere fact that the judge has decided in a specific way in the past may not be enough as an argument to predict his future actions. It might be that in all the past cases, one and the same officer gave testimony, and that he trusted this specific officer, not the police in general. In this case, the trial in question may well yield a different result, if a different officer is the crown's witness. But once we move beyond a mere list of past events and explain or understand his past actions as the expression of an underlying reason, we can make a more confident prognosis. (Here: his political conviction and family relation). This focus on actual advice in everyday legal practice, centred around the attempt to predict how a specific court or judge will decide, is particularly closely associated with the American legal realist movement. The movement's shift away from the abstract meaning of laws to the empirical reality of the courts, and its understanding of the role of lawyers to predict the behaviour of judges makes its approach a particularly suitable field for thinking about inductive reasoning in law. This type of prediction is particularly relevant when judges are awarded significant discretion, so that a mere knowledge of the law may be insufficient to predict their behaviours with sufficient detail. This was recognised in particular by Zeleznikow, whose approach to legal artificial intelligence uses rule induction to model decision-making in discretionary domains. In this approach, large numbers of "mundane" cases, cases that do not set precedents, are data mined inductively to discover underlying patterns (Zeleznikow 2000).

3 A Primer in Probability Theory in Law

We can briefly recap some of the key ideas of the previous sections. Inductive inferences are non-monotonic, non-truth preserving inferences that allow the reasoner to "add" new knowledge to their beliefs that, in a significant way, "goes beyond" the information, evidence E, that was contained in the premisses. In all of the examples we looked at, E provided some *degree of support* for the conclusion. The degree of support is provided by the ratio of the probability of the evidence if the conclusion is true to the probability of the evidence if the conclusion is false. While the truth of the premisses does not guarantee the truth of the conclusion, the premisses provide a *degree of support for* the truth of the conclusion.

We have seen that this type of inference occurs frequently in legal reasoning, both in reasoning about facts and about law. The next obvious question then is:

- How can we distinguish a convincing from an unconvincing (or maybe even a valid from an invalid) inductive inference? An obvious legal application is for instance post-conviction scrutiny through an appeal court that analyses the reasoning of the court of first instance.
- As we saw in the examples above, it is often possible to construct valid inductive arguments for opposing conclusions, can we in these situations say that one argument is better (more convincing) than the other? Similarly, as we discussed, law often takes place in an adversarial setting, where both sides will present contradictory accounts. How can we decide between them?
- Notice that the "strength" of the support does not in itself provide a probability for the truth of the conclusion. Prior probabilities (these can be everything from scientifically established "*base artes*" to our general knowledge of the world—in the worst case, the prejudices and assumptions of the jurors) for the truth of the conclusion and for some mutually exclusive alternative are also needed. The ratio of these prior probabilities and the strength of the support together provide posterior odds for the truth of the conclusion. It is these posterior odds that we are interested in when evaluating the strength of the prosecutor's or the defence's case. Do these posterior odds need to meet a certain pre-defined threshold? In legal settings, this threshold is often mandated through law. For instance, there is a requirement to prove the guilt of the accused "beyond reasonable doubt" in a criminal case, and, by contrast, to find for the plaintiff or defendant in civil litigation only "on the balance of probabilities."

To answer this question, it is helpful to look at the paradigmatic example of induction, generalising from a finite series of observations to a universal conclusion, and change a bit the way we write them down. This will later allow us in a particularly natural way to build the bridge to probabilistic reasoning.

Example 1 Every pill in a random sample of 100 pills from a confiscated consignment of a large number (much greater than 100) of suspected pills tests positive for LSD. This strongly supports the hypothesis that all pills in the consignment were LSD.

Example 2 Registered users (62%) in a random sample of 400 registered users of a peer-to-peer platform have violated copyright law as regards downloads at some point in time. It can then be said with 95% confidence that between 57 and 67% of all registered users have violated copyright law.

Examples of the first type often occur in evidential reasoning, especially when the legal punishment is linked to quantitative aspects of the crime. If the law for instance has a mandatory 10-year term for possession of more than 10,000 LSD tablets, how many tablets should be tested out of a confiscated batch of 20,000 tablets to determine if the accused was above or below this threshold? (see, e.g., Aitken 1999). In the second case, the amount of damages that the platform owner may have to pay will often depend on an estimation of the economic harm that the copyright holders suffered. In both situations, it is often not feasible to test the entire consignment, either because of costs or because there is only limited evidence to analyse.

Even though the way we have now written the induction may look a bit different, it is still a version of *induction by enumeration* of cases that we discussed above. We can now abstract from these examples a general form for such inferences:

- *Premiss*: In a random sample S consisting of n members of population B, the proportion of members that have attribute A is x.
- Therefore, with a degree of support of pre-specified size p it can be concluded that:
- *Conclusion*: The proportion of all members of B with degree of support p that have attribute A is between $x - q$ and $x + q$.
- We also call q the *margin of error* of x for degree of support p.

Bayesian approaches to inductive reasoning can provide us therefore with the tools that we need to answer the questions lawyers are interested in—what are the respective strengths of the various lines of argument, what support for propositions is provided by the evidence, which line of argument is sound, which line of investigation should be followed. However, the issue has been described in rather abstract terms. We conclude this chapter by looking in more detail at one specific application of inductive probabilistic reasoning in law, the evaluation of DNA in a criminal trial, to illustrate how exactly the abstract ideas may be put into practice.

3.1 Probabilistic Conceptions of Evidential Value

The admission of evidence describing the relative frequency of a DNA profile has encouraged the courts to be somewhat more open to the admission of probabilistic evidence in general than perhaps they were before the advent of DNA profiling. However, there is still much confusion surrounding the interpretation of evidence to which a measure of uncertainty is attached in explicitly probabilistic terms. Consideration in detail of one specific example helps us to highlight the legal issues together with the issues regarding a theory of legal argumentation.

A small probability for finding incriminating evidence on an innocent person does not imply a large probability of guilt for a person on whom the evidence is found. This seemingly innocuous statement has been the source of much confusion in the interpretation of evidence in which probabilities have been mentioned. John Darroch,

Table 1 Data on the frequency of trace evidence amongst Scots and other males

Trace evidence	Guilt	Innocence	Total
Present	1	199	200
Absent	0	19,800	19,800
Total	1	19,999	20,000

an Australian statistician, used an example similar to this one in his evidence to an Australian Royal Commission (Darroch 1985).

Consider again our example from the very beginning, the murder–robbery in an apartment block. Trace evidence has been found at the crime scene which indicates that there was one, and only one, perpetrator. From a witness who overheard the perpetrator speak, we know that he was likely to be Scottish. There are 200 immigrants from Scotland (and no other Scots) in the town and 19,800 other men capable of having committed the crime, the nature of the crime being such that only a man could have committed it. Now, there is obviously room here for a debate as to the number of people in the population to which the criminal may be presumed to belong, i.e. the *relevant population* for our purposes. It is a matter of contention whether such a population can be defined and, if so, whether its size may be determined (Aitken and Taroni 2004).

In the example, suspects may be identified by means other than consideration of the trace evidence. Any individual suspect must be Scottish, otherwise he would have been excluded from the investigation earlier, because of the eyewitness. Trace evidence is found in his environment (on his clothes, in his car, in his home, etc.) which in some sense matches the trace evidence at the crime scene, but which is trace evidence that is also associated with all other Scots in the town and no one else. This numerical information can be represented in a table with two rows and two columns, known as a two-by-two (or 2 × 2) table, as shown in Table 1.

Table 1 illustrates why a small probability for finding incriminating evidence on an innocent person does not imply a large probability of guilt for a person on whom the evidence is found. Consider the "Innocence" column, containing in total 19,999 males. Of these, 199 have trace evidence present in their environment. The ratio of these two numbers is 199/19,999, which approximately equals 1/100 or 0.01. This small number is an estimate of the probability of finding the trace evidence on an innocent person. To reiterate: *there is this small probability of finding the evidence* on an innocent *person.* Now consider the probability of *guilt* of a person for whom trace evidence is found in his environment. There are 200 such people, as shown in the row labelled "Present." Yet *ex hypothesi* one, and only one, of these individuals is guilty. The ratio of these two numbers (1/200, or 0.005) estimates the probability of guilt of a person on whom the evidence is found. Thus, we can now clearly see that the small probability (0.01) of finding the evidence on an innocent person is not equivalent to a large probability that a person on whom the evidence is found is actually guilty. In other words, the probability 199/19,999 is not the probability of innocence for a person on whom the evidence is found, but the probability that traces evidence will be detected on an innocent person.

Conversely, the complement of this fraction, 1-199/19,999 or 19,800/19,999, is not the probability of guilt for a person on whom the trace evidence is found. To the contrary, given that there are 199 “false positives” in the total sample of 20,000 males and only one truly guilty individual, it is very likely (199/200) that any person chosen at random with trace evidence in their environment *is innocent.*

The crucial consideration is to distinguish correctly between (1) the event about whose outcome one is uncertain and (2) the known—or presumed known—event on which one is conditioning to obtain an assessment of probability. Suppose that the event on which one is conditioning is the innocence of the person of interest. We might then say: assuming that this individual is innocent, what is the probability of trace evidence being found in his environment? There are 19,999 innocent people in the relevant sample, as depicted in Table 1’s “Innocence” column. The event, the outcome of which is uncertain, is the presence or absence of the evidence. There are 199 people, known by assumption to be innocent, on whom the evidence is present. The probability of the event of interest (presence of the evidence on an innocent person) is therefore 199/19,999.

Now consider, again, the probability of guilt of a person on whom the evidence is found. The event on which one is conditioning is the discovery of the evidence; that is, we are now assuming that trace evidence has been found on a particular individual. We can see from Table 1’s “Present” row that there are 200 people on whom the evidence will be discovered. The event which is uncertain is the guilt or innocence of the suspect. There is one, and only one, person, known by assumption to have the evidence on him, who is guilty. The probability of the event of interest (guilt of a suspect with trace evidence in his environment) is then 1/200. The complementary probability, the probability of innocence, given discovery of the evidence, is 1-1/200, or 199/200. The fallacy of equating (1) the probability of the presence of the evidence, assuming innocence, to (2) the probability of innocence, assuming the presence of the evidence, is known to statisticians as the fallacy of the transposed conditional. In legal circles, it is more popularly known as “the prosecutor’s fallacy.” Use of the prosecutor’s fallacy was denounced by the Court of Appeal in *Adams,* but the lesson has apparently not always been properly understood or taken to heart. A report produced by the Nuffield Council on Bioethics (2007, para 5.20) states that “[t]he prosecutor’s fallacy has bedevilled the use of DNA evidence in courts.” During the course of its investigations, the Nuffield Council made the startling discovery that one accredited forensic laboratory was routinely committing the prosecutor’s fallacy in its written reports adduced in evidence in criminal trials. As the Council remarks (ibid., para 5.30–5.32), if forensic scientists can make such errors, it is hardly surprising that lawyers, judges and arguably also jurors are susceptible to making them as well.

The crucial distinction between different forms of conditioning is further illustrated by another memorable example. The statement “If I am a monkey, then I have two arms and two legs” is true. However, the statement “If I have two arms and two legs, then I am a monkey” is not always true! There is a direct analogy to the interpretation of DNA profiles in a forensic context. The statement “If I am guilty, then my DNA profile will match that of the DNA profile at the crime scene" (let us stipulate) is true. The statement "If my DNA profile matches that of the crime scene DNA profile, then I am guilty" is not always true. This is not a question of innocent contamination or deliberate evidence tampering, etc., possibilities which can be set aside for present purposes. The point is that DNA evidence is inherently probabilistic, such that the simple fact of a match does not exclude even a strong likelihood of innocence—as Table 1 demonstrated.

3.2 The Language of Probability

The examples presented in the last section made use of terminology which statisticians routinely employ when discussing probability. This section will introduce some of the basic concepts of probability and in doing so also explore their legal applications.

As a starting point, we suppose, simplistically, that before any evidence is heard "innocent until proven guilty" means that every person in the relevant population is equally likely to be guilty. If the relevant population were taken to be the population of the whole world, it is fairly straightforward to think of evidence that will eliminate most of the people in the world from serious consideration as potential suspects, thus reducing the size of the relevant population to a more manageable size.

3.2.1 Proposition

A proposition is taken to be the hypothesis put forward by one side in adversarial trial proceedings. In criminal cases, there are prosecution propositions and defence propositions. Examples of the prosecution proposition include "the defendant is guilty," "the defendant was at the scene of the crime," "the DNA at the scene of the crime is that of the defendant." The defence does not necessarily need to have a proposition, since a blanket denial of the charges puts the prosecution to proof. Where the defence chooses to argue an affirmative case, however, examples of the defence proposition could include "the defendant is innocent," "the defendant was not at the scene of the crime," "the DNA at the scene of the crime is not that of the defendant." These are all complements of the corresponding prosecution proposition, but the defence proposition need not be the complement of the prosecution proposition in every case. Examples include "my brother committed the crime" and "I acted in self-defence." If the defence does not put forward an explicit proposition, it may not be possible to evaluate the evidence as there is no feasible alternative. If this is the situation, then the scientist can assign a probability to the evidence but comment that evaluation is not possible since only one proposition has been considered.

In recent years, an important improvement in the way in which probabilistic arguments are presented and analysed in court was a distinction in "levels of propositions." As lawyers, we are typically interested most in the "ultimate probandum," the question whether the accused is guilty or not. This, however, is rarely ever a purely factual question, let alone one that can be decided by forensic science alone, as questions of “reasonableness” in a self-defence situation may play a crucial role. The questions asked of a forensic scientist by contrast are more likely to be of the form: Does the saliva that was found at the crime scene come from the suspect? Several further inferences are needed to connect such a “source level” proposition to the ultimate target, the offence level proposition. These inferences in turn will often be “warranted” or supported by inductive generalisations about the way we think the world works—sometimes called “scripts” or “stories.” For instance, the expert witness might testify that the saliva that was recovered from the clothing of the victim provides a certain level of support for the proposition that it came from the accused in a racially aggravated assault case in contrast to coming from an individual unrelated to the accused. At this point, our general world knowledge tells us that a common way of transferring saliva is by a certain *activity*, that is spitting at someone. From this activity, we then reason again abductively to the ultimate probandum: we know from experience (inductively) that in our culture, spitting is an expression of ultimate disdain, which then “anchors” the emerging narrative according to which the suspect spitted at the victim due to his hatred of his ethnicity. We can distinguish (at least) four basic levels of proposition:

(i) offence level propositions;
(ii) activity level propositions;
(iii) source level propositions; and
(iv) sub-source level propositions.

This taxonomy combines a mixture of ordinary linguistic usage and more specialist, technical terminology. The establishment of the truth of an offence level proposition (or in civil litigation, claim level proposition) is the aim of the trial; it is the determination of facts that directly trigger legal consequences. In their simplest form, offence level propositions are the antecedents of legal norms, in our example a conditional norm that says that if someone commits racially aggravated assault, then they are liable for imprisonment.

Activity level propositions address whether the suspect (or some other person of interest) performed a relevant action. In criminal law, the action in question will often be the actus reus (conduct elements) of an offence. Thus, we might ask whether the presence of the victim’s blood on the suspect’s clothes is probative evidence that the suspect assaulted the victim. Even if the blood recovered from the suspect’s clothes “matches” (i.e. is indistinguishable from) the blood of the victim, it does not necessarily follow that the suspect performed the relevant action—here, assaulting the victim. He might equally have been a first aider who tried to resuscitate the victim, transferring his blood in the process.

Source level propositions are addressed to the source of particular physical evidence. In relation to biological evidence, source level propositions address whether the suspect (or other relevant person) is the source of an identifiable body fluid such as blood, semen, saliva or hair, the latter also being biological material from which

DNA may be extracted. Thus, we might ask whether the blood found on the hilt of the knife might have been donated by the suspect, the victim, or some other person.

Finally, sub-source level propositions consider just the DNA in isolation, without attributing it to a particular body fluid. DNA extracted from a crime stain does not necessarily share a common donor with the fluid containing it—as is obvious in cases of mixed profiles, implying two or more donors, derived from a single bloodstain, semen sample, or other bodily fluid.

The conceptual distinction between different layers of propositions allows us to place probabilistic reasoning in the wider context of theories about evidential reasoning in law. We used the term "ultimate probandum" as an alternative expression for "offence level proposition." This echoes the usage by the early twentieth century evidence scholar John Henry Wigmore, whose "argument charts" gained renewed interest in the new evidence scholarship movement, and which have been referred to elsewhere (see, e.g., Twining 1985) in this book. Wigmore allowed potentially many more levels of propositions, without assigning special categories to any but the top level proposition. A particularly detailed analysis on how Wigmore charts and probabilistic analysis can work together to rationally reconstruct arguments about evidence is the analysis of the Sacco and Vanzetti trial by Kadane and Schum (2011).

Second, in our reconstruction the transition between different levels of propositions is in turn supported by certain experience based and hence inductive generalisations or "stories." This connects our analysis with the highly influential concept of "anchored narrative (Wagenaar et al.1993). In their analysis, trials can be understood as a conflict between narratives or stories told by the two opposing sides (our prosecutor and defence "theories") that are in various ways "anchored" in reality by evidence. In the case of the racially aggravated assault, the prosecution might tell a "story" about a typical conflict between football hooligans whose teams are divided by sectarian lines that started when one team spit at the other. The reported match of a DNA profile from the saliva and a DNA profile from the accused is a necessary, but in itself insufficient, anchor of this story or theory through evidence. In addition however further inductive generalisations are needed that anchor the transition between different levels of propositions, for instance the experience the jurors have about typical forms of aggression between football fans. A particularly explicit and formally rigorous combination of story-based and probabilistic accounts of evidential reasoning has been developed by Bex (2013) and Verheij et al. (2016).

Finally, our analysis also allows us to briefly point out the contested nature of formal probabilities in evidence law. This chapter tried to show the continuum between informal, everyday induction and mathematical, rigorous probabilities. Induction, so we have argued, is at the heart of our attempts to make sense of our environment, and it is ubiquitous and essential for all forms of reasoning about the world. Some, but not all forms of evidence, allow the assignation of precise numbers to these probabilities, but even where this is impossible, the Bayesian approach allows us to distinguish correct from incorrect inductive reasoning. Some legal systems have been traditionally hostile towards the use of numerical evaluation of the probative weight of evidence by expert witnesses, not (necessarily) because of concerns about the correctness of this type of inference, but because such use was seen as an illegitimate intrusion into the territory of the juror. In this account, the task of the expert is to give “only the facts,” while it is the task of the juror to evaluate how convincing or unconvincing they are. By assigning precise probabilities to the evidence, experts could be seen as taking on a role that for sound political and societal reasons, we reserve to the juror (so, e.g., the early yet still highly influential paper by Tribe 1971 written in response to Finkelstein and Fairley (1970). For critical replies, see, e.g., Finkelstein and Fairley 1971, or more recently Kaye 1986). What we have hoped to show is that, not only is the separation of “facts” and “weight” for many forms of scientific evidence artificial, the fear of “guilt by numbers” is unfounded; this type of probabilistic assessment is encountered typically in the “lower level propositions.” Such an encounter leaves jurors ample space in which to exercise their role when it comes the degree to which knowledge of these propositions assist them in their determination of the ultimate issue.

3.2.2 Relevant Population

While probabilistic evaluation of evidence is essential for reasoning about facts, undoubtedly it can be a source of frequent mistakes (see, e.g., Schneps and Colmez 2013). We continue our introduction into the basic vocabulary of probability theory for legal argumentation with some key concepts that, if misapplied, can be a source of fallacious reasoning.

The *relevant population* is that population of individuals to which the criminal, as yet unknown, belongs. It is determined from the circumstances of the crime. This population may be used to help determine—or, more strictly speaking, *estimate* the probability of particular evidence, for example DNA frequencies in the population (e.g. Caucasian, Hispanic or Negroid though sub-populations of these major groupings may also be relevant). The relevant population is partly defined by the defence proposition, as an example proposed by Robertson and Vignaux (1995, 36–37) demonstrates. This real case loosely inspired the Scottish burglar from our fictitious example above. An English tourist was murdered in Hamilton, New Zealand, in 1992. A man of Maori appearance was seen running away from the scene and subsequently washing himself in a nearby river. Blood which did not belong to the victim was found at the scene and analysed to produce a DNA profile. The frequency of DNA profiles is known to vary between ethnic groups. A Maori male was subsequently arrested and identified by the eyewitness as the person seen running away from the scene. In this example, the prosecution proposition is, obviously, that this individual was the perpetrator of the murder. The defence might have argued that the accused was indeed the person seen running away from the crime scene, but that he was not the murderer, but trying to get help quickly. If this proposition were accepted, and the murderer was not, in fact, the man seen running away from the scene, then it turns out that we have no information about the murderer. In particular, we have no information about the ethnicity of the murderer. The conditioning event for the assessment of the DNA profile in this case should therefore be that the murderer is a person of unknown ethnic origin. The probability of the DNA profile should be determined from consideration of the population of New Zealand as a whole (assuming for the sake of argument that it was a New Zealander who committed the crime). Alternatively, the defence might argue that the accused was *not* the person seen running away from the scene, and that the eyewitness's identification of the defendant as that man was mistaken. Were this proposition accepted, then the accused is innocent but there is still reason to believe that the murderer was a man of Maori origin. The conditioning event for the assessment of the DNA profile in this case is that the murderer is a person of Maori origin. The probability of the DNA profile should be determined from consideration of the Maori population of New Zealand, not from the entire population containing Maoris and non-Maoris. The "relevant population" is thus determined, not only by the nature of the charge and the evidence adduced to prove it—as reflected in the prosecution proposition—but also by the arguments and evidence advanced by the defence.

3.2.3 Prior Odds

The probability that a person chosen at random from the relevant population is guilty can be calculated, in the absence of any other information, by dividing 1 by the number of people in the relevant population. Thus, if all we know is that there are 1,000 individuals in the relevant population, of whom one and only one is guilty, the probability of any individual chosen at random being the guilty individual is 1/1000. This may be taken as a numerical representation of the belief that the person chosen at random is just as likely, and no more likely, to be guilty as anyone else similarly chosen at random from the relevant population. This is an obvious simplification of reality, but it nonetheless supplies a useful working assumption, sometimes designated the "prior probability" (that is, prior to considering the impact of any other evidence).

The complement of the (prior) probability of guilt is the (prior) probability that a person chosen at random from the relevant population is innocent. This probability may be taken to be the number of people in the population minus 1 (to account for the guilty person), divided by the total number of people in the relevant population. Where there are 1,000 individuals in the relevant population, the prior probability of innocence is (1,000-1)/1,000 = 999/1,000.

Referring back to Table 1, the prior probability of guilt for a person chosen at random from the relevant population would be 1/20,000. The prior probability of innocence for a person chosen at random from the relevant population would be (20,000-1)/20,000 = 19,999/20,000. Now, the ratio of these two prior probabilities is 1 to 19,999 (i.e. 1:19,999), which can also be written as 1/19,999 (because 1/20,000 divided by 19,999/20,000 = 1/19,999). This result is known as the "prior odds" *in favour of* guilt. Notice that the prior odds are very small, much less than 1 but fractionally larger than 1/20,000. The reciprocal of the prior odds is 19,999 (because 19,999/1 = 19,999), which can be read as 19,999 to 1 *against* guilt. Odds of 1 (sometimes expressed as 50-50) are equivalent to a prior probability of 0.5 for each of the two relevant events (here, guilt and innocence) since 0.5/0.5 = 1. Odds in favour of guilt greater than 1 arise when the prior probability of guilt is greater than the prior probability of innocence. Whenever, conversely, the prior probability of innocence is greater than the prior probability of guilt (which is the assumption at the start of all criminal trials), the prior odds will be some fraction (usually much) less than 1. As odds are ratios of two probabilities, they are never negative and take values between zero and infinity and only equal zero when the numerator is equal to zero, i.e. when there are no individuals of interest in the relevant population. This is equivalent to saying, in the forensic context, that the perpetrator is not within the relevant population, such that the probability of guilt for any individual selected at random is zero. When the denominator equals zero (no chance of innocence), the corresponding odds are infinite (guilty by definition).

3.2.4 Posterior Odds

The probability of guilt for a person chosen at random from the population on whom the evidence has been found is taken to be 1 divided by the number of people in that population. The population of individuals on whom the evidence is found is always a subset of the relevant population used in the determination of the prior odds. In other words, the members of the population on whom the evidence has been found constitute a subset of the initial relevant population. This subset may be denoted the *posterior population.* By extension, a probability determined from this population is known as a *posterior probability.* It is determined after the evidence in question has been heard (i.e. posterior to the consideration of that evidence).

It is expected that the size of the posterior population will be very much smaller than that of the original population used in the determination of prior odds. The complement of the posterior probability of guilt (for a person on whom the evidence has been found) is the probability that a person chosen at random from the population on whom the evidence has been found is nonetheless innocent. This complement is taken to be the number of people in the posterior population minus 1 (to account for the guilty person), divided by the number of people in the posterior population. In Table 1, the probability of guilt for a person chosen at random from the population on whom the evidence has been found would be 1/200. The probability that a person chosen at random from this population is innocent would then be (200-1)/200 = 199/200. The ratio of these two probabilities is 1: 199, or 1/199. This number represents the *posterior odds* in favour of guilt. Conversely, its reciprocal is 199, which can be expressed in words as "the posterior odds are 199 to 1 against guilt." Notice that this analysis assumes that all 200 people on whom the evidence is found are equally likely to be guilty, without (yet) having taken into account any other evidence that might bear on the issues in the case. This is a very big assumption, which may or may not be warranted in the instant case. Reliance should only be placed upon it with appropriate circumspection.

3.2.5 False Negatives and False Positives

Theoretical probabilities are axiomatic, true by definition. In the empirical world, however, allowance has to be made for errors of various kinds. Two particularly significant kinds of error, which we must mention in order to set aside, are known as "false negatives" and "false positives." These two types of error can be illustrated by considering the results of comparisons between DNA crime-stain samples and comparison samples taken from known suspects. In some circumstances, a negative (non-match) result may be reported where the two samples have a common source, and hence should provide a positive (match) result. This is a "false negative": the result reported is negative, but that result is false. A "false-positive" report, conversely, occurs when a positive (match) result is reported when the two samples in reality have different sources, and therefore should have been reported as a non-match. False-negative and false-positive reports can arise for a host of reasons (the details of which need not concern us here) including contamination of samples, laboratory testing error and misinterpretation of test results.

Analogous terms are employed in medical diagnosis. A false-negative outcome of a test for the presence of a particular disease would be one in which the patient does, in fact, have the disease but the test appears to rule it out. A false-positive outcome of a test for the presence of a particular disease would be one in which the patient is actually disease-free but the test indicates they do have the disease. The consequences of such errors are patently potentially very serious. Errors in medical diagnosis can lead to inappropriate treatment, or to vital treatment being withheld until it is too late to intervene successfully. In a forensic context, a false-negative report may result in a guilty suspect being excluded from an ongoing investigation, while a false-positive report could potentially implicate an innocent suspect and precipitate a serious miscarriage of justice.

3.2.6 Probability of the Evidence if the Person is Guilty (or is Innocent)

For the purposes of undertaking probabilistic analyses for DNA profiles with a stain from a single source, it is conventional to assume that false negatives cannot occur, since the profile is a discrete entity rather than a measurement with which random error is associated. On this assumption, if an individual is guilty, then the evidence found in that individual's environment will be certain to match that found at the crime scene; for example, it is certain that a guilty suspect's DNA will match the crime-stain sample. The probability of an event which is certain is 1. Thus, assuming no false negatives, the probability of the evidence (e.g. a DNA match) if the person is guilty is 1. All our calculations here are premissed on this assumption.

Referring back to the Scottish immigrant example summarised in Table 1, the assumption of no false negatives is reflected in the fact that trace evidence will definitely be found on the guilty suspect (depicted by "1" in the first cell of the "present" row). There are also 19,999 innocent people in the relevant population of males in the town. Of these, 199 have evidence on them which is linked to the crime scene (as shown in the second cell of the "present" row). Thus, the probability of finding the evidence in the environment of a person who is (nonetheless) innocent, that is the probability of the evidence assuming innocence, equals 199/19,999.

3.2.7 Likelihood Ratio

The ratio of the probability of the evidence if the person is guilty, divided by the probability of the evidence if the person is innocent, is known as the *likelihood ratio.* In the example of the immigrant Scots, the likelihood ratio is calculated by dividing 1 by (199/19,999), which is 19,999/199. This in turn is approximately equal to 20,000/200, or—simply—100. A verbal interpretation of this result is that "the evidence is 100 times more likely if the person is guilty than if he is innocent."

Prior odds (of guilt, for a person chosen at random) = 1/19,999.
Likelihood ratio (of the evidence) = 19,999/199.
Posterior odds (of guilt, for a person chosen at random) = 1/199.
Note that:
1/199 = (19,999/199) × (1/19,999).
or in words:
Posterior odds = Likelihood ratio × Prior odds.

The verbal statement is Bayes' theorem, to which earlier reference has been made. The likelihood ratio is the factor which converts prior odds into posterior odds. More fully, the posterior odds in favour of guilt are equal to the product of (i) a ratio of the probability of the evidence if the suspect is guilty, to the probability of the evidence if the suspect is innocent; and (ii) the prior odds in favour of guilt. The numerical example provides verification of the theoretical result. Consider two propositions, the proposition put forward by the prosecution and the proposition put forward by the defence. In their most simple forms, these two propositions may be, respectively, that the suspect is guilty and that the suspect is innocent. There could just as easily be other pairs of propositions: the suspect was at the scene of the crime and the suspect was not at the scene of the crime, or the DNA of the crime-stain sample came from the suspect and the DNA of the crime-stain sample came from some other person (unrelated to the suspect), for example. The general result given by Bayes' theorem may then be written as:

The posterior odds in favour of the prosecution's proposition are equal to the product of (i) a ratio of the probability of the evidence if the prosecution's proposition is true, to the probability of the evidence if the defence proposition is true; and (ii) the prior odds in favour of the prosecution's proposition.

This result has several interesting implications, some of which have important forensic applications:

(i) A likelihood ratio greater than one means that the posterior odds are greater than the prior odds. Evidence for which the likelihood ratio is greater than one may be said to support the prosecution's proposition.

(ii) A likelihood ratio less than one means that the posterior odds are less than the prior odds. Evidence for which the likelihood ratio is less than one may be said to support the defence proposition.

(iii) A likelihood ratio equal to one means that the posterior odds are equal to the prior odds. Evidence for which the likelihood ratio equals one may be said to be *irrelevant,* both logically and legally, in that the evidence leaves the probability of the truth of either proposition exactly the same as it was before the evidence was taken into account.

Evidence which does not alter the prior odds, either in favour of the prosecution (likelihood ratio greater than 1 or in favour of the defence (likelihood ratio smaller than 1), has no utility in adjudication. It cannot logically assist the fact-finder to arrive at a decision, because the probability of the accused's guilt or innocence is wholly unaffected by that evidence.

(iv) A likelihood ratio is a ratio of two probabilities. It takes the value zero when the probability of the evidence if the prosecution proposition is true equals zero, implying that its complement, the defence proposition, is certainly true. It takes the value infinity when the probability of the evidence if the defence proposition is true equals zero, implying that the prosecution proposition is certainly true. Note that, whereas probabilities take values between 0 and 1, likelihood ratios take values between 0 and infinity. The likelihood ratio is sometimes taken to be a measure of support for the relevant proposition. To facilitate reasoning with probabilities for non-experts, the numerical values are routinely translated into verbal scales. For instance, instead of reporting that the test showed that the likelihood ratio makes it 10–100 times more likely that the fingerprints on the knife come from the same person than from different persons, the expert may report this as saying that the evidence lends "moderate support" for the hypothesis that the two fingerprints originate from the same person. To ensure consistency, some jurisdiction will use published verbal scales, though how to phrase them in a way that prevents misinterpretation by laypeople remains an open research question (see, e.g., Martire et al. 2014). We note here though that verbal scales show the relation between scientific, probabilistic reasoning and the informal inductive inferences with which we started this chapter even better.

(v) There is symmetry about 1 in the values of the likelihood ratio. A value of, say, 1000 means that the posterior odds are greater than the prior odds by a factor of 1000. As a numerical illustration, prior odds of 1/10, for example, would be converted to posterior odds of 1000/10, that is, 100. A value of 1/1000, by contrast, means that the posterior odds are smaller than the prior odds by a factor of 1000: thus, prior odds of 10 would be converted to posterior odds of 10/1000, that is, 1/100.

(vi) A relative frequency determined from a sample from a relevant population may be used to estimate a probability applicable to the population. A sample from the relevant population is sufficient. Thus, frequencies derived from forensic databases of fingerprints, shoe-prints, glass, and DNA samples, etc., can be used for probabilistic inference. It is not necessary to collect samples from every conceivable source or donor. However, there is scope for considerable debate about the relevance of the population from which the sample was drawn.

(vii) In criminal adjudication, the values of the prior odds and the posterior odds are matters for the judge and jury, in accordance with the normal division of labour in forensic fact-finding. The value of the likelihood ratio, however, is a matter for the forensic scientist or other expert witnesses, as it is an assessment of the objective probative value of their evidence. Assessments of prior and posterior odds require subjective opinions which are the responsibility of the fact-finders. The scientist does not need to know values for either the prior or the posterior odds. The likelihood ratio can be calculated on the basis of the assumed truth of the propositions put forward by the prosecution and defence.

(viii) If a value of zero is assigned to the prior probability either of the truth of the prosecution's proposition or the truth of the defence proposition, the corresponding posterior probability will also necessarily be zero, regardless of the value of the likelihood ratio. This is a logical consequence of the arithmetic result that the product of zero and any number is zero. It follows that any potential juror who believed that "innocent until proven guilty" equates to a probability of zero for the truth of the prosecution's proposition should be barred from jury service, because such a person is also logically committed to finding a posterior probability of guilt to be zero, *regardless of the strength of the evidence against the defendant.* As noted above, an alternative interpretation of the dictum "innocent until proven guilty" is to interpret it to mean that the accused is "just as likely to be guilty as anyone else." This interpretation has been challenged on the basis that it is not normally realistic to assume that the accused is just as likely to be guilty, no more and no less, than any other person in the relevant population, which could conceivably be the population of the entire world. The choice of the whole world population may only realistically be the case when nothing is known about the crime except that it has happened. Nothing is known about where or when it happened. Even then, certain subsets of the population can be eliminated very quickly, such as those people under a certain age. It must be understood, however, that what is being advocated is a default interpretation to be adopted prior to the consideration of any detailed information (i.e. evidence) actually bearing on the case. Once evidence is led, for example in relation to the location of the crime, the vast majority of theoretical suspects will be eliminated from any further consideration, and most of those still remaining will have probabilities of guilt considerably lower than that of the defendant.

(ix) The conversion of odds in favour of a proposition to the probability of the proposition can only be made if the two propositions being considered are exhaustive as well as mutually exclusive. For non-exhaustive propositions all that can be be said is that the probability of one proposition is bigger than the other by a factor equal to the odds.

4 Conclusion

We started our investigation by noting that even though inductive and probabilistic inferences are central to our cognition and a ubiquitous aspect of our capacity to reason, they have been for a long time the Cinderella of theories about legal reasoning. We traced this back to the pre-eminence of intellectual traditions that treated them with suspicion either because they were "too scientific" and beholden to a vision of a mechanistic calculation of results that is quintessentially inimical to the value-driven and pragmatic attempt of the justice system to come to equitable solutions *for this individual* case, or conversely, falling short of the ideal of certainty that the imposition of harm on citizens through court decisions requires. It was only recently, with the ever-increasing importance of forensic evidence in the fact-finding stage of the trial, that interest in inductive, abductive and probabilistic reasoning took more central stage in legal-theoretical debate, especially in the wake of the new evidence scholarship movement. Our chapter has tried to show that this neglect is both regrettable and indeed dangerous. It is regrettable because the rejection of inductive inferences by theories of legal argumentation was based more on "guilt by association" than any intrinsic shortcomings or philosophical commitments of this mode of reasoning. It is dangerous because of the increased risk of fallacious reasoning. Rather, given the centrality of induction to our cognition, both formalist and realist theories of legal reasoning would benefit from paying greater attention to the potential of reconstructions of legal reasoning that uses the conceptual tools of induction and probability theory. Our capacity to reason reliably about our environment forms a seamless web where we typically utilise different reasoning strategies in conjunction, and often interchangeably, depending on the way we phrase our research questions. In most complex reasoning tasks, as we have tried to show, inductive, abductive and deductive modes of argumentation complement each other and indeed can often be translated into each other depending on the context. The neglect of inductive and probabilistic reasoning is also dangerous. As the importance of forensic evidence in litigation increases, the often counterintuitive nature of probabilistic reasoning, together with the often hidden and unstated assumptions that it requires—we discussed, e.g., the problem of the independence of factors—all too often leads to fallacious inferences with often devastating consequences for the parties concerned. Recent initiatives such as that by the Royal Statistical Society in the UK to develop minimum standards of reasoning competency for lawyers are to be welcomed. However, just as recognition of the importance of inductive and statistical methods for the administration of justice is gaining ground, new developments, driven by increasingly powerful computational tools, cast a new shadow on this endeavour. "Data science," the analysis of large data sets with tools that include statistics, but may also use heuristics and pattern matching algorithms that go beyond statistical analysis as it is commonly understood (see, e.g., the opinion piece by Wickham 2014), is mooted as an entirely new paradigm to think about all aspects of our lives. This will necessarily also impact on the administration of justice. Some of the debates that, as we have seen, were fought about the role of induction in legal reasoning may well have to be revisited, also because the proprietary nature of these computational tools prevent the type of simple, "pen and paper," analysis of argumentative validity that as we argued is not only possible, but essential. It is essential that the legal process does not just try to give the right result. As a core requirement for the transparent administration of justice, the process has also to justify the result in a public way and to give reasons that can, at least in principle, be checked universally for correctness.

References

Aitken, C.G.G. 1999. Sampling—How big a sample. *Journal of Forensic Sciences* 44: 750–760.

Aitken, C.G.G., and F. Taroni. 2004. *Statistics and the evaluation of evidence for forensic scientists*. Chichester: Wiley.

Aliseda, A. 2004. Logics in scientific discovery. *Foundations of Science* 9: 339–363.

Aliseda, A. 2006. *Abductive reasoning: logical investigations into discovery and explanation*. Heidelberg: Springer.

Amaya, A. 2007. Formal models of coherence and legal epistemology. *Artificial Intelligence and Law* 15 (4): 429–447.

Ashley, K.D. 1992. Case-based reasoning and its implications for legal expert systems. *Artificial Intelligence and Law* 1 (2-3): 113–208.

Bacon, F. 1887. *Novum Organum*, 1st ed, vol. 1620, ed. T. Fowler. Oxford: Clarendon Press.

Baldwin, S.E. 1903. The study of elementary law, the proper beginning of a legal education. *Yale Law Journal* 13: 1–15.

Barnard, G.A. (1958). Thomas Bayes—a biographical note (together with a reprinting of Bayes 1764) *Biometrika*, 45: 293–315.

Bayes, T. 1764. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London for 1763* 53: 370–418.

Bex, F. 2013. Abductive argumentation with stories. In *ICAIL-2013 workshop on formal aspects of evidential inference*.

Boutilier, C. 1996. Abduction to plausible causes: An event-based model of belief update. *Artificial Intelligence* 83: 143–166.

Boyle, D. 2012. The ways of the wise: Hume's rules of causal reasoning. *Hume Studies* 38: 157–182.

Brewka, G. 1991. *Nonmonotonic reasoning: Logical foundations of commonsense*. Cambridge: Cambridge University Press.

Brewka, G., J. Dix, and K. Konolige. 1997. *Nonmonotonic reasoning - An overview*. Stanford University Press: CSLI publications, Redwood City, Calif.

Cairns, J.W. 1984. Institutional writings in Scotland reconsidered. *The Journal of Legal History* 4 (3): 76–117.

Carnap, R. 1952. *The continuum of inductive methods*. Chicago, III: The University of Chicago Press.

Cooper, G.F., and E. Herskovits. 1992. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 9 (4): 309–347.

Dale, A.I. 1982. Bayes or Laplace? An examination of the origin and early application of Bayes' theorem. *Archive for the History of the Exact Sciences* 27: 23–47.

Darboux, J.G., P.E. Appell, and J.H. Poincaré. 1908. Examen critique des divers systèmes graphologiques auxquels a donné lieu le bordereau de (l'affaire Dreyfus). In *L'affaire Dreyfus, La revision du procès de Rennes, Enquête de la chambre criminelle de la Cour de Cassation*, 499–600. Paris: Ligue française des droits de l'homme et du citoyen.

Darroch, J. 1985. Probability and criminal trials. *Newsletter of the Statistical Society of Australia* 30: 1.

Darroch, J. 1987. Probability and criminal trials: Some comments prompted by the Splatt trial and the Royal Commission. *The Professional Statistician* 6: 3.

Daston, L.J. 1981. Mathematics and the Moral Sciences: The Rise and Fall of the Probability of Judgments, 1785-1840. In *Epistemological and Social Problems of the Sciences in the Early Nineteenth Century*, ed. H.N. Jahnke, and M. Otte, 287–309. Heidelberg: Springer.

Dawid, P., D.L. Faigman, and S. Fienberg. 2014. Fitting science into legal contexts: Assessing effects of causes or causes of effects. *Sociological Methods and Research* 43: 359–421.

Dawid, P., M. Musio, and S. Fienberg. 2016. From statistical evidence to evidence of causality. *Bayesian Analysis* 11: 725–752.

Einhorn, H.J., and R.M. Hogarth. 1981. Behavioral decision theory: Processes of judgment and choice. *Journal of Accounting Research* 19 (1): 1–31.

Fienberg, S.E. 2006. When did Bayesian inference become "Bayesian"? *Bayesian Analysis* 1: 1–40.

Finkelstein, M.O., and W.B. Fairley, 1970. A Bayesian approach to identification evidence. *Harvard Law Review* 83: 489–517.

Finkelstein, M.O., and W.B. Fairley. 1971. The continuing debate over mathematics in the law of evidence: A comment on (trial by mathematics). *Harvard Law Review* 8: 1801–1809.

Froeb, L.M., and B.H. Kobayashi. 1996. Naïve, biased, yet Bayesian: Can juries interpret selectively produced evidence? *Journal of Law Economics and Organization* 12 (1): 257–276.

Gillies, D. 2000. *Philosophical Theories of Probability*. London: Routledge.

Goel, V., B. Gold, S. Kapur, and S. Houle. 1997. The seats of reason? An imaging study of deductive and inductive reasoning. *NeuroReport* 8: 1305–1310.

Goldberg, D.E., and J.H. Holland. 1998. Genetic algorithms and machine learning. *Machine Learning* 3 (2): 95–99.

Gordon, T.F. 1988. The importance of nonmonotonicity for legal reasoning. In *Expert systems in law: Impacts on legal theory and computer law*, ed. H. Fiedler, F. Haft, and R. Traunmüller, 111–126. Tübingen: Attempto Verlag.

Halford, G.S., J.D. Bain, M.T. Maybery, and G. Andrews. 1998. Induction of relational schemas: Common processes in reasoning and complex learning. *Cognitive Psychology* 35 (3): 201–245.

Hawthorne, J. 2012. Inductive logic. In *The Stanford Encyclopedia of Philosophy*, ed. E.N. Zalta. http://plato.stanford.edu/archives/win2012/entries/logic-inductive/.

Heit, E. 2000. Properties of inductive reasoning. *Psychonomic Bulletin and Review* 7: 569–592.

Hoeflich, M.H. 1986. Law and geometry: Legal science from Leibniz to Langdell. *American Journal of Legal History* 30: 95–121.

Holland, J.H., K.J. Holyoak, R.E. Nisbett, and P. Thagard. 1998. *Induction: Processes of inference, learning, and discovery*. Cambridge, Mass: MIT Press.

Howson, C. 2000. *Hume's problem: Induction and the justification of belief*. Oxford: Oxford University Press.

Hoyningen-Huene, P. 2006. Context of discovery versus context of justification and Thomas Kuhn. In *Revisiting discovery and justification*, ed. J. Schickore, and F. Steinle, 119–131. Heidelberg: Springer.

Hunter, D. 1998. No wilderness of single instances: Inductive inference in law. *Journal of Legal Education* 48: 365.

Jackson, J.D. 1996. Analysing the new evidence scholarship: Towards a new conception of the law of evidence. *Oxford Journal of Legal Studies* 16: 309.

Jacobs, S. 1991. John Stuart Mill on induction and hypotheses. *Journal of the History of Philosophy* 29 (1): 69–83.

Jaynes, E.T. 2003. *Probability theory: The logic of science*. Cambridge: Cambridge University Press.

Johnson-Laird, P.N. 1993. *Human and machine thinking*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Kadane, J.B., and D.A. Schum. 2011. *A probabilistic analysis of the Sacco and Vanzetti evidence*. Chichester: Wiley.

Kakas, A.C., and F. Riguzzi. 2000. Abductive concept learning. *New Generation Computing* 18 (3): 243–294.

Kaye, D.H. 1986. The admissibility of "probability evidence" in criminal trials—Part I. *Jurimetrics* 26: 343–346.

Kaye, D.H. 2004. On falsification and falsifiability: The first Daubert factor and the philosophy of science. *Jurimetrics* 45: 473.

Keener, W.A. 1894. The inductive method in legal education. *Reports of the American Bar Association* 17: 473–494.

Keppens, J., and B. Schafer. 2006. Knowledge based crime scenario modelling. *Expert Systems with Applications* 2: 203–222.

Lakatos, I. 1970. Falsification and the methodology of scientific research programmes. In *Criticism and the growth of knowledge*, ed. I. Lakatos, and A. Musgrave. Cambridge: Cambridge University Press.

Landman, J.H. 1927. Anent the case method of studying law. *New York University Law Review* 4: 139–160.

Langdell, C.C. 1871. *A selection of cases on the law of contracts: With references and citations*. Boston, Mass: Little, Brown, and Company.

Langley, P., and H.A. Simon. 1995. Applications of machine learning and rule induction. *Communications of the ACM* 38 (11): 54–64.

Laplace, Marquise de, P.S. 1814. *Essai philosophique sur les probabilités*. Paris: Courcier.

Lempert, R. 1977. Modeling relevance. *Michigan Law Review* 75: 1021–1057.

Lukaszewicz, W. 1990. *Non-monotonic reasoning*. Chichester: Ellis-Horwood.

Maher, P. 1993. *Betting on theories*. Cambridge: Cambridge University Press.

Martire, K.A., R.I. Kemp, M. Sayle, and B.R. Newell. 2014. On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect. *Forensic Science International* 240: 61–68.

McCormick, N. 1987. Why cases have rationes and what these are. In *Precedent in law*, ed. L. Goldstein, 155–182. Oxford: Oxford University Press.

McDermott, D., and J. Doyle. 1980. Non-monotonic logic I. *Artificial Intelligence* 13: 41–72.

McGinnis, J. 2003. Scientific methodologies in medieval Islam. *Journal of the History of Philosophy* 41: 307–327.

Medawar, P.B. 2013. *Induction and intuition in scientific thought*. London: Routledge.

Michalski, R.S. 1980. Pattern recognition as rule-guided inductive inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 4: 349–361.

Michalski, R.S. 1983. A theory and methodology of inductive learning. *Artificial Intelligence* 20: 111–161.

Mill, J.S. 1843. *System of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation in two volumes*, vol. 1. London: John W. Parker.

Milton, J.R. 1987. Induction before Hume. *British Journal for the Philosophy of Science* 38: 49–74.

Mozer, M.C. 1984. *Inductive information retrieval using parallel distributed computation*. No. ICS-8406. California Univ San Diego La Jolla Inst For Cognitive Science.

Murray, J.R. 1982. The role of analogy in legal reasoning. *UCLA Law Review* 29: 833–847.

Neyman, J. 1977. Frequentist probability and frequentist statistics. *Synthese* 36: 97–131.

Nisbett, R.E., D.H. Krantz, C. Jepson, and Z. Kunda. 1983. The use of statistical heuristics in everyday inductive reasoning. *Psychological Review* 90 (4): 339.

Nuffield Council on Bioethics. 2007. *The forensic use of bioinformation: Ethical issues*. https://nuffieldbioethics.org/wp-content/uploads/The-forensic-use-of-bioinformation-ethical-issues.pdf.

Oberhofer, R. 1992. Rechtsanwendung und Auslegung. In *Allgemeiner Teil des bürgerlichen Rechts*, 54–90.

Osherson, D.N. 1990. Category-based induction. *Psychological Review* 97: 185–200.

Pearson, E.S., and Kendall, M.G. (1970) *Studies in the History of Statistics and Probability*. 131–153. Charles Griffin, London.

Peirce, C.S. 1903. *Pragmatism as a principle and method of right thinking: The 1903 Harvard lectures on pragmatism*, ed. P.A. Turrisi. Albany, N.Y.: State University of New York Press.

Peirce, C.S. 1956. The probability of induction. In *The world of mathematics*, vol. 2, ed. J.R. Newman, 1341–1354. New York, N.Y.: Simon and Schuster (1st ed. 1878).

Plott, C. 2000. *Global history of philosophy: The period of scholasticism*. Delhi: Motilal Banarsidass Publ.

Prakken, H., and G. Sartor. 1997. A dialectical model of assessing conflicting arguments in legal reasoning. In *Logical models of legal argumentation*, ed. H. Prakken, and G. Sartor, 175–211. Heidelberg: Springer.

Rissland, E.L., and Friedman, M.T. 1995. Detecting change in legal concepts. In *Proceedings 5th international conference artificial intelligence and law, Melbourne, Australia, June 30–July 4,* 127–136. New York, N.Y.: ACM Press.

Robertson, B.W., and G.A. Vignaux. 1995. *Interpreting evidence*. Chichester: Wiley.

Schneps, L., and C. Colmez. 2013. *Math on Trial. How numbers get used and abused in the courtroom. New York,* N.Y.: Basic Books.

Shnee, A. 1997. Logical reasoning obviously. *Legal Writing: The Journal of Legal Writing Institute* 3: 105–126.

Stigler, S.M. 1982. Thomas Bayes's Bayesian inference. *Journal of the Royal Statistical Society,* Series A 145: 250–258.

Tenenbaum, J.B., T.L. Griffiths, and C. Kemp. 2006. Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences* 10 (7): 309–318.

Tribe, L.H. 1971. Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review* 84: 1329–1393.

Twining, W. 1984. Taking facts seriously. *Journal of Legal Education* 34: 22.

Twining, W. 1985. *Theories of evidence: Bentham and Wigmore*. Stanford, Cal.: Stanford University Press.

Verheij, B., F. Bex, S.T. Timmer, C.S. Vlek, J.-J. Meyer, S. Renooij, and H. Prakken. 2016. Arguments, scenarios and probabilities: Connections between three normative frameworks for evidential reasoning. *Law, Probability and Risk* 15: 35–70.

Wagenaar, W.A., P.J. van Koppen, and H.F. Crombag. 1993. *Anchored narratives: The psychology of criminal evidence*. New York, N.Y.: St Martin's Press.

Whewell, W. 1840. *The philosophy of the inductive sciences: Founded upon their history*, vol. 1. London: JW Parker.

Wickham, H. 2014. Data science: How is it different to statistics? IMS Bulletin online, http://bulletin.imstat.org/2014/09/data-science-how-is-it-different-to-statistics%E2%80%89/.

Zeleznikow, J. 2000. Building decision support systems in discretionary legal domains. *International Review of Law, Computers and Technology* 14 (3): 341–356.

Zeleznikow, J. 2004. The split-up project: Induction, context and knowledge discovery in law. *Law, Probability and Risk* 3: 147.

Defeasibility in Law

Giovanni Sartor

1 The Idea of Defeasibility

In a very broad sense, the idea of defeasibility may be applied to any process that responds to its normal inputs with certain outcomes (the default results), but which delivers different outcomes when such inputs are augmented with further, exceptional or abnormal elements.

The computer scientist and theorist of complexity John Holland argues that complex systems—such as a cell, an animal, or an ecosystem—can be characterised "in terms of a set of signal-processing rules called classifier rules" (Holland 2012, 28). Each such rule represents a mechanism which "accepts certain signals as inputs (specified by the condition part of the rule) and then processes the signals to produce outgoing signals (the action part of the rule)." He observes that complex systems need to address different situations, requiring different responses, which are triggered by rules having different levels of generality. In many cases, the most efficient way to cover multiple different contingencies consists in constructing "a hierarchy of rules, called a default hierarchy, in which general rules cover the most common situations and more specific rules cover exceptions" (Holland et al. 1989). Thus, a general rule would provide the normal response to a certain input, but more specific rules would override the general rule in exceptional situations, in which a different response is needed. The emergence of default hierarchies may be favoured by evolution, since such hierarchies may contribute to the fitness of the systems using them.

G. Sartor (✉)
Dipartimento di Scienze giuridiche, Università di Bologna, Bologna, Italy
e-mail: giovanni.sartor@gmail.com

G. Sartor
European University Institute, Florence, Italy

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_12

> Default hierarchies offer several advantages to systems that learn or adapt:
>
> – A default hierarchy has many fewer rules than a set of rules in which each rule is designed to respond to a fully specified situation.
>
> – A higher-level rule [...] is easier to discover (because there are fewer alternatives) and it is typically tested more often (because the rule's condition is more frequently satisfied.
>
> – The hierarchy can be developed level by level as experience accumulates. (Holland 2012, 122)

This perspective can be applied to different domains at different levels of abstraction. For instance, at the cellular level rule mechanisms specify the catalytic and anti-catalytic processes that induce and inhibit chemical reactions. At the DNA level, rule mechanisms are provided by genes (and parts of them). Each gene delivers the protein matching the sequence of the gene's bases and it may be regulated by other genes that send signals which under particular conditions repress (turn off) or induce (turn on) the functioning of the gene rule at issue. Animal behaviour is also largely governed by systems of reflex rules defining reactions: to heat or cold; or to the sight, smell, or taste of food; or to the perception of incoming dangers; and so on. Such reflexes can be innate or learned by experience, i.e. by conditioning and reinforcement. They may interact in complex patterns: some reflexes are stronger than others, so that they determine the response in cases of conflict, and some reflexes have an inhibitory impact on others blocking them under particular situations. In humans, reflexes are integrated with deliberative processes and means-end reasoning, but still they govern a large part of human behaviour.

As Holland et al. (1989, 38) argue, not only instinctive reflexes but also mental models can be based on sets of prioritised default rules:

> The rules that constitute a category do not provide a definition of the category. Instead they provide a set of expectations that are taken to be true only so long as they are not contradicted by more specific information. In the absence of additional information these "default" expectations provide the best available sketch of the current situation. Rules and rule clusters can be organized into default hierarchies, that is, hierarchies ordered by default expectations based on subordinate/superordinate relations among concepts. For example, knowing that something is an animal produces certain default expectations about it, but these can be overridden by more specific expectations produced by evidence that the animal is a bird.

In conclusion, a defeasible process can be characterised as a mechanism which responds to its normal inputs with certain default outcomes, but which may fail to respond in this way when the input is accompanied by certain additional exceptional elements.

2 Defeasibility in Reasoning and Nonmonotonic Inference

Though defeasibility also applies to reactive agents, it acquires its fullest meaning in cognitive agents: defeasible cognition consists in achieving certain cognitive states (beliefs, intentions, etc.) when provided with certain normal cognitive inputs (perceptions, beliefs, intentions), but refraining from adopting these states, or abandoning them. when the normal inputs are accompanied by further elements. More specifically, the idea of defeasability takes on a more precise content when referred to reasoning, i.e., to inference or argumentation. A defeasible reasoning process (an inference or argument pattern) responds to typical input premises with certain default conclusions, but fails to deliver those conclusions when the typical input premises are accompanied by further premises, indicating exceptional circumstances.

The most cited example of a default inference concerns Tweety the penguin. Let us assume that we are told that Tweety is a bird. Given this information and knowing that birds usually fly, we would normally conclude that Tweety flies. Assume, however, that we are later told that Tweety is a penguin. Given this additional piece of information and knowing that penguins are birds which do not fly, we should refrain from endorsing the conclusion that Tweety flies. In fact, we now know that he is a special kind of bird, namely a penguin, to which the default rule does not apply.

As this example shows, the addition of premises in a defeasible reasoning may lead to the withdrawal of a conclusion. This aspect of defeasible reasoning is conceptualised through the distinction between monotonic reasoning and nonmonotonic reasoning. In general, we say that an inference method is *monotonic* when it behaves as follows: any conclusion that can be obtained from an initial set of premises can still be obtained whenever the original set is expanded with additional premises. More precisely, all conclusions that are derived through monotonic inferences from a premise set S_1 can also be derived from any larger (more inclusive) premises set S_2 ($S_1 \subseteq S_2$).

Correspondingly, an inference method is *nonmonotonic* when it behaves as follows: a conclusion that can be obtained from an initial set of premises may no longer be obtainable when the original set is expanded with additional premises. More precisely, conclusions that are derived through nonmonotonic inferences from a premise set S_1 may no longer be derivable from a larger (more inclusive) set of premises S_2.

Deduction is monotonic: as long as we accept all premises of a deductive inference, we must continue to accept its conclusion. Therefore, we also say that deductive inference is *conclusive*: as long as we maintain the premises, any additional information will not affect the conclusion.

By contrast, defeasible inferences are nonmonotonic: we may reject the conclusion of a defeasible inference while maintaining all of its premises. This may indeed happen when further premises are provided that substantiate exceptions to the defeasible inference. In defeasible reasoning "if the premises hold, the conclusion also holds tentatively, in the absence of information to the contrary" (Walton 2008, 161). Thus, defeasible inference

> relies on absence of information as well as its presence, often mediated by rules of the general form: given *P*, conclude *Q* unless there is information to the contrary. (Horty 2001, 337)

Defeasible reasoning is not only a matter of practice but also one of rational justification, as stated in the following definition:

> Reasoning is *defeasible* when the corresponding argument is rationally compelling but not deductively valid. The truth of the premises of a good defeasible argument provide support

> for the conclusion, even though it is possible for the premises to be true and the conclusion false. In other words, the relationship of support between premises and conclusion is a tentative one, potentially defeated by additional information. (Koons 2017)

As we shall see in what follows, in many situations we are entitled or justified to derive default conclusions and to maintain those conclusions until we come to appreciate that circumstances obtain under which such conclusions should not be retained.

3 Conclusive and Defeasible Arguments

Different approaches to defeasible (nonmonotonic) reasoning and its formalisation have been developed (see Ginzberg 1987; Horty 2001; Prakken and Vreeswijk 2001). Here, I shall approach defeasible reasoning as argumentation, namely as the derivation of provisionally justified conclusions through the dialectical opposition of competing arguments (on argumentation, see Walton 2013). This is indeed the perspective that better fits the argumentative and dialectical nature of legal reasoning, as it emerges in analysis, advocacy, and decision-making.

A *valid argument* can be said to consist of three elements: a set of premises, a conclusion, and a support relation between premises and conclusion. In a *deductively valid argument,* the premises provide *conclusive* support for the conclusion: if we accept the premises, we must necessarily accept the conclusion. In a *defeasibly valid argument*, the premises only provide *presumptive* support for the conclusion: if we accept the premises, we should also accept the conclusion, but only so long as we do not have prevailing arguments to the contrary. We can extend the notion of an argument to unsupported claims: such a claim can be viewed as argument only consisting in the assertion of a conclusion. The unsupported claim of a proposition will be sufficient to substantiate it, when the truth of the proposition is evident or is anyway agreed upon.

Defaults usually have a general form and consequently have to be mapped onto or instantiated to the specific propositions to which they are applied. For instance, to apply the general default “pet dogs are presumably unaggressive,” i.e. in a conditional form, “if something is a pet dot, then presumably it is nonaggressive,” to Fido, we must specify or “instantiate” the default to the case of Fido, i.e., generate the following specification: “if Fido is a pet dog, then presumably Fido is not aggressive.” This specification, in combination with the premise that Fido is a pet dog, leads us to the presumable conclusion that Fido is not aggressive, through defeasible modus ponens. In the examples that follow, I will omit the specification step, presenting the conclusion as directly resulting from the general default and the specific conditions matching its antecedent. In fact, a general default can be seen as the set of all of its specific instances, which include the one applied to the case at hand.

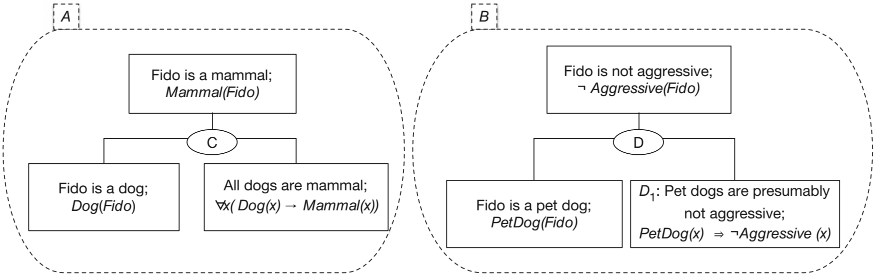


Fig. 1 Conclusive and defeasible arguments

I shall use a diagrammatic representation for arguments, as exemplified below, where boxes include premises or conclusions, and combinations of premises are linked to the conclusion they conjunctively support. In the diagram of Fig. 1, we can see a conclusive (deductive) argument (A) supporting the conclusion that Fido, being a dog, is a mammal and a defeasible argument (B) supporting the conclusion that Fido, being a *pet* dog, is presumably not aggressive. I have represented the premises both in natural language and in the usual formalism of predicate logic and have labelled the connection between premises and conclusion by the letters C and D to distinguish conclusive from defeasible arguments.

Arguments in natural language usually have an *enthymematic* form, meaning that they may omit some of the premises that are needed to support their conclusions. Here, I shall present all arguments in their complete form, that is as including all premises that are sufficient to conclusively or defeasibly establish their conclusion.

In particular, I assume that each defeasible argument includes (a) a set of antecedent conditions, and (b) a defeasible conditional, called a *default*, according to which the (conjunction of the) conditions presumably determine the argument's conclusion. I represent defaults in the form "if P_1 and … and P_n, then presumably Q," in formula $P_1 \wedge \cdots \wedge P_n \Rightarrow Q$, where the arrow ⇒ denotes defeasible conditionality (I will use the arrows $\Rightarrow$, $\rightarrow$, and $\twoheadrightarrow$ to denote defeasible, material, and strict conditional, respectively). Thus, a single-step defeasible argument has the following form:

1. $P_1, \ldots, P_n$(the antecedent conditions), and
2. if P_1 and … and P_n, then presumably Q (the default, in formula: $P_1 \wedge \cdots \wedge P_n \Rightarrow Q$).
   therefore
3. Q.

This inference is called *defeasible modus ponens* to distinguish it from the conclusive modus ponens inference of deductive logic. We can represent a defeasible argument by providing the set of its premises (conditions and default): $\{P_1, \ldots, P_n, P_1 \wedge \cdots \wedge P_n \Rightarrow Q\}$, the conclusion of the argument being the conclusion of the default. Given a defeasible modus ponens inference (argument) $\mathcal{A} = \{P_1, \ldots, P_n, P_1 \wedge \cdots \wedge P_n \Rightarrow Q\}$, I will say that the conjunction of the

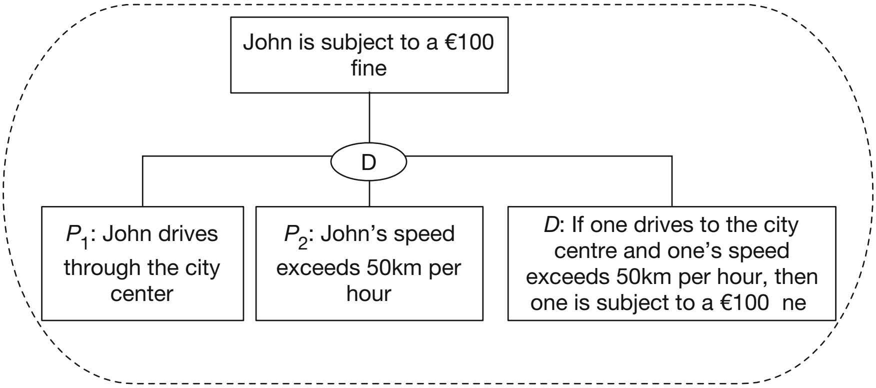

John is subject to a €100 fine
D
P_1: John drives through the city center
P_2: John's speed exceeds 50km per hour
D: If one drives to the city centre and one's speed exceeds 50km per hour, then one is subject to a €100 ne

Fig. 2 Linked argument

$P_1, \ldots, P_n$ conditions is the *reason* for (concluding that) Q and that the default $P_1 \wedge \cdots \wedge P_n \Rightarrow Q$ is the *warrant* for Q. I will also say that Q is *warranted* by that default.

For instance, given argument B in Fig. 1, we can say that the fact that Fido is a pet dog is the reason for concluding that he is not aggressive and that this conclusion is warranted by the default that pet dogs are not aggressive. As example of conjunctive reason, consider the argument in Fig. 2, where the conjunction of the two premises P_1 and P_2 provides the reason for the conclusion warranted by the default D. Note that, I freely use symbols $P_1, \ldots, P_n$ as names for propositions and $D_1, \ldots, D_n$ as names for defaults, whenever needed.

The notion of a defeasible argument can be generalised to cover multistep defeasible arguments, which consist of a set of the arguments providing the conditions of a top default, plus that default. For instance, if $\{P, P \Rightarrow Q\}$ is a defeasible argument, so is also $\{\{P, P \Rightarrow Q\}, Q \Rightarrow R\}$ (for an example of multistep defeasible argument, see Fig. 16, and for a formal definition of the general notion of an argument, possibly including both defeasible and deductive steps, see Prakken 2010, Sect. 3.2).

4 Linked Arguments and Convergent Argument Structures

Besides the distinction between defeasible and conclusive arguments, a second categorisation of arguments is relevant to our purposes, namely the distinction between linked arguments and convergent argument structures (see Walton 2006, 139 ff.; Hitchcock 2017, Chap. 2).

A linked argument is an argument that includes, beside a conditional warrant, more than one premises. None of these premises is sufficient to trigger on its own the conjunctive antecedent of the conditional warrant. Therefore, in isolation, each of them fails to provide any (presumptive or conclusive) support to the conclusion of

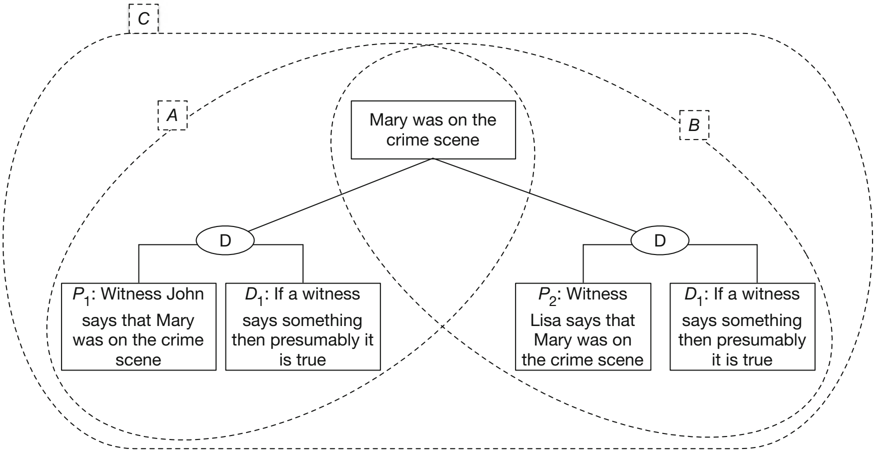

C
A
B
Mary was on the crime scene
D
D
P_1: Witness John says that Mary was on the crime scene
D_1: If a witness says something then presumably it is true
P_2: Witness Lisa says that Mary was on the crime scene
D_1: If a witness says something then presumably it is true

Fig. 3 Convergent factual argument

that warrant. For instance, assume the following premises (P_1) John drives through the city centre, (P_2) his speed exceeds 50 km per hour, and (D) if one drives through the city centre, and his or her speed exceeds 50 km per hour, then one is subject to a 100 € fine. Only the joint combination of premises P_1 and P_2 triggers (presumably) the conclusion that John is subject to a 100 € fine (Q). The resulting argument is depicted in Fig. 2.

A convergent argument structure is a combination of multiple arguments, each leading to the same conclusion. Often, but not always, a convergent argument structure provides a stronger support to the common conclusion of its component arguments than each of these arguments would do in isolation (see Prakken 2005; Bench-Capon and Prakken 2006). In Fig. 3 you can see how two witness testimonies originate separate arguments A and B which merge into the convergent argument C, which provides a stronger support to the common conclusion of A and B.

Figure 4 shows a combination of independent arguments pointing to the same practical conclusion (a conclusion concerning what should be done): being asked the way by driver John, I should not direct him in a wrong direction (tell him that he will get to destination by going to the left), given that my false statement would both be a lie and harm John, impeding him from reaching its destination. Arguments A and B refer to two parallel principles: the duty to be truthful and the duty not to harm others (a foremost requirement of the law according to Justinian's Digest, D 1.1.1).

The distinction between linked arguments and convergent argument structures enables us to distinguish two important concepts: the concept of a contributory condition and the concept of a contributory reason.

A *contributory condition* for a conclusion is a necessary element of a (presumably or conclusively) sufficient condition for that conclusion. This concept applies to each element of a conjunctive warrant supporting that conclusion. Given a warrant "ïf P_1

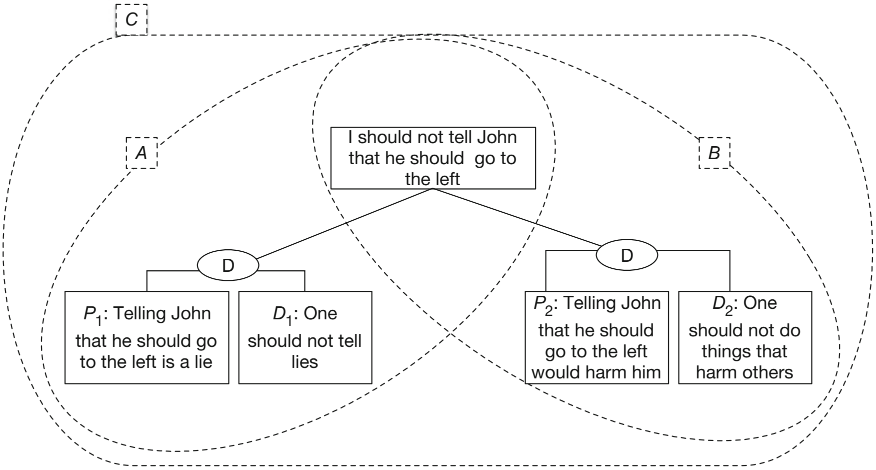

C
A
B
I should not tell John that he should go to the left
D
D
P_1: Telling John that he should go to the left is a lie
D_1: One should not tell lies
P_2: Telling John that he should go to the left would harm him
D_2: One should not do things that harm others

Fig. 4 Convergent practical argument

and …, and P_n then (presumably) Q," each P_1, …, P_n is a contributory condition for that warrant.

A *contributory reason*, relative to a conclusion, is a presumably sufficient condition of that conclusion. This concept only applies the whole antecedent of a warrant supporting that conclusion.

Thus, premises P_1 and P_2 in Fig. 2 are contributory conditions, but fail to qualify as contributory reasons, since neither of them can separately trigger the conclusion of the argument including both of them: both are needed to satisfy the conjunctive antecedent of the argument's warrant. Therefore, neither of them can be properly characterised as a reason for that conclusion. On the other hand, premises P_1 and P_2 in Figs. 3 and 4 do qualify as (contributory) reason for their common conclusion, since each of them (together with the corresponding default warrant) independently triggers that conclusion, besides contributing to provide a stronger joint support to the same conclusion.

In the legal domain, the idea of a contributory reasons applies to the domain of principles, understood as optimisation requirements (Alexy 2002, Chap. 4) or value-norms (Sartor 2013, Section D). The fact that a choice advances a principle (a legal value) is a contributory reason for adopting that choice (of for its constitutional legitimacy). When the same choice advances multiple principles, this originates multiple convergent arguments—the advancement of each principle being a contributory reason—that join to provide a stronger support to that choice. Similarly, the fact that a choice negatively affects the realisation of a principle is a contributory reason for not adopting the choice or against its legitimacy. When multiple principles are negatively affected, this originates multiple convergent arguments against that choice.

The idea of a contributory reason also applies to the antecedent of legal rules. As I shall argue in the following, the antecedent of a legal rule usually only provides a presumably sufficient condition for the conclusion of that rule. For instance, a driver exceeding a speed limit may not be subject to sanction in case that his behaviour is justified by state of necessity (he was transporting a person to the hospital for an emergency). Rule-warranted arguments and principle-warranted arguments, while sharing the same basic logical structure, present some relevant differences. Firstly, rule-warranted arguments may "exclude" (undercut, in our terminology, see Sect. 5), rather than oppose (rebut), certain contrary arguments warranted by principles (if we follow the idea of Raz 1985, also adopted by Hage 2000). Secondly, convergent rule-based argument structures usually do not provide a stronger support to their conclusion than the constituting arguments do. For instance, assume that a person has committed a violation that triggers his or her liability both in contract and in torts. This provides for a converging argument structure for the liability of this person. However, this convergent argument structure arguably does not provide a stronger support to the liability conclusion than the strongest of the two separate arguments for that conclusion. I have preferred to speak of a convergent argument structure, rather than or a convergent argument, to denote the combination of arguments leading to the same conclusions, to maintain the concept of an argument I introduced above, that requires a single warrant linking premises and conclusion.

5 Attacks Against Arguments: Rebutting and Undercutting

An argument can be attacked in any of three ways: by attacking its premises, by attacking its conclusions, or by attacking the support relation between premises and conclusions. Conclusive arguments can only be attacked by challenging their premises, since, if the premises are accepted, then the conclusion must also be accepted. So, for instance, if we accept that Fido is a dog and that all dogs are mammals, we must also accept that Fido is a mammal (as soon as we are aware of the logical connection between premises and conclusion). In fact, it may also be possible to attack the conclusion of the argument—i.e. to deny that Fido, who is a robot in the likeness of a dog, is a mammal—but if we reject the argument's conclusion we must also reject either the premise that Fido is a dog (we exclude that dog-like robots count as dogs), or the premise that all dogs are mammals (we also include dog-like robots among "dogs").

By contrast, a defeasible argument can be attacked by denying its conclusion, even if its premises are not questioned. For instance, let us assume that Fido is not only a pet dog but also a Doberman and that Dobermans are presumably (normally) aggressive. Then, as shown in Fig. 5 we can build an argument that attacks the previous argument by contradicting its conclusions (attack is expressed by the jagged arrow), and reject the conclusion of the attacked arguments, while still accepting its premises.

Clearly, we cannot endorse both arguments A and B at the same time (their conclusions are contradictory), and so we must either choose between them or remain uncertain as to which one we should choose. When two arguments conflict in such a way that the (final or intermediate) conclusion of one of them contradicts a (final or intermediate) conclusion of the other, we have a *rebutting conflict* between two

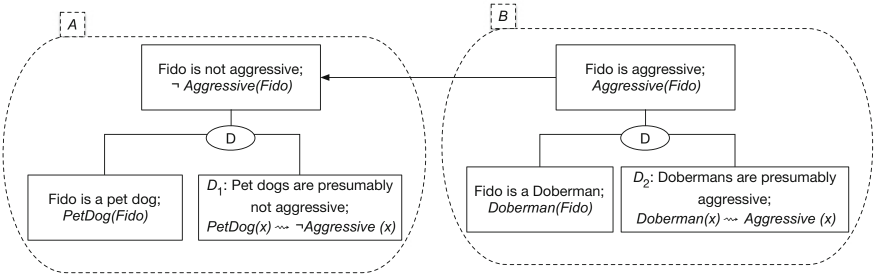


Fig. 5 Rebutting attack

arguments. To determine the outcome of a rebutting conflict, we must consider the comparative strength of the two arguments. If one argument is stronger than the other (at the juncture at which the conflict takes place), then it prevails; i.e. it defeats its opponent without being defeated by it. In this case, the prevailing argument is said to *strictly defeat* its opponent. If neither of the conflicting arguments is stronger than the other, they each *weakly defeats* the other; i.e. their conflict remains undecided (for a logical analysis of these notions, see Prakken and Sartor 1997; Prakken 2010). To compare arguments, we adopt here the so-called last-link principle, which affirms that when two defeasible arguments contradict each other, to determine the comparative strength of the two argument, at the point of where they clash, we must compare only the defaults that directly deliver the conflicting conclusions (possibly with the help of deductive inferences). We do not consider the defaults eventually used, in multistep arguments, to establish the preconditions of the directly conflicting defaults (for a discussion of the last-link principle and a formal definition, see Prakken 2010, Sect. 6).

In our example, let us assume that we consider that the argument on the right in Fig. 5 (Fido presumably is aggressive, given than it is a Doberman) is stronger that the argument on the left (Fido presumably is not aggressive, being a pet dog). According to this priority relation between the two arguments, the first can be said to strictly defeat the second: we should accept the conclusion that Fido is indeed aggressive (and be careful in approaching him).

A second kind of attack against defeasible arguments consists in contesting the support link between the premises and the conclusion of the argument, namely in denying that, in the case at hand, these premises provide sufficient support for the conclusion. This kind of conflict is called *undercutting* (on undercutting, see Pollock 2008). Let us assume that we are dealing with another dog—let us call him Tommy—and let us assume that we know Tommy to be a pet dog, but we also that he has been reared in an isolated mountain hut, having had contact only with his owner, and that we believe that the nonaggressiveness of pet dog towards strangers is mainly due to their experience in previous interactions with a large enough set of humans. We can then reasonably claim that, under these particular circumstances, the fact that

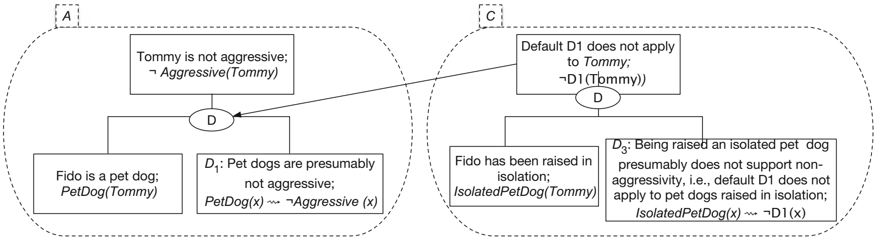

A
Tommy is not aggressive;
¬ Aggressive(Tommy)
D
Fido is a pet dog;
PetDog(Tommy)
D1: Pet dogs are presumably not aggressive;
PetDog(x) ⇝ ¬Aggressive (x)
C
Default D1 does not apply to Tommy;
¬D1(Tommy))
D
Fido has been raised in isolation;
IsolatedPetDog(Tommy)
D3: Being raised an isolated pet dog presumably does not support non-aggressivity, i.e., default D1 does not apply to pet dogs raised in isolation;
IsolatedPetDog(x) ⇝ ¬D1(x)

Fig. 6 Undercutting attack

Tommy is a pet dog does not adequately support the conclusion that he is friendly towards strangers. (see Fig. 6).

An undercutting argument always strictly defeats the argument it attacks, since (contrary to what happens in rebutting) it is not counterattacked by the latter argument. In fact, the undercutter says that the undercut argument does not work in the case at hand, while the undercut argument does not say anything about its undercutter. Note that the undercutter could also be viewed as attacking the particular instance of the default that is applied in the inference. For instance, it may be said that undercutter in Fig. 6 denies that the conditional "if Fido is a pet dog, then presumably he is not aggressive" holds; i.e. it denies that the fact that Fido is a pet dog is a reason for him not to be aggressive (given the conditions in which Fido has been raised). However, I prefer to view the undercutter as an attack against a particular inference applying the general default, to stress that the general conditional, stating a presumptive connection, is not affected by the attack.

Let us consider an example pertaining to the epistemology of perception (see Fig. 7). Assume that in the park I see a bird that to me looks pink (I perceive it in this way), and therefore, I conclude that the bird is pink. However, assume that I also see that the bird is a swan, which leads me to conclude that the bird is white, as swans normally are. However, since I know little about swans, I may remain in doubt about the colour of the bird: am I seeing a special swan (are there any pink swans around?) or is my perception of pink misleading me. Assume, however, that I notice that there is a red sunset. Then, as I know that even white things (not only pink ones) look pink under a red light, I will conclude that the fact that the bird looks pink under these conditions does not guarantee that it is indeed pink (it might as well be white): this undercuts the inference from *looking* pink to *being* so.

6 Rebutting and Undercutting in the Legal Domain

Let us now take up rebutting and undercutting in the legal domain. Consider three norms dealing with civil liability (they are somewhat simplified versions of rules in the Italian Civil code; note that I assume that all such norms express the presumptive

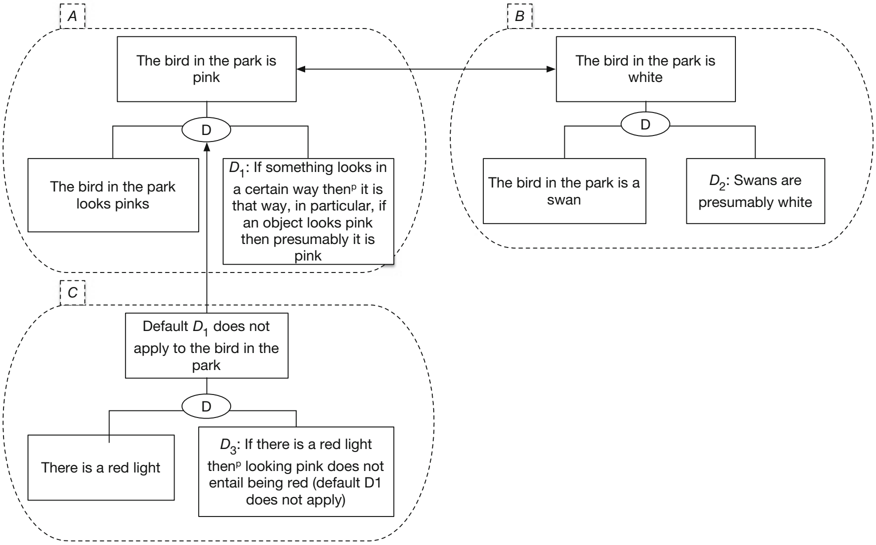

A
The bird in the park is pink
D
The bird in the park looks pinks
D_1: If something looks in a certain way then^p it is that way, in particular, if an object looks pink then presumably it is pink
B
The bird in the park is white
D
The bird in the park is a swan
D_2: Swans are presumably white
C
Default D_1 does not apply to the bird in the park
D
There is a red light
D_3: If there is a red light then^p looking pink does not entail being red (default D1 does not apply)

Fig. 7 Undercutting attack: defeasible perception

conditional connection "then presumably," which is abbreviated as then^p in the figures): the first rule (D1) says that those who cause damage to another through their fault are presumably liable, the second (D2) that persons lacking capacity are presumably not liable, and the third (D3) that the incapacity exception presumably does not apply to those who find themselves in a state of incapacity through their own fault (see Fig. 8). Assume that we know that John culpably caused damage to Tom (e.g. by deliberately smashing his car). On the basis of this information and of the first norm, we can conclude that John is liable to pay damages (argument A). However, assume that it appears that John lacked capacity at the time of the incident. Then, we can have an argument as to why John is not liable (argument B). Indeed, the incapacity exception takes priority over the general liability rule, so that argument B defeats argument A without being defeated by it. Assume, however, that John's incapacity was due to his fault, e.g. to his taking illegal drugs. This provides us with a third argument (*C*) that undercuts (makes irrelevant) the incapacity exception.

In legal contexts, a different way of undercutting can also be found. This involves those cases in which a legal norm explicitly includes among its preconditions the absence of an "impeditive fact," namely a fact such that if were established, it would prevent the norm's conclusion being derived (on impeditive facts, see Sartor 1993). This is conveyed by stating that the norm's consequent follows from certain conditions, unless the impeditive fact holds, or by stating that it follows from such conditions if the impeditive fact is not established. The norm's consequent can be derived without needing to establish the absence of the impeditive fact, while establishing that fact would prevent that derivation.

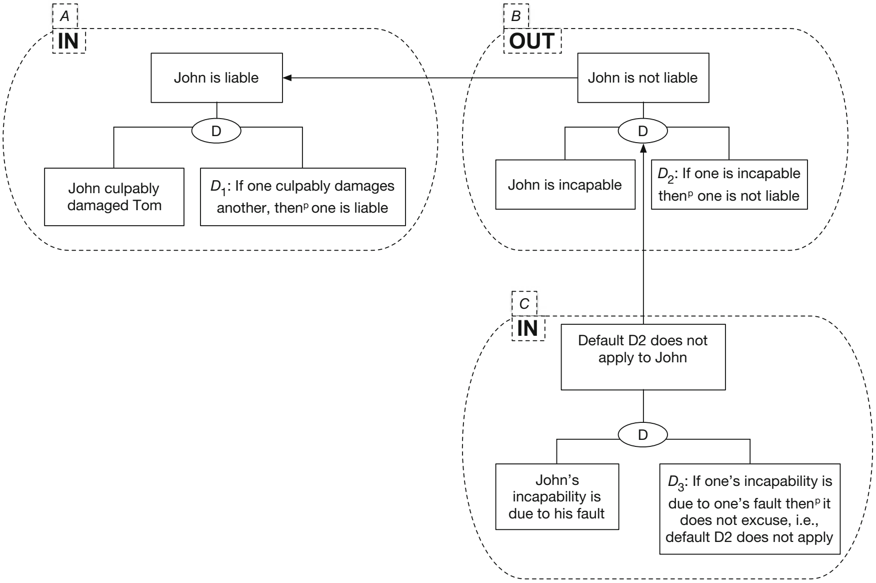

A
IN
John is liable
D
John culpably damaged Tom
D_1: If one culpably damages another, then^p one is liable
B
OUT
John is not liable
D
John is incapable
D_2: If one is incapable then^p one is not liable
C
IN
Default D2 does not apply to John
D
John's incapability is due to his fault
D_3: If one's incapability is due to one's fault then^p it does not excuse, i.e., default D2 does not apply

Fig. 8 Undercutting attack: inapplicability rule

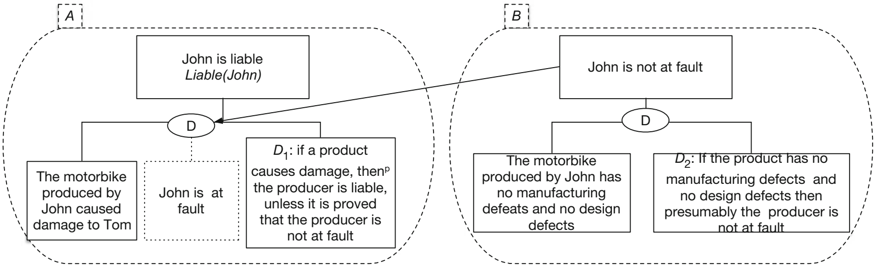

A
John is liable
Liable(John)
D
The motorbike produced by John caused damage to Tom
John is at fault
D_1: if a product causes damage, then^p the producer is liable, unless it is proved that the producer is not at fault
B
John is not at fault
D
The motorbike produced by John has no manufacturing defeats and no design defects
D_2: If the product has no manufacturing defects and no design defects then presumably the producer is not at fault

Fig. 9 Undercutting attack: impeditive fact

Consider, for instance, the rule in Italian law under which a producer is liable when a product it manufactures harms a consumer, unless it is shown that the producer is not at fault (took all reasonable precautions). Here, the impeditive fact is the absence of fault on the producer's side. Let us consider the issue of whether John may be liable as the producer of the motorbike which caused Tom's accident by failing to come to a stop before an obstacle (see Fig. 9). It is not necessary to establish John's fault to determine his liability as a producer. In other words, John's liability can be presumed by applying this norm (this is denoted by the dotted lines around this premise). However, if it is established that the motorbike was not defective, John may avoid liability.

7 Levels of Abstraction of Arguments

Defaults can have different levels of abstraction, some representing general patterns of inference or inference schemes (Walton et al. 2008), others representing more specific connections between preconditions and conclusions. Indeed, the same conclusion can often be argued by using either a general inference scheme or a more specific rule. Consider, for instance, the issue of the morality of lying, which was the object of a famous controversy between Emmanuel Kant and Benjamin Constant (see Kant 1949). Assume that John shows up at Mary's door and asks her whether Bob is at her place. Assume that Bob is in the house, that Mary is aware of this, and that Mary knows that John is armed and intends to kill Bob. The issue is whether Mary should lie, saying that Bob is away so as to save his life.

Let us first consider the argument according to which Mary should not lie. One way to frame it is as an argument pertaining to the implementation of moral rules in general. In that case, the premises of the argument could be presented as follows:

(1) If rule "if P then Q" is a moral principle, and P is the case, then presumably Q.
(2) The rule "if a statement is a lie, then one should not make the statement" is a moral principle.
(3) The statement that Bob is away is a lie.

By defeasible modus ponens, these premises lead to the conclusion that

(4) Mary should not make the statement that Bob is away.

However, the argument can also be framed in a more specific way, taking the prohibition on lying for granted and using it directly as a premise:

(1) The statement that Bob is away is a lie.
(2) If a statement is a lie, then presumably one should not make the statement.

It seems to me that this second approach fits better our commonsense reasoning, in which we directly use the warrants we endorse, to derive specific conclusions. Considerations pertaining to the foundation or the nature of such warrants are brought in through further arguments. For instance, the adoption of a warrant may be supported by arguments pointing to the consequences of its practice (e.g. in a rule-utilitarian perspective, the prohibition to lie may be supported by considering the benefit deriving from his generalised practice). Similarly, the strength and function of a warrant can be supported by arguments pointing to its nature (e.g. the fact that a principle pertains to morality may support its superiority over self-interested reasons, or the fact that it belongs to the law may support its coercive enforceability or its exclusionary nature).

Again, by defeasible modus ponens, premises 1 and 2 lead to the presumable conclusion that

(3) Mary should not make the statement that Bob is away.

Let us now consider an argument why Mary should, on the contrary, lie. To build this argument, we can appeal to a different pattern of defeasible inference: call it

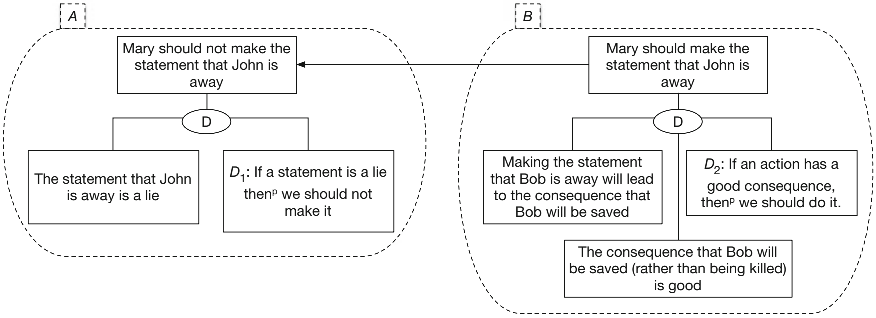

A
Mary should not make the statement that John is away
D
The statement that John is away is a lie
D1: If a statement is a lie thenp we should not make it
B
Mary should make the statement than John is away
D
Making the statement that Bob is away will lead to the consequence that Bob will be saved
D2: If an action has a good consequence, thenp we should do it.
The consequence that Bob will be saved (rather than being killed) is good

Fig. 10 Conflicting arguments: strict defeat

"inference from good consequences" or teleological argument. According to this pattern, the premises

(1) Making the statement that Bob is away will lead to the consequence that Bob will be saved (rather than being killed by John).
(2) This consequence is good.
(3) If an action has a good consequence, then presumably we should do it.

lead to the conclusion that

(4) Mary should make the statement that Bob is away.

The two arguments and their conflict are represented in Fig. 10.

If we agree that argument *B* is stronger than argument *A*, we should maintain that it strictly defeats argument *A*, and consequently, we should endorse the consequence of *B*; i.e. Mary should tell John that Bob is away (even if it is a lie).

8 Reinstatement

So far, we have only considered relations between *pairs* of arguments. However, this is insufficient to determine the status of an argument, namely whether we should accept it or not. More precisely, this is insufficient to establish whether an argument is: justified, such that we should accept its conclusion; overruled, such that we should not pay attention to it; or merely defensible, such that we should remain uncertain as to whether to accept it or not (on justified, overruled, and defensible arguments, see Prakken and Sartor 1997). This is because an argument *A* that is defeated by a counterargument *B* can still be acceptable when *B* is in turn defeated by a further argument *C*: we would have rejected *A* if we had accepted *B*, but since we do not accept *B* (given that it is defeated by *C*), then *A* remains acceptable.

To clarify this point, it is useful to specify the conditions that an argument should meet to be IN (acceptable) or OUT (inacceptable). The basic idea is that only a

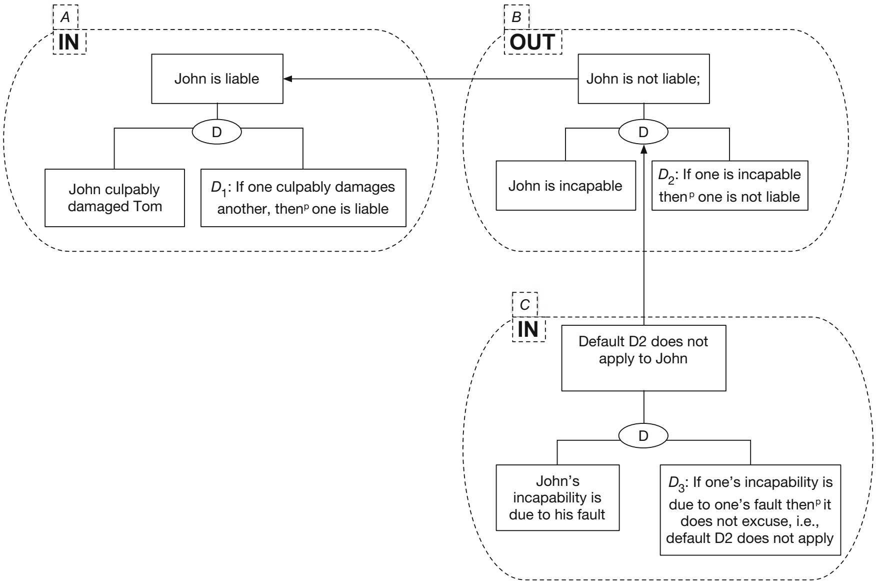

A
IN
John is liable
D
John culpably damaged Tom
D_1: If one culpably damages another, then^p one is liable
B
OUT
John is not liable;
D
John is incapable
D_2: If one is incapable then^p one is not liable
C
IN
Default D2 does not apply to John
D
John's incapability is due to his fault
D_3: If one's incapability is due to one's fault then^p it does not excuse, i.e., default D2 does not apply

Fig. 11 Reinstatement

defeater which is IN can turn OUT the argument it attacks; a defeater which is OUT is not relevant to the status of the argument it attacks. Thus, we can state the following rules:

(1) An argument $\mathcal{A}$ is IN iff no argument which defeats $\mathcal{A}$ is IN.
(2) An argument $\mathcal{A}$ is OUT iff an argument which defeats $\mathcal{A}$ is IN.

To clarify our analysis, let us consider the legal example in Fig. 11, which extends Fig. 8 with labels denoting the statuses of the corresponding arguments.

Relative to the set of arguments in Fig. 11 (A, B, and C), argument C is necessarily IN, since no defeater questions its status. Therefore, argument B is OUT (having a defeater, namely A, which is IN). Consequently, argument A is IN, since it has no defeater which is IN. This is the only assignment of IN and OUT labels that is consistent with rules (1) and (2). Consequently, A is justified, and so is its conclusion (John is liable), B is overruled, and C is justified.

This example shows the connection between dialectics and nonmonotonicity. By introducing new arguments into an argument framework (typically, the set of the arguments proposed in a debate or constructible from a given set of premises), the status of the pre-existing arguments may change relative to that framework: arguments that were justified may now be overruled and arguments that were overruled may now be justified.

Rules (1) and (2) above fail to univocally determine the status of those arguments in cases where we have an unresolved conflict (see Fig. 13, which depicts the divergent opinions of two experts).

In such a case, which arguments are justified depends on where we start from: if arguments $\mathcal{A}$ and $\mathcal{B}$ attack each other (and neither of them is OUT on other grounds), then if we assume that $\mathcal{A}$ is IN, then $\mathcal{B}$ will be OUT, and if we assume that $\mathcal{B}$ is IN, then $\mathcal{A}$ will be OUT. We can deal with this situation by considering all possible assignments of IN and OUT labels to the arguments at stake, consistently with rules (1) and (2) above: an argument is justified if it is IN according to every assignment; it is overruled if it is OUT according every assignment; it is defensible if it is IN according to some assignment and OUT according to some other assignment (Pollock 2008). An equivalent approach by which to assess the status of arguments consists in constructing alternative extensions, namely maximal sets of consistent arguments: justified, defensible, and overruled arguments are contained in all, some, or no extensions (Dung 1995).

Unresolved conflicts concerning legal and factual issues are addressed in different ways in the adversarial context of legal disputes, where the judge is assumed to know the applicable law, while the parties should bring evidence on the facts of the case. If an unresolved conflict between competing arguments concerns a legal issue (e.g. there are arguments supporting alternative interpretations of the same source of law), the decision-maker (the judge) should resolve the conflict by assigning priority to one of the conflicting arguments (on defeasible reasoning in legal interpretation, see Walton et al. 2016). If the unresolved conflict concerns a factual premise that is needed to construct an argument, it will be assumed that the factual premise have not been legally substantiated; therefore the factual argument will be OUT.

The dialectical interaction between arguments and counterarguments is reflected in the allocation of burdens of proof and, more generally, of burdens of argumentation. The idea of the burden of proof applies to many dialectical interactions, in context-dependent ways (see Walton 2008, 59), but it acquires a specific significance in the law (see Sartor 1993; Prakken and Sartor 2009). In a legal case, the party that is interested in establishing a legal outcome bears the burden of presenting and substantiating an argument supporting that outcome. For instance, in the example in Fig. 11, plaintiff Mary must provide argument A, establishing John's liability for negligence. She must substantiate the argument's normative premises (the general rule of civil liability) by referring to sources of law, and its factual premise (John culpably damaged Mary) by bringing in appropriate evidence. This argument will be sufficient for Mary to win the case if its premises are accepted and no counterarguments defeating it are provided by John (at least with regard to factual premises, since the judge may independently bring in legal information).

Thus, Mary can be said to bear the burden of proving that John did damage to her, since without establishing this fact she will not be able to construct argument A, which supports the outcome she favours. She does not bear the burden of proving that John was not incapable, since to build argument A, she does not need to establish that fact. On the contrary, John bears the burden of proving that he was incapable, since without establishing this fact, he will not be able to substantiate argument B, which could defeat Mary's argument A (switching A's status from IN to OUT, in the absence of further interfering arguments).

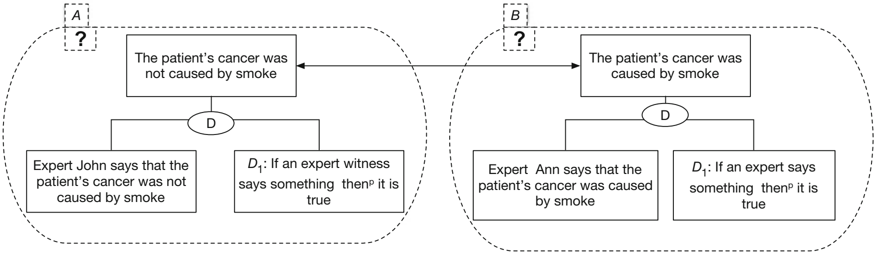

A
?
The patient's cancer was not caused by smoke
D
Expert John says that the patient's cancer was not caused by smoke
D_1: If an expert witness says something then^p it is true
B
?
The patient's cancer was caused by smoke
D
Expert Ann says that the patient's cancer was caused by smoke
D_1: If an expert says something then^p it is true

Fig. 12 Undecided conflict

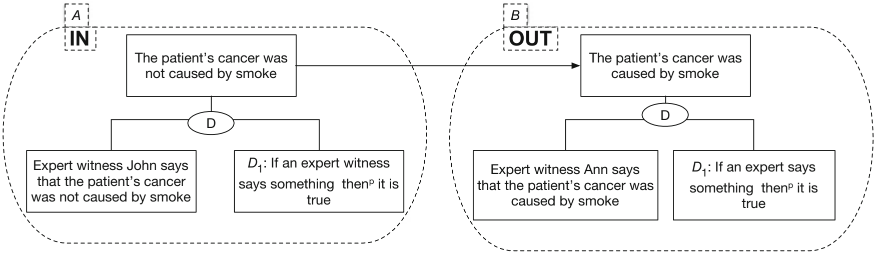

A
IN
The patient's cancer was not caused by smoke
D
Expert witness John says that the patient's cancer was not caused by smoke
D_1: If an expert witness says something then^p it is true
B
OUT
The patient's cancer was caused by smoke
D
Expert witness Ann says that the patient's cancer was caused by smoke
D_1: If an expert says something then^p it is true

Fig. 13 Decided conflict

In general, when a party π_1 fails to construct a certain legally acceptable argument $\mathcal{A}$ supporting her side unless evidence is provided for premise P, we say that π_1 has the burden of proof regarding P. This does not mean that the counterparty π_2 has no interest in P. It is true that π_1 will fail to build the argument based on P if π_1 fails to provide evidence for P, even if π_2 remains inactive. However, if π_1 provides sufficient evidence for P (and the other premises of $\mathcal{A}$ have been established), then π_2 must provide evidence against P, or other counterarguments against $\mathcal{A}$, if he does not want to lose on the basis of $\mathcal{A}$ (on the logic of the burden of proof, and for further refinements, including the distinction among the burden of production, the burden of persuasion, tactical burden, and standards of proof, see Prakken and Sartor 2009, on the connection between defeasibility and proof, see also Sartor 1993; Duarte de Almeida 2013; Brewer 2011). Figure 13 exemplifies the context of the burden of proof. Let us assume that the plaintiff (the alleged victim) has the burden of showing that his cancer was caused by smoke and that the standard of preponderance of the evidence applies. Then, even if the two arguments have equal weight, the plaintiff's argument would be strictly defeated by the defendant's argument (to defeat the defendant's argument, the plaintiff's argument must meet the required standard of proof). Thus, in an adversarial legal context governed by the burden of proof, the status assignment of Fig. 12 (no justified arguments, two defensible one) would be transformed into the assignment of Fig. 13 (one justified and one overruled argument).

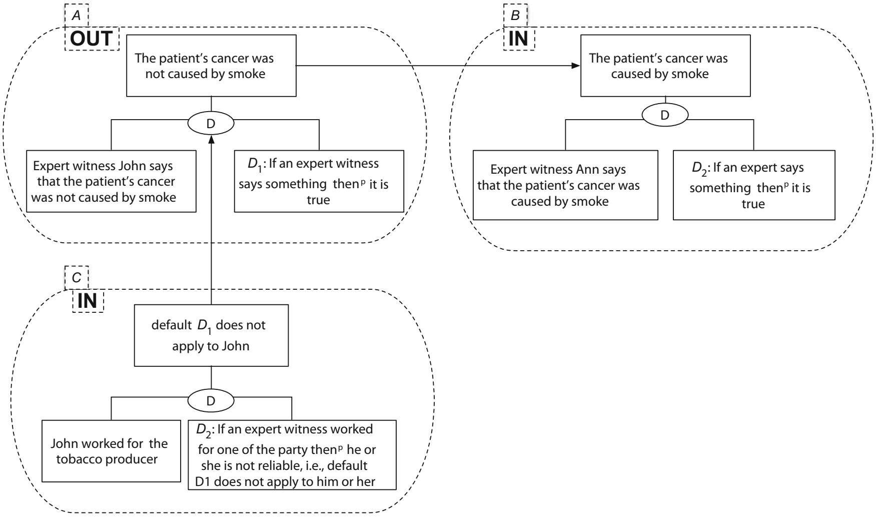

A
OUT
The patient's cancer was not caused by smoke
D
Expert witness John says that the patient's cancer was not caused by smoke
D_1: If an expert witness says something then^p it is true
B
IN
The patient's cancer was caused by smoke
D
Expert witness Ann says that the patient's cancer was caused by smoke
D_2: If an expert says something then^p it is true
C
IN
default D_1 does not apply to John
D
John worked for the tobacco producer
D_2: If an expert witness worked for one of the party then^p he or she is not reliable, i.e., default D1 does not apply to him or her

Fig. 14 Burden of proof and reinstatement

However, the patient may address this situation not only by providing additional evidence, so that his argument outweighs the doctor's argument, but also by undercutting the doctor's argument, e.g. by successfully contesting the reliability of the expert testimony in defence, as shown in Fig. 14.

9 Dynamic Priorities

In the previous examples involving priorities over arguments, we assumed that priorities were given. However, even priorities may be determined by (defeasible) arguments. Usually, a conflict between competing arguments is adjudicated according to the comparative strength of the defaults included in such arguments. Therefore, priority arguments aim to establish the comparative strength of conflicting defaults. In the legal domain, where legal norms provide the relevant defaults, priority arguments may appeal to formal legal principles—i.e. criteria which do not refer to the content of the norms at issue—such as the preference accorded to the more recent laws (*lex posterior derogat legi priori*), to the more specific ones (*lex specialis derogat legi generali*), or to those issued by a higher authority (*lex superior derogat legi inferiori*). Priority arguments may also be supported by textual clues; e.g. norms having negative conclusions are usually meant to override norms having the corresponding positive conclusions. Finally, priority arguments may refer to the substance of the norms at issue, e.g. assigning priority to the norm that promotes the most important values (legally valuable interests) to a greater extent.

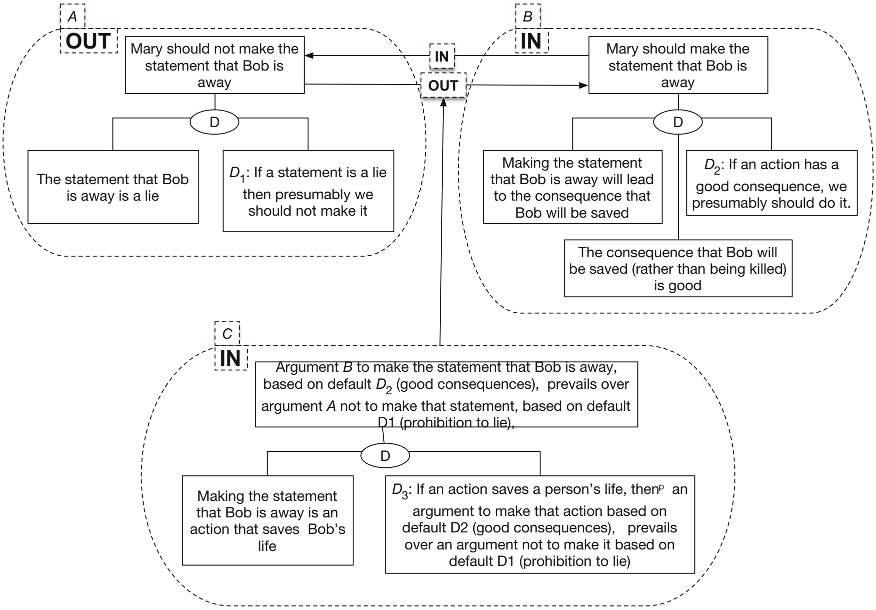

A
OUT
Mary should not make the statement that Bob is away
D
The statement that Bob is away is a lie
D_1: If a statement is a lie then presumably we should not make it
IN
OUT
B
IN
Mary should make the statement that Bob is away
D
Making the statement that Bob is away will lead to the consequence that Bob will be saved
D_2: If an action has a good consequence, we presumably should do it.
The consequence that Bob will be saved (rather than being killed) is good
C
IN
Argument B to make the statement that Bob is away, based on default D_2 (good consequences), prevails over argument A not to make that statement, based on default D1 (prohibition to lie),
D
Making the statement that Bob is away is an action that saves Bob's life
D_3: If an action saves a person's life, then^{p} an argument to make that action based on default D2 (good consequences), prevails over an argument not to make it based on default D1 (prohibition to lie)

Fig. 15 Dynamic priorities

One way to deal with the argumentative role of priority arguments consists in extending the IN and OUT labelling to defeat links between arguments. The previous rules can then be rewritten as follows:

(1) An argument $\mathcal{A}$ or a defeat link d is IN iff no argument which is IN defeats $\mathcal{A}$ or d through a defeat link which is IN.

(2) An argument $\mathcal{A}$ or a defeat link d is OUT iff an argument which is IN defeats $\mathcal{A}$ or d through a defeat link which is IN.

We need to specify when a defeat link is defeated: an argument $\mathcal{C}$ defeats the defeat link d denoting a rebutting attack from $\mathcal{A}$ to $\mathcal{B}$ when $\mathcal{C}$ states that $\mathcal{B}$ prevails over $\mathcal{A}$.

To clarify this idea, let us return to the issue of the admissibility of lying to save a person's life. Let us now add a priority argument (C), stating that, since the statement that Bob is away will save Bob's life, the duty to make the statement, as supported by the argument from good consequences, outweighs the duty not to make it, as supported by the prohibition on lying (Fig. 15).

Argument C affirms that argument B (for the duty to say that Bob is away) is stronger than argument A (for the duty not to make that statement). Therefore, we can conclude that the defeat link from A to B is OUT (as a weaker argument cannot rebut a stronger one), while the defeat link from B to A remains IN. Therefore, B strictly defeats A (it defeats it without being defeated by it). Consequently, B is IN and A is OUT: we should tell the lie.

Obviously, the opposite conclusion would follow if we took a different view of priorities, such as the view that deontological arguments, warranted by generalisable rules, always have priority over consequentialist arguments.

10 Patterns of Defeasible Reasoning

Various warrants (general defaults) for defeasible reasoning can be identified. The following ones are discussed by Pollock (1998, 2008):

- *Perceptual inference*. If I have a percept with content P, then I can presumably conclude that P is true. For instance, if I have an image of a book at the centre of my field of vision, I can conclude that there is a book in front of me. This conclusion is defeated if I become aware of circumstances that do not ensure the reliability of my perceptions (I am watching a hologram).
- *Memory inference*. If I remember P, then I can presumably conclude that P is true. For instance, my recollection that yesterday I had a faculty meeting lends presumptive support to the conclusion that there was such a meeting. This inference is defeated if I come to believe that my supposed recollection was an outcome of my imagination.
- *Enumerative induction*. If I observe a large enough sample of Fs, all of which are Gs, then I can presumably conclude that all Fs are Gs. For instance, if all crows I have ever seen are black, then I can presumably conclude that all crows are black. This inference is defeated if I should see a white crow.
- *Statistical syllogism*. If most Fs are Gs and an individual a is an F, then I can presumably conclude that a is a G. For instance, assume that (1) the pages of most printed books are even-numbered on their verso side and that (2) the bound pages on my table are a printed book. I can then conclude that these bound pages are even-numbered on their verso side. This inference is defeated if I discover that these bound pages were incorrectly printed with even numbers pages on their recto side.
- *Temporal persistence*. If it is the case that P at time t_1, then presumably P is still the case at a later time t_2. For instance, if my computer was on my table yesterday evening (when I last saw it), then presumably it will still be there. This inference is defeated if I come to know that the computer was moved from the table after I last saw it, and more generally if I have any reason to believe that its location may have changed.

General processes of human cognition, such as abduction and analogy (see Walton et al. 2008) can support further schemes for defeasible arguments, such as the following:

- *Abduction of a cause.* If Q is the case, and P causes Q, then presumably P was the case. For instance, if the grass is wet, and rain causes the grass to be wet, then presumably it has rained. Arguments based on this warrant can be defeated in different ways: by indicating alternative, no less probable, causes of the effect (e.g. somebody has watered the grass), or by showing an inconsistency between the cause (rain) and other states of affairs (e.g. the street is not wet, as it should be if it had rained), etc.

Basic analogy. If P is relevantly similar to Q, and P has property R, then presumably also Q has property R. For instance, if detecting something by just seeing does not count as search, and detecting something by just seeing is relevantly similar to detecting drug with a sniffing dog, then also detecting drug with a sniffing dog does not count as search (for refinements of the analogy pattern, and for a discussion of the sniffing dog case, see Walton et al. 2008, Chap. 2). Arguments based on analogy can be attacked by questioning or denying that there is a relevant similarity, by pointing to relevant differences, by bringing counterexamples, etc (I cannot enter here the discussion on what may count as a relevant similarity or difference, on analogy see Brozek, Chap. 4, part II, this volume, on "Analogical Arguments"). In many cases, an analogical conclusion can (or should) also be supported by a more elaborate piece of reasoning, where the aspects that make the similarity relevant are presented as the antecedents of a general warrant (e.g. detecting something without actively interfering does not count as search) which is abducted to explain the common conclusion (see Walton et al. 2008, Chap. 2; Brewer 1996).

These defeasible warrants are not meant to substitute the logical, philosophical, or psychological theories of the phenomena they address, such as perception, induction, abduction, or analogy (see for instance, for analogy, Holyoak and Thagard 1996). They should be rather viewed as rules of thumb that may be supported, explained, and constrained by such theories.

In the previous examples, I have considered further general defaults, such as those enabling the argument from good or bad consequences or the argument from expert testimony. I have also observed that more specific defaults may be used to construct defeasible arguments: empirical generalisations, as well as legal and moral norms, can be viewed as defaults. In fact, the set of the defaults that may be used in individual and social cognition cannot be reduced to an exhaustive list, since default warrants are justified pragmatically, i.e. because of how well they serve the needs of different practical or epistemic activity types (Walton and Sartor 2013). The successful use of a default warrant in a social activity (such as legal reasoning) critically depends on the extent to which the scheme enjoys shared acceptance, as providing valid support to its conclusions (since the default's acceptance is a crucial precondition of its successful use in arguments meant to convince other people, or to converge with them into shared conclusions). Thus, even abstract legal principles, such as interpretive canons, only justify their conclusions in those legal systems in which they are in fact endorsed and deployed, so as to enjoy the status of social and institutional normative principles.

It is important to stress that defeasible arguments can include multiple steps. For instance, in an argument culminating in the conclusion of a rule, the rule may be supported by an interpretive argument, while rule's factual antecedent may result from

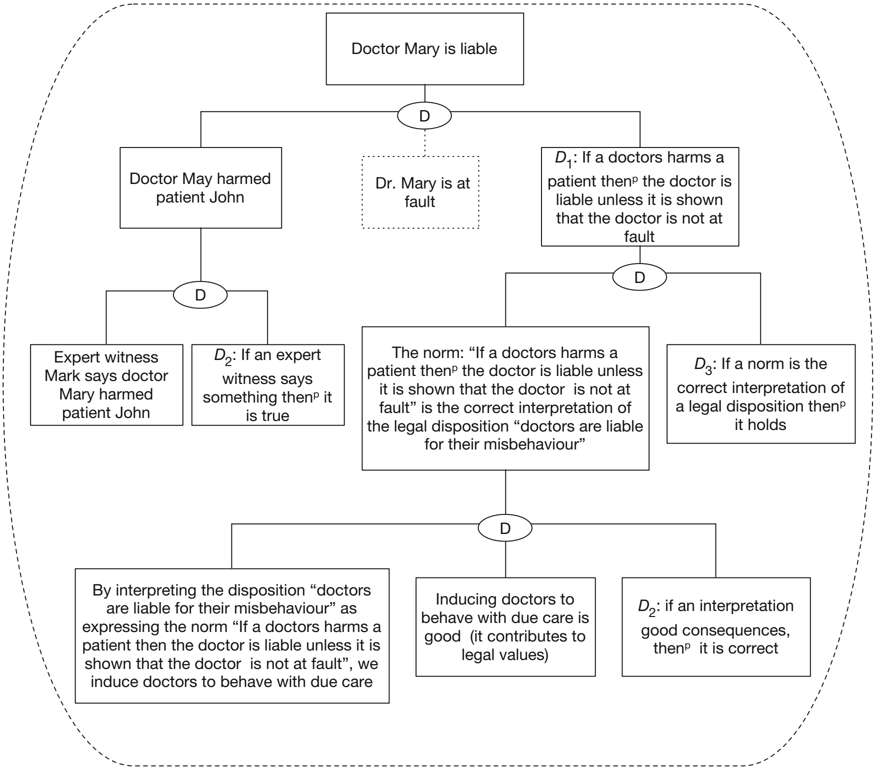

Doctor Mary is liable
D
Doctor May harmed patient John
Dr. Mary is at fault
D1: If a doctors harms a patient thenp the doctor is liable unless it is shown that the doctor is not at fault
D
Expert witness Mark says doctor Mary harmed patient John
D2: If an expert witness says something thenp it is true
D
The norm: "If a doctors harms a patient thenp the doctor is liable unless it is shown that the doctor is not at fault" is the correct interpretation of the legal disposition "doctors are liable for their misbehaviour"
D3: If a norm is the correct interpretation of a legal disposition thenp it holds
D
By interpreting the disposition "doctors are liable for their misbehaviour" as expressing the norm "If a doctors harms a patient then the doctor is liable unless it is shown that the doctor is not at fault", we induce doctors to behave with due care
Inducing doctors to behave with due care is good (it contributes to legal values)
D2: if an interpretation good consequences, thenp it is correct

Fig. 16 Multistep argument

arguments assessing the available evidence. Consider the liability case illustrated in Fig. 16. The argument for the liability of Doctor Mary includes the following:

- A norm-based argument that Mary is liable, since she harmed her patient and doctors are liable for harming their patients unless they are shown not to be at fault.
- A teleological interpretive argument (a subspecies of the argument from good consequences): the law on doctors' liability must be interpreted in this way, since this interpretation contributes to increasing diligence in the medical profession, which is a good thing.
- An empirical argument based on an expert testimony supporting the conclusion that there was a causal link between the Mary's behaviour and the patient's harm.

This argument is subject to a series of possible attacks, against each of its subarguments (a subargument being an argument which is included in a larger argument): its top subargument may be undercut by establishing that Mary was not at fault (she used the available medical knowledge correctly); the interpretive subargument can be attacked by contesting the very idea that the proposed interpretation promotes careful behaviour among doctors (on the contrary, it may undermine patient care, since doctors may become too risk-averse, knowing that they may face the difficult task of proving a negative, namely, that they did not act negligently); the empirical subargument can be rebutted by providing a contrary expert opinion, or it can be undercut by challenging the expert's reliability, among other options.

11 Legal Systems as Argumentation Bases

We have so far considered arguments and their interactions, i.e. conflicts giving rise to defeat relations. Let us now look at the set of premises that provide the ingredients for constructing a set of interacting arguments.

A set of such premises is not a consistent set of deductive axioms but is rather a repository of materials to be used to build competing arguments and counterarguments. It is an *argumentation basis,* in the sense of a knowledge base (a set of premises) that can be used for constructing an *argumentation framework* (a set of interacting arguments). In Sartor (1994), I used the term argumentation framework (see also Stone Sweet 2004, 34) to denote what I here call *argumentation basis* Here I reserve the term *argumentation framework* to the set of arguments that are constructible from the argumentation basis in order to be consistent with the prevailing terminology (Baroni et al. 2011).

Figure 17 (adapted from Baroni et al. 2011) shows the process to determine the inferential semantics of an argumentation basis, namely the set of all conclusions that are supported by that basis. First, we construct the maximal argumentation framework resulting from the argumentation basis; i.e. we build all arguments that can be obtained by using only the premises in the basis and we identify all defeat relations between such arguments. Then, we determine what arguments and defeat links are IN or OUT (for all or some labellings) and consequently establish the status of each argument, i.e. whether the argument is justified, defensible, or overruled relative to the given argumentation basis. Finally, we identify the status of the conclusions of these arguments, the conclusion of the justified or defeasible arguments being respectively justified or defensible relative to the argumentation basis. A different (but equivalent) approach is described in Prakken and Sartor (1997), where the proof of a defeasible conclusion takes place in a game where the proponent of that conclusion has to build an argument (from the argumentation base) and defend it against all possible direct and indirect counterarguments an opponent may construct (from the same argumentation base).

I shall argue that a legal system itself—considered from an argumentation standpoint, and complemented with the relevant factual evidence—indeed appears to be an argumentation basis rather than a deductive system. In fact, if we accept that the legal system contains general rules and exceptions, conflicting norms, and principles expressing incompatible legal interests, then we must reject the traditional postulate of the consistency of the law, and consequently, we must reject its image as

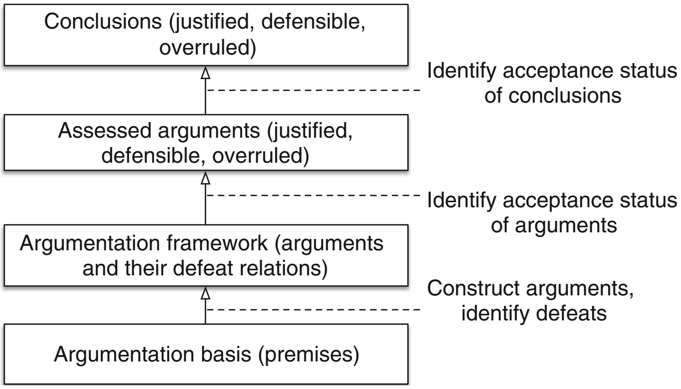

Conclusions (justified, defensible, overruled)
Identify acceptance status of conclusions
Assessed arguments (justified, defensible, overruled)
Identify acceptance status of arguments
Argumentation framework (arguments and their defeat relations)
Construct arguments, identify defeats
Argumentation basis (premises)

Fig. 17 Inferential semantics of an argumentation basis

an axiomatic base that, when combined with the relevant facts, yields conclusive deductive implications.

On the contrary, a legal system is a heterogeneous, stratified, and conflicting set of legal defaults (legal rules and principles, metarules, accepted argument schemes, etc.) which, when combined with the relevant facts, make it possible to derive presumptive conclusions. By complementing a legal system (the relevant portion of it) with the evidence establishing the operative facts of a case (facts that match the antecedents of some of the system's norms), we obtain an argumentation basis from which competing presumptive arguments may be constructed. To clarify this idea, let us assume, for simplicity's sake, that the legal system $\mathbb{L}$ in question only contains the three defeasible rules on civil liability included in the arguments in Fig. 11 above:

D_1: If one culpably damages another, one is liable: $CulpablyDamages\,(x, y) \Rightarrow Liable(x)$.
D_2: If one is incapable, one is not liable: $Incapable\,(x) \Rightarrow \neg Liable\,(x)$.
D_3: If one's incapability is due to one's fault, then it does not excuse; i.e. default D_2 does not apply: $IncapableByFault(x) \Rightarrow \neg D_2\,(x)$.

The three factual propositions (possible operative facts) that match the antecedents of these three rules are the following:

P_1: John culpably damages Tom: $CulpablyDamages(John, Tom)$.
P_2: John was incapable: $Incapable(John)$.
P_3: John's incapability is due to his fault: $IncapableByFault\,(John)$.

By complementing $\mathbb{L}$ with appropriate facts (any combination of P_1, P_2, and P_3), we obtain argumentation bases that makes it possible to construct different combinations of arguments A, B, and C (different facts being required for each of these arguments).

All these arguments are in principle defeasible, being susceptible to rebuttal or undercutting by appropriate counterarguments, should the latter become available. However, only A and B and can be defeated by counterarguments constructed with the norms in $\mathbb{L}$, plus corresponding operative facts, since $\mathbb{L}$ does not contain any default that may be used to build a defeater to C.

Let us consider, for instance, argument A in Fig. 11. This argument can be constructed from $\mathbb{L}$, complemented by the factual proposition F_1, since the premises for A are constituted by default D_1, which belongs to $\mathbb{L}$, and fact F_1. We can say that argument A can be defeated in $\mathbb{L}$, to mean that $\mathbb{L}$, complemented with appropriate facts, provides the resources for constructing a defeater to A. In fact, A is strictly defeated by B, which can be constructed from $\mathbb{L}$, complemented with factual proposition P_2. Also B can be defeated in $\mathbb{L}$, since B is defeated by C, which can be constructed from $\mathbb{L}$, complemented with the factual proposition P_3. On the other hand, C, while also being a defeasible argument, cannot be defeated in $\mathbb{L}$, since there is no operative fact that would make it possible to rebut or undercut C using only the rules in $\mathbb{L}$.

Note that the fact that an argument can be defeated in $\mathbb{L}$ does not mean that the argument fails to be justified in every argumentation basis obtainable by adding an appropriate set of operative facts to $\mathbb{L}$. For instance, if only the fact that John culpably damaged Tom is added to $\mathbb{L}$, we obtain the argumentation basis $\mathbb{L} \cup \{P_1\}$, from which we can only build argument A. Since no counterargument to A can be constructed from $\mathbb{L} \cup \{P_1\}$, A is justified relative to argumentation basis $\mathbb{L} \cup \{P_1\}$ and so is his conclusion: John is liable. If we also add the fact that John was incapable, we obtain the argumentation basis $\mathbb{L} \cup (P_1, P_2\}$, relative to which A is no longer justified, since A's strict defeater B can be constructed. Relative to $\mathbb{L} \cup \{P_1, P_2\}$, B is justified and so is his conclusion: John is not liable. Similarly, A would again be justified, and B would be overruled, relative to the argumentation basis $\mathbb{L} \cup \{P_1, P_2, P_3\}$, which makes it possible to construct argument C. Thus, relative to $\mathbb{L} \cup \{P_1, P_2\}$, which originates the argumentation framework $\{A, B, C\}$, A's conclusion is justified: John is liable.

An argument that cannot be defeated in a normative system $\mathbb{L}$ may be defeated in larger normative system. Assume, for instance, that through a legislative act or through judicial interpretation, a new norm D_4 is introduced, which is stronger than D_3:

D_4: If one's incapacity is due to a chronic condition (alcoholism or drug addiction), then the incapacity excuse, i.e. default D_2, does apply: $IncapableByChronicalCondition\,(x) \Rightarrow D_2\,(x)$.

Then, argument C, which could not be defeated in $\mathbb{L}$, can be strictly defeated in $\mathbb{L}' = \mathbb{L} \cup \{D_4\}$. In fact, $\mathbb{L}'$, in combination with the operative fact:

P_4: John is incapable by a chronical condition (e.g. alcoholism): $IncapableByChronicalCondition\,(John)$.

enables us to construct a further argument, let us call it G, that strictly defeats C. Thus, relative to the argumentation basis $\mathbb{L}' \cup \{P_1, P_2, P_3, P_4\}$, that originates the argumentation framework $\{A, B, C, G\}$, argument A is overruled, while argument B is justified, and so is its conclusion that John is not liable, as shown in Fig. 18.

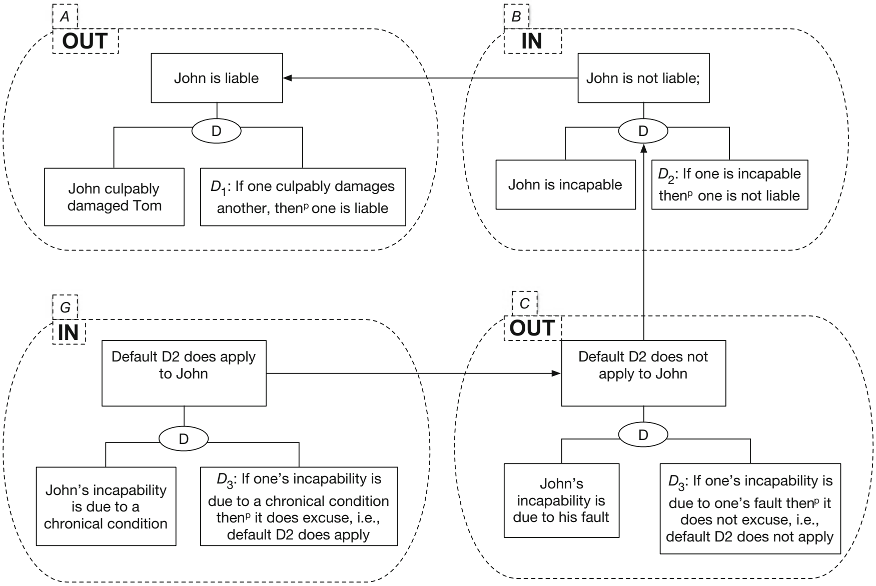

A
OUT
John is liable
D
John culpably damaged Tom
D_1: If one culpably damages another, then^p one is liable
B
IN
John is not liable;
D
John is incapable
D_2: If one is incapable then^p one is not liable
G
IN
Default D2 does apply to John
D
John's incapability is due to a chronical condition
D_3: If one's incapability is due to a chronical condition then^p it does excuse, i.e., default D2 does apply
C
OUT
Default D2 does not apply to John
D
John's incapability is due to his fault
D_3: If one's incapability is due to one's fault then^p it does not excuse, i.e., default D2 does not apply

Fig. 18 Defeat relative to an argumentation basis

12 The Rationale for Defeasibility

Pollock (1998) argues that defeasibility is a key aspect of human cognition (and more generally, of the cognition of any boundedly rational agent). We start with perceptual inputs and proceed by inferring beliefs from our current cognitive states (our percepts plus the beliefs we have previously inferred). A process so described must satisfy two apparently incompatible desiderata:

- We must form our beliefs on the basis of partial perceptual input (we cannot wait until we have a complete representation of our environment).
- We must be able to take an unlimited set of perceptual inputs into account.

According to Pollock, the only way to reconcile these requirements is by defeasible reasoning. We must adopt beliefs on the basis of a small set of perceptual inputs, but then we must be ready to retract these beliefs in the face of additional perceptual inputs, whenever these additional inputs conflict with the initial basis for our beliefs.

Thus, defeasible reasoning appears to have different, but related, functions (see Sartor 2005, Sects. 2.2 and 2.3). The first function consists in providing us with provisional beliefs, on the basis of which we can reason and act, until we gain information to the contrary.

The second function consists in activating a structured process of inquiry that consists in drawing *pro tanto* conclusions, looking for their defeaters, for defeaters of defeaters, and so on, until stable outcomes are obtained. This process has two main advantages: (1) it focuses the inquiry on relevant knowledge, and (2) it continues to deliver provisional results while the inquiry moves on.

A third function of defeasibility consists in enabling our collective knowledge structures to persist in time, i.e. to continue to work as a shared communal asset, even though each of us is exposed to new information, often challenging the information we already have.

We indeed have two basic strategies for coping with the provisional nature of human knowledge: revision and defeasibility.

Revision assumes that our general knowledge is a set of universal laws. When we discover a case where such universal laws lead us to a false (unacceptable or absurd) conclusion, we must conclude that our theory (or the subsets of it entailing the false conclusion) has been falsified, becoming thus unacceptable (Popper 1959). Thus, we must abandon some propositions in that theory and replace them with new universal propositions, from which the false conclusion is no longer derivable. Rational strategies for revising a theory have been the object of several studies (see, for instance, Alchourrón et al. 1985; Gärdenfors 1987). In the legal domain, this idea was originally proposed Alchourrón and Makinson (1981) and was subsequently developed by Maranhao (2013).

The other strategy, *defeasibility*, assumes that general propositions are defaults, which are meant govern most cases or the normal cases. Thus, we can consistently endorse such propositions and deny that they apply to certain cases: the exception serves the rule, or at least it does not compromise the rule. To deal with an anomalous case on a defeasibility strategy, we do not abandon the default or change its formulation, but instead we assume that the default's operation is limited on grounds that are different from those that support the use of the default itself. As we saw in the previous example, these grounds may provide an argument that undercuts or rebuts the argument warranted by the default. The idea that legal norms are defaults (rather than strict rules) makes possible a certain degree of stability in legal knowledge: we do not need to change our norms whenever their application is limited through subsequent exceptions or distinctions. However, this perspective does not exclude the need to abandon a norm, when it no longer reflects a "normal" connection, being superseded by subsequent norms (as in implicit derogation), or when it is explicitly removed from the knowledge base (as in explicit derogation: see Governatori and Rotolo 2010).

13 Defeasible Reasoning and Probability

Probability calculus—especially its versions based on the idea of subjective probability—provides an attractive alternative to defeasible reasoning as a method for dealing with limited and provisional information. It has a rich history of successful applications in many domains of science and practice, including legal practice (though its legal applications are still controversial: see Fenton et al. 2016) and has recently found many applications in artificial intelligence.

Consider, for instance, a case where Tom was run over by a car carrying Mary and John, and in which it is not clear who was driving at the time of the accident.

On the probabilistic approach, conflicting evidence does not lead us to incompatible belief—like the belief that John was driving the car when the car ran over Tom, and the belief Mary was driving the car on the same occasion—between which a choice is needed. We rather come to the consistent view that incompatible hypotheses have different probabilities. For instance, on the basis of the available evidence, we may consistently conclude that there is a 40% probability that John was driving, and a 60% probability that Mary was doing it. Probabilistic inference uses probability calculus to determine the probability of an event on the basis of the probability of other events. For instance, if there is an 80% probability that Tom will have problems in walking because he has been run over, there is a 32% probability (40% * 80%) that Tom will have such problems having been run over by John, and a 48% chance (60% * 80%) that he will have such problems having been run over by Mary. Here, I cannot enter probability calculus or discuss the many difficult issues related to it, especially when ideas of probability and causation are combined, or when Bayesian reasoning is used to determine the probability of a hypothesis in the light of the evidence. I will merely highlight three issues that make probability calculus inadequate as a general approach for dealing with uncertainty in legal reasoning.

The first issue is that of practicability: we often do not have enough information to assign numerical probabilities in a sensible way. For instance, how do I know that there is a 40% probability that John was driving and a 60% probability that Mary was driving? In such circumstances, it seems that we must attribute probabilities arbitrarily or, no less arbitrarily, we must assume that all alternative ways in which things may have turned out have the same probability.

The second issue is conceptual: although it makes sense to ascribe probabilities to factual propositions, it makes little sense to assign probabilities to legal rules and principles, unless we are making predictions. A legal decision-maker does not usually decide to use a normative premise by assessing the probability that the premise holds.

The third issue relates to psychology: humans tend to face situations of uncertainty by choosing to endorse hypothetically one of the available epistemic or practical alternatives (while keeping open the chance that other options may turn out to be preferable), and by applying their reasoning to this hypothesis (while possibly, at the same time, exploring what would be the case if things turn out to be different). We do not usually assign probabilities and then compute what further probabilities follow from such an assignment. When we have definite beliefs or hypotheses, we are usually good at developing inference chains, storing them in our minds (keeping them dormant until needed), and then retracting any such chains when one of its links is defeated. Conversely, we are bad at assigning numerical probabilities and even worse at deriving further probabilities and revising probability assignments in the light of further information.

Our inability to work with numerical probabilities certainly figures among the many failures of human cognition (like our inability to quickly execute large arithmetical calculations). In fact, computer systems exist which can handle efficiently complex probability networks (otherwise termed *belief networks* or *Bayesian networks*). They perform very well in certain domains by manipulating numerical probabilities much faster and more accurately than a normal person (see Russell and Norvig 2010, Chap. 13). However, our bias towards exploring alternative scenarios, and defeasibly endorsing one of them, does have some advantages: it focuses cognition on the implications of the most likely situations, it supports making long reasoning chains, it facilitates building scenarios (or stories) which may then be evaluated according to their coherence, and it enables us to link epistemic cognition with binary decision-making (it may be established that we have to adopt decision Q if P is the case, and NON-Q if P is not the case). There is indeed psychological evidence that humans develop theories even under situations of extreme uncertainty, when no reasonable probability assignment can be made.

The limited applicability of probability calculus in many domains does not exclude that there may be various practical and legal issues where statistics and probability provide decisive clues, as when scientific evidence is at issue.

Recently, approaches have been developed that try to combine defeasible reasoning and probability by working out the likelihood that different premises and combinations of them will be used in making arguments and that these will interact with other arguments. Such approaches would lead to probabilistic refinements of the IN and OUT labelling previously considered: rather than just saying that an argument is IN or OUT, we could establish that it has a certain probability of being IN or OUT relative to an argumentation basis whose premises or combinations of them are assigned certain probabilities (Riveret et al. 2012; Hunter 2013).

14 Defeasibility in the Law

Defeasible reasoning characterises the law at different levels.

First, clues to the defeasibility of legal reasoning are embedded in the very language of legal sources. As we saw in the previous example, the legislator itself often suggests how to construct defeaters to certain arguments. For example, to indicate that liability in tort can be excluded by appealing to self-defence or a state of necessity, the legislator may use any of the following formulations:

- *Unless clause*. One is liable if one voluntarily causes damage, unless one acts in self-defence or in a state of necessity.
- *Explicit exception*. One is liable if one voluntarily causes damage. One is not liable for damages if one acts in self-defence or in a state of necessity.
- *Presumption*. One is liable if one voluntarily causes damage and one does not act out of self-defence or a state of necessity. The absence of both is presumed.

According to all these formulations, to build an argument to the effect that one must make good some damage, it is normally sufficient to ascertain that one voluntarily caused that damage, but this argument is defeated by counterarguments appealing to the fact the person turns out to have acted either out of necessity or in self-defence.

Defeasibility is also an essential feature of *conceptual constructions* in the law. Legal concepts must be applied to such a diverse range of instances that they can at best offer a tentative and generic characterisation of the objects to which they apply, a characterisation that must be supplemented with exceptions. General legal concepts presuppose defeasibility: the requirement of absolute rigour in defining and applying concepts—the demand that all features which are included in, or entailed by, a concept apply to each of its instances—would paradoxically run counter to the very possibility of being "logical" in the sense of using general concepts. In fact, even the definitions of the legal concepts that can be found in statutes and codes reflect the stepwise defeasible process of establishing legal qualifications: first, a general discipline is established for a certain legal genus (e.g. the genus "contract"); special exceptions are then introduced for species within this genus (e.g. the species contract of sale); finally, further exceptions may be introduced for specific subspecies (e.g. the sale of real estate). Consequently, when using conceptual hierarchies, we must apply to a certain object the rules governing the category in which it is included, but only insofar as no exceptions emerge concerning a subcategory in which that object is also included.

Defeasibility can be deliberately established by the legislator, but it may also result from the evolution of legal knowledge: after a general rule has been established, exceptions are often provided for those cases where the rule appears to be inadequate.

This is typically the evolution of judge-made law, where general *rationes decidendi* are often limited by way of *distinctions*, that is by way of exceptions introduced for specific contexts (on defeasibility and precedents, see Prakken and Sartor 1998; Horty 2011). In such cases, judges often leave the original default rule unchanged and add a new prevailing rule that addresses the specific situations requiring a distinction. For instance, in the *Monge* case (US Supreme Court, 28 February 1974, No. 6637), the judges introduced an exception to the idea that contracts of employment at will (lacking any set term) could be terminated by both parties regardless of the reason ("for any reason or no reason at all"). They stated that "a termination by the employer of a contract of employment at will which is motivated by bad faith or malice or based on retaliation […] constitutes a breach of the employment contract." Correspondingly, on the basis of this rule the dismissed employee Olga Monge could build an argument (her dismissal was a breach of contract, being based on malice and retaliation) that could defeat the employer's argument that she could be legitimately dismissed on the ground that her contract was at will. Note that the judges could also have revised the original rule into a new rule: "a contract can be terminated by both parties for any reason unless the employer is terminating the contract motivated by bad faith or malice or based on retaliation." The new rule would have triggered the same dialectical exchange, as long as the unless clause was interpreted as attributing to the employee the burden of proving bad faith, malice, or retaliation.

Finally, we need to also consider the procedural aspect of defeasibility. As noted, this aspect concerns the fact that defeasible reasoning activates a structured process of inquiry in which we draw prima facie conclusions, look for their (prima facie) defeaters, look for defeaters of defeaters, and so on, until stable results can be obtained. A process like this one reflects the natural way in which legal reasoning proceeds. This is especially the case in the law's application to particular situations, when we have to consider the different, and possibly conflicting, legal rules that apply to such situations and must work out conflicts between these rules.

The defeasibility of legal reasoning also reflects the dialectics of judicial proceedings, where each party provides arguments supporting his or her position, and these arguments conflict with the arguments made by the other party. The debate of the parties is usually transferred to the judicial opinion that takes in the results of the dispute and determines its output. To convincingly justify a judicial decision in a case involving genuine issues, it is not sufficient to state a single argument; it is necessary to establish that the winning argument prevails over all arguments to the contrary, especially those that have been presented by the losing party, or that the latter arguments have to be rejected on other grounds.

Finally, doctrinal work cannot avoid being contaminated by the dialectics of legal proceedings, since its main function consists in providing general arguments and points of view to be used in judicial debates. From this perspective, doctrinal reasoning may be viewed as consisting in an exercise in *unilateral dialectics*, understood as a disputational model of inquiry in which "one develops a thesis against its rivals, with the aim of refining its formulation, uncovering its basis of rational support, and assessing its relative weight" (Rescher 1977, 47).

The significance of defeasibility in legal reasoning has been recently confirmed by the psychological experiments by Gazzo Castañeda and Knauff (2016), which show how both lawyers and laypersons reason defeasibly when applying legal norms. When presented with a legal conditional, in its usual formulation (If somebody kills a person, then he or she should be punished for manslaughter), and with an instance of the antecedent condition (Bert killed a persons), most participants in the experiment conclude for the conditional's conclusion (Bert should be punished for manslaughter), but withdraw this conclusion when told that an exculpatory circumstance (because of a psychological disorder, Bert was unable to control his actions) also obtains. The experiments also show that lawyers are better than laypersons in withdrawing legal conclusion when faced with legally recognised exceptions, having a more precise knowledge of such exceptions and of their role in legal reasoning.

15 Overcoming Legal Defeasibility?

Some authors have suggested that the law ought to be recast into a set of deductive axioms that would lead to consistent outcomes in any possible factual situation. This reformulation of the law would eliminate normative conflicts and therefore would leave no room for legal defeasibility. This idea has been affirmed by Alchourrón and

Bulygin (1971): the legislator and the doctrinal jurist should combine their efforts towards providing axiomatic reformulation of the law, or at least of particular sections of it. Just as Euclid developed an axiomatic model of geometry, and as modern natural science and social science (especially economics) have developed axiomatic models for their theories, so the legislator and the jurist should axiomatise the law. By adding to such an axiomatisation a description of a specific case, we should obtain a set of premises from which the obligations and entitlements of the parties in the case can be deduced.

Alchourrón (1996a, b) claimed that the ideal of the axiomatisation of the law should inspire legislation and doctrine. It could contribute to bringing legal studies and scientific method together: just as in science the phenomena to be explained, the *explanandum*, should be the logical consequences of a set of premises, the *explanans*, containing scientific laws and the description of particular facts, so in law the content of a legal conclusion (the decision) should be the deductive consequence of a set of premises including both general norms and the description of specific facts. Systemic interpretation should have the task of making exceptions explicit, by embedding their negation into the antecedent of the concerned legal norm (a prima facie norm "if φ then ψ," which is subject to exception χ, should be rewritten as "if φ and not χ then ψ").

It seems to me that even if such a reformulation of the law were feasible (with regard to all exceptions that could be identified by legal scholars), it is doubtful that it would be useful, i.e., that it would make the law easier to understand and apply. Legal prescriptions would need to become much more complex, since every rule would have to incorporate all its exceptions. In addition, such a representation of the law would not be able to model the dynamic adjustment that takes place—without modifying the wording of existing rules—whenever new information concerning the conflicting rules and the criteria for working out their conflicts is taken into consideration. Finally, by rejecting defeasible reasoning, we would forfeit the law's ability to provide provisional outcomes while legal inquiry moves on.

The need to represent the law in ways that facilitate defeasible reasoning does not imply that the current way of expressing legal regulations in statutes and regulatory instruments cannot be improved. On the contrary, considerable improvements in legislative technique are required to cope with the many tasks entrusted to modern legal systems. However, such improvements should not be aimed at producing a conflict-free set of legal rules, just for the sake of logical consistency. They should rather be aimed at producing legal texts that can more easily be understood and applied. This objective requires skilful use of the very knowledge structures (such as conceptual hierarchies, speciality, or the combination of rules and exceptions) that enable defeasible reasoning.

Accepting defeasibility in the law has significant implications both for the way we use legal knowledge and for the structure of such knowledge. On the one hand, deductive inference can be complemented with defeasible arguments. On the other hand, the acceptance of defeasibility leads us to view the law as an argumentation basis containing conflicting pieces of information as well as the criteria for resolving some of these conflicts. It is important to stress the difference between an argumentation basis and a deductive axiomatic base. While a deductive axiomatic base is consistent and flat, an argumentation basis is conflictive and possibly hierarchical: it includes reasons clashing against one another, reasons for preferring one reason to other reasons, and reasons for applying or not applying certain reasons given particular conditions.

Both strategies just mentioned—representing the law as an axiomatic base and representing it as an argumentation basis—may be justified in different contexts. The first strategy may be appropriate when we want to deepen our analysis of a small set of norms and anticipate as much as possible all instances of their application, finding a precise solution for each of them. The second strategy, however, more directly corresponds to the logical structure of nonformalised legal language (which expresses the law as setting out rules and exceptions, principles, preference criteria, etc.), and it reflects the ways in which legal reasoning proceeds when dealing conflicting pieces of information: rules and exceptions, different values needing to be balanced, different norms implementing different values, competing standards indicating what norms and values ought to prevail in case of conflict, and so on.

An argumentation basis may be transformed into an axiomatic knowledge base whose deductive conclusions include all outcomes that would be defeasibly justified relative to the given argumentation basis (assuming that all the facts of the case are known). The dialectical interaction between reasons for and against certain conclusions, and between grounds for preferring one argument to another, would be transformed into a set of conclusive connections between legal preconditions and legal consequences. Flattening legal information in this way, however, would entail a loss of information: the deductive knowledge base would not include a memory of the choices from which it derives, and therefore, it would not contain the information needed to reconsider such choices—it would not, for example, contain the information on which it was decided that a certain principle would outweigh a competing principle or that a certain interpretation was preferable. To understand the articulation of the relevant legal reasons, we would need to go back to the original argumentation basis.

Consider, for instance, the domain of privacy. Under EU regulation law, the processing of personal data is admissible only for a specific purpose that is communicated to the person concerned. Moreover, such processing is in general admissible only with that person's consent. These constraints are justified by the need to protect values such as individual self-determination and dignity. However, there is a large set of exceptions to the consent principle, namely different scenarios in which data can be processed without consent. These exceptions are justified by the need to protect the competing rights of others, as well as certain social values. Moreover, we have cases where consent alone is insufficient to make data processing permissible, further requirements being necessary (like the authorisation of a data protection authority for genetic data), and for each such exception specific rationales can be found that guide interpreters in determining the contents and limits of the exception. Finally, there may be cases where personal data may be processed even beyond the explicitly stated legislative scenarios, on the basis of an authorisation which a data protection authority issues to protect the rights of others, but which overrides the right to privacy. To determine whether a data protection authority has made legitimate use of its powers, we need to consider the importance of the values at stake (privacy, freedom of expression, economic freedom, health, etc.) and evaluate whether they have been balanced in a way that respects legal (in particular, constitutional) constraints. We could try to reduce this multilevel argumentation basis to a set of flat rules, but what we would obtain is a representation removed from the original legal texts (laws, regulations, authorisations), and whose contents and rationales are much more difficult to grasp.

16 The Emergence of the Idea of Defeasibility in Law and Ethics

Though formal logics for defeasible reasoning have been developed only recently, we can find references to defeasibility in the history of philosophical and legal reasoning.

A famous fragment by Aristotle apparently characterises legal reasoning as defeasible (Aristotle, *Nicomachean Ethics*, 1137b), in the sense that legal conclusions derived from general norms may have to be rejected in the face of particular cases having exceptional features that make those conclusions inadequate:

> All law is universal, and there are some things about which it is not possible to pronounce rightly in general terms; therefore, in cases where it is necessary to make a general pronouncement, but impossible to do so rightly, the law takes account of the majority of cases, though not unaware that in this way errors are made. And the law is none the less right; because the error lies not in the law nor in the legislator, but in the nature of the case, for the raw material of human behaviour is essentially of this kind. So, when the law states a general rule, and a case arises under this that is exceptional, then it is right, where the legislator, owing to the generality of his language, has erred in not covering that case, to correct the omission by a ruling such as the legislator himself would have given if he had been present there, and as he would have enacted if he had been aware of the circumstances. (Aristotle, *Nicomachean Ethics*, 1137b)

Cicero distinguishes presumptive (probabilis) and necessary argumentation (Cicero, *De inventione*, Book 1, Section 44). He provides various patterns (warrants) for presumptive inferences: the (natural) meaning of a sign (e.g. blood traces indicate participation in a violent action), what happen usually (e.g. mothers love their children), common opinion (e.g. philosophers are atheists), or similarity (if it is not discreditable to the Rodians to lease their port-dues, then it is not discreditable even to Hermacreon to rent them). Moreover, he considers how (defeasible) arguments may be refuted:

> All argumentation is refuted when one or more of its assumptions is non granted, or when, the assumptions having been granted, it is denied that the conclusion follows from them, or when it is shown that the kind itself of the argumentation is faulty, or when against a strong argumentation another argumentation equally strong or stronger is put forward (Cicero, *De inventione*, Book 1, Section 79).

The second and the fourth items in Cicero's list seem to correspond to what we called undercutting and rebutting, respectively, namely those attacks that are peculiar to defeasible arguments.

The Aristotelian approach to the dialectics of rule and exception is developed by Aquinas:

> [I]t is right and true for all to act according to reason: And from this principle it follows as a proper conclusion, that goods entrusted to another should be restored to their owner. Now this is true for the majority of cases: But it may happen in a particular case that it would be injurious, and therefore unreasonable, to restore goods held in trust; for instance, if they are claimed for the purpose of fighting against one's country. And this principle will be found to fail the more, according as we descend further into detail, e.g., if one were to say that goods held in trust should be restored with such and such a guarantee, or in such and such a way; because the greater the number of conditions added, the greater the number of ways in which the principle may fail, so that it be not right to restore or not to restore. (Aquinas 1947, *Summa Theologiae*, I–II, q. 94, a. 4)

The idea of defeasibility in the legal domain is precisely outlined by G. W. Leibniz, who characterises legal presumption as defeasible inference, arguing that in presumptions

> the proposed statement necessarily follows from what is established as true, without any other requirements than negative ones, namely, that there should exist no impediment. Therefore, it is always to be decided in favor of the party who has the presumption unless the other party proves the contrary. (Leibniz 1923, De Legum Interpretatione, A VI iv C 2789)

Leibniz argues that all laws are defeasible: legal norms support presumptive conclusions, which are subject to exceptions established by other norms. He also points at the connection between defeasibility and burden of proof:

> every law has a presumption, and applies in any given case, unless it is proved that some impediment or contradiction has emerged, which would generate an exception extracted from another law. But in that case the charge of proof is transferred to the person who adduces the exception. (Leibniz 1923, De Legum Interpretatione, A VI iv C 2791)

Turning from law to morality, we can find a notion of defeasibility in the work of David Ross, an outstanding Aristotelian scholar and moral philosopher who developed a famous theory of prima facie moral obligations (Ross 1930, 1939). Espousing a pluralist form of moral intuitionism, Ross relates defeasibility to the possibility that, in concrete cases, moral principles may be overridden by other moral principles:

> Moral intuitions are not principles by the immediate application of which our duty in particular circumstances can be deduced. They state […] prima facie obligations. […] [We] are not obliged to do that which is only prima facie obligatory. We are only bound to do that act whose prima facie obligatoriness in those respects in which it is prima facie obligatory most outweighs its prima facie disobligatoriness in those aspects in which it is prima facie disobligatory. (Ross 1939, 84–85).

Ross links the notion of defeasibility to the idea of outweighing, a key notion in reason-based approaches to practical reasoning (see Bongiovanni, Chap. 1, part I, this volume, on "Reasons (and Reasons in Philosophy of Law)." The idea that moral reasoning consists in balancing reasons and the idea of defeasibility are indeed connected, under the assumption that we can legitimately make moral assessments also on the basis of partial knowledge of the situations we face, i.e. even when we are not guaranteed to have taken into account all relevant reasons. The fact that certain reasons support a certain action only provides a defeasible support to that action: these reasons justify that action in the absence of outweighing reasons to the contrary, but would fail to support the outcome in the presence of the latter reasons. Consequently, if we believe that the reasons justifying the actions are present and we are not aware of reasons to the contrary, we should conclude that the action is presumably justified (on the basis of the information we have). If we come to believe that outweighing reasons are present, we should withdraw this conclusion.

Indeed, defeasibility may make the appeal to general ethical principles compatible with the particularistic view that any moral principle or reason may be overridden or be inapplicable depending on the circumstances (Dancy 2004). As Horty (2007, 2012) has argued, moral principles should be viewed as defaults, that link reasons to actions (or to obligations to act), and support such actions as long as they are not rebutted by reasons to the contrary or undercut by reasons against their application. Logics for defeasible reasoning provide formal accounts of the view that practical reasoning consists in the assessment of competing reasons for action by a bounded cogniser.

Although the notion of defeasibility is quite familiar in legal practice and in doctrinal work, it was not extensively discussed and analysed in legal theory until recently. This notion was brought to the attention of legal theorists by H. L. A. Hart (1951, 152):

> When the student has learnt that in English law there are positive conditions required for the existence of a valid contract, [. . .] he has still to learn what can defeat a claim that there is a valid contract, even though all these conditions are satisfied. The student has still to learn what can follow on the word "unless," which should accompany the statement of these conditions. This characteristic of legal concepts is one for which no word exists in ordinary English. [. . .] [T]he law has a word which with some hesitation I borrow and extend: This is the word "defeasible," used of a legal interest in property which is subject to termination of "defeat" in a number of different contingencies but remains intact if no such contingencies mature.

References to the defeasibility of legal arguments can be found in important approaches to legal reasoning. For instance, Viehweg (1965) argued that lawyers approach specific problem situations, not by reasoning from a complete and consistent system of universal axioms, but by referring to an open, unordered, inconsistent, undetermined list of *topoi* (points of view, usually expressed as maxims) addressing the relevant features of the different situations that come up. Such *topoi* are usually defeasible, since they may fail to apply under particular situations. Consider, for instance, the legal *topos* that nobody can transfer to another person more rights than those he or she possesses (*nemo plus juris in alium transferre potest quam ipse habet*). This rule does not apply to some exceptional cases in which a buyer in good faith acquires property from an apparent seller that is not the actual owner.

Similarly, Perelman and Obrechts-Tyteca (1969) focused on the distinction between deductive demonstration and argumentation, and affirmed that, contrary to demonstration, argumentation is always in principle open to challenge or reconsideration (see Blair 2012, 127).

17 The Idea of Defeasibility in Logic and AI

Logic and artificial intelligence have played a key role in providing a precise analysis of defeasibility (see Ginzberg 1987 for a collection of seminal contributions on nonmonotonic reasoning, Horty 2001 and Koon 2009 for a discussion of nonmonotonic logics, and Blair 2012, Chap. 9, on defeasibility in the context of argumentation theories).

Pollock (2010) observes that Chisholm (1957) was the first epistemologist to use the term *defeasible*, taking it from Hart (1951). Among the philosophers who have addressed aspects of defeasibility is Stephen Toulmin, whose approach to reasoning is based on the idea that inference rules or warrants connect data and conclusions of arguments. In the following passage, he claims that some of these warrants are defeasible:

> Warrants are of different kinds, and may confer different degrees of force on the conclusions they justify. Some warrants authorise us to accept a claim unequivocally, given the appropriate data […]; others authorise us to make the step from data to conclusion either tentatively, or else subject to conditions, exceptions, or qualifications (Toulmin 1958, 100).

According to Toulmin, defeasibility has a special place in the law:

> Again, it is often necessary in the law-courts, not just to appeal to a given statute or common-law doctrine, but to discuss explicitly the extent to which this particular law fits the case under consideration, whether it must inevitably be applied in this particular case, or whether special facts may make the case an exception to the rule or one in which the law can be applied only subject to certain qualifications (Toulmin 1958, 101).

Defeasibility is also addressed by Nicholas Rescher, who deals with it in connection with dialectics (Rescher 1977) and presumptive reasoning (Rescher 2006). Rescher (1977, 6) describes defaults as "provisoed assertions," having the logical form P/Q and meaning that:

> "*P* generally (or usually or ordinarily) obtains provided that *Q*" or "*P* obtains, other things being equal, when *Q* does" or "when *Q*, so ceteris paribus does *P*" or "*P* obtains in all (or most) ordinary circumstances (or possible worlds) when *Q* does" or "*Q* constitutes prima facie evidence for *P*."

The assertion of *P* under proviso *Q*, combined with the assertion of *Q*, constitutes an argument for *P*, though *Q* does not "entail, imply or ensure P," but makes *Q* only "normal, natural, and only to be expected" (Rescher 1977, 7).

The most influential and comprehensive model of defeasibility is the one provided by John Pollock, who as noted introduced the ideas of undercutting and rebutting, as well as the technique of labelling defeasible inference graphs to determine their justification status (see Pollock 1995, 2010).

Particularly influential in contemporary research on informal logic has been the account of defeasible reasoning provided by Doug Walton. According to Walton (1996, 42–43)

> presumptive reasoning is neither deductive nor inductive in nature, but represents a third distinct type of reasoning of the kind classified by Rescher (1976) as plausible reasoning, an inherently tentative kind of reasoning subject to defeat by special circumstances (not defined inductively or statistically) or a particular case. (pp. 42–43)

Walton et al. (2008) identify a number of distinct argumentation patterns, called argument schemes, each of which can be challenged by appropriate critical questions acting as pointers to possible defeaters.

In artificial intelligence and logic, some formal approaches have been developed to capture the normality assumption embedded in defeasible reasoning: things are assumed to be normal unless we have evidence to the contrary. This assumption can be modelled by minimising the extension of predicates that express abnormality conditions (McCarthy 1980). A similar idea underlies negation by failure, used in logic programming: atomic propositions are assumed to be false unless they can be shown to be true (Clark 1978). Preferential defeasible logics (see Kraus, Lehmann, and Magidor 1990) are based on the idea that the defeasible implications of a set of premises are those propositions that are true in the most normal models (situations) that satisfy those formulas.

The idea of defeasible reasoning as the application of default inference rules supporting nondeductive presumptive inferences has been developed by Reiter (1980). An elegant and broadly scoped model of reasoning with defaults, meant to capture the link between reasons and the conclusions they favour, has recently been proposed by Horty (2007, 2012).

A large amount of AI research has been recently developed which merges defeasible reasoning and argumentation (Rahwan and Simari 2009). In particular, the abstract account of argumentation proposed by Dung (1995) has been very influential. Its abstractness lies in the fact that it focuses on attack (defeat) relations between arguments without considering these arguments' internal structure.

18 Defeasibility in Research on AI and Law

In AI and law, defeasible reasoning has been the subject of much research starting from the end of the 1980s. Much of this work focuses on defeasible argumentation (for a survey, see Prakken and Sartor 2015). The possibility of using negation by failure to model defeasible reasoning and burdens of proof in the law was suggested by Sergot et al. (1986). The issue of defeasibility in legal reasoning was first identified by Gordon (1988, 1995), who later developed the Carneades system into a computable framework for defeasible reasoning (Gordon et al. 2007).

Hage (1997) proposed the idea of rule application as a general pattern for defeasible reason, where rules deliver their consequences only when they are shown to be both valid and applicable (applicability meaning, in Hage's terminology, that the rule's antecedent conditions are satisfied). In his framework, a legal rule works as an exclusionary reason, such that arguments applying the rule defeat arguments based on excluded reasons (but they may be defeated by arguments based on other reasons).

Prakken and Sartor developed the first model of defeasible reasoning in law which includes reasoning with (defeasible) rules and with priorities among such rules (Prakken and Sartor 1996). The model has been extended to cover the burden of proof (2009) and has been applied to various aspects of legal reasoning, such as reasoning with precedents (Prakken and Sartor 1998). Prakken has further developed the idea of prioritised argumentation in several technical contributions (Prakken 2010; Modgil and Prakken 2010).

The idea of legal reasoning as defeasible argumentation has also been developed by Loui and Norman (2005), who have analysed the way a single defeasible legal inference may result from the compression of various inference steps, and may be attacked by unpacking it and addressing these steps.

Bench Capon (2003) has developed the idea of value-based argumentation, namely the idea that preferences between arguments are determined by the values endorsed by the audience to which the arguments are directed. Bench-Capon and Sartor (2003) have studied how alternative defeasible theories (sets of premises) can be constructed to explain cases and how they may be prioritised.

Governatori et al. (2004) have shown how defeasible argumentation can be captured by using defeasible logic, in the manner originally proposed by Nute (1994). Extensions of defeasible logic have been used to capture different aspects of legal reasoning, such as the timing of legal effects (Governatori et al. 2005) and changes in the law (Governatori and Rotolo 2010).

Finally, I should mention the rich research line on the use of defeasible legal argumentation in the evidence domain and its connections with other approaches to evidence (see Verheij et al. 2016).

19 Defeasibility in Legal Theory

The idea of defeasibility remains highly controversial, as evidenced by the contributions contained in a recent collection (Ferrer Beltran and Ratti 2012a).

Carlos Alchourron, a leading legal logician, has opposed the ideal of defeasible reasoning, arguing for a combination of systematic interpretation and deduction: systematic interpretation should merge rules and exceptions into a coherent whole to which deduction could be applied (see Alchourron 1996a, b). Other legal theorists, such as Alexander Peczenik (Peczenik 2005, 115ff.; Hage and Peczenik 2000) and Neil MacCormick (1995), have on the contrary argued that defeasibility plays a significant role in legal reasoning (see also Brożek 2004).

It is no easy task to review the legal theorists' approaches to defeasibility, since such theorists have advanced different understandings of defeasibility, which often do not comport with the idea of defeasibility as nonmonotonic reasoning. Brożek (2014) has pointed out different ways in which defeasibility is understood by the authors of different contributions in Ferrer Beltran and Ratti (2012a):

> Ferrer Beltran and Ratti consider, inter alia, the following formulation: "a norm is defeasible when it has the disposition not to be applied even though it is indeed applicable" (Ferrer Beltran and Ratti 2012b: 31). Frederick Schauer, in turn, claims that "the key idea of defeasibility [...] is the potential for some applier, interpreter, or enforcer of a rule to make an ad hoc or spur-of-the-moment adaptation in order to avoid a suboptimal, inefficient, unfair, unjust, or otherwise unacceptable, rule-generated outcome," and concludes that "defeasibility is not a property of rules at all, but rather a characteristic of how some decision-making system will choose to treat its rules" (Schauer 2012, 81 and 87). Jorge L. Rodríguez says that "when we express a conditional assertion, we assume the circumstances are normal, but admit that under abnormal circumstances the assertion may become false," and—transferring this characteristic of defeasibility into the domain of law—claims that "legal rules [are defeasible since they] specify only contributory, yet not sufficient, conditions to derive the normative consequences fixed by legal system" (Rodríguez 2012, 88). [...] Finally, Riccardo Guastini claims that legal rules are defeasible since "there are fact situations which defeat the rule although they are in no way expressly stated by normative authorities in such a way that the legal obligation settled by the rule does not hold anymore." (Guastini 2012, 183)

All the foregoing formulations point to interesting aspects of legal reasoning and to the practice of defeasible reasoning in the law. I would argue, however, that they fail to provide convincing redefinitions or clarifications of the notion of defeasibility. I have argued that defeasibility applies to three objects:

- *Arguments*. A defeasible argument is an internally valid argument that may be defeated by counterarguments that do not challenge the argument's premises but rebut its conclusions or undercut the link between its premises and its conclusion.
- *Inference*. A defeasible inference is nonmonotonic, in the sense that it makes it possible to derive conclusions that may no longer be derivable if additional premises are added.
- *Conditionals*. A conditional is defeasible when it has the logical structure of a default, i.e. when it links a merely presumptive (nonconclusive) consequent to its antecedent.

These three aspects are different faces of the same issue. A defeasible argument $\mathcal{A}$ consists in a nonmonotonic inference: if we expand the argumentation basis from which $\mathcal{A}$ is constructed with premises that enable the construction of a defeater $\mathcal{B}$ to $\mathcal{A}$, the conclusion of $\mathcal{A}$ will be no longer justified relative to the expanded argumentation basis, and in this sense, no longer derivable from it. Correspondingly, default conditionals make it possible to construct defeasible arguments, i.e. nonmonotonic inferences: the results obtained through defeasible modus ponens can be defeated by rebutters or undercutters.

According to this idea of defeasibility, a legal norm can be said to be *defeasible* whenever all the following conditions are jointly possible:

- The norm is accepted (being valid and being generally applicable in the spatio-temporal domain under consideration).
- The norm's antecedent is also accepted.
- The norm's consequent is rejected.

As we have seen in the examples above (for instance, in Fig. 8), a defeasible norm N can be modelled as a default, i.e. in the logical form "if $P\ (\boldsymbol{x})$, then presumably $Q\ (\boldsymbol{x})$," i.e. as $N\ (\boldsymbol{x})\ :\ P\ (\boldsymbol{x})\ \Rightarrow\ Q\ (\boldsymbol{x})$, where $\boldsymbol{x}$ is the list of the variables in the norm (and the default would stand for the set of its ground instances). In fact, the inferences (arguments) warranted by that norm—i.e. arguments having the form $(P(\boldsymbol{a}),\ N\ (\boldsymbol{x})\ :\ P\ (\boldsymbol{x})\ \Rightarrow\ Q(\boldsymbol{x}),\ therefore\ Q\ (\boldsymbol{a}))$ where **a** is an individual case, namely a list of values for variables $\boldsymbol{x}$—can be rejected, given appropriate conditions, without rejecting the norm or is antecedent. This would happen whenever the premise for building a rebutter (a stronger norm having the form $N_1\ (\boldsymbol{x})\ :\ R(\boldsymbol{x})\ \Rightarrow\ \neg Q\ (\boldsymbol{x})$ or an undercutter (a norm having the form $N_2\ (\boldsymbol{x})\ :\ R\ (\boldsymbol{x})\ \Rightarrow\ \neg N\ (\boldsymbol{x})$) is available. This notion of the defeasibility also applies to the more abstract view of the application of a norm as involving a metalevel warrant such as "If the norm 'If $N\ (\boldsymbol{x})\ :\ P\ (\boldsymbol{x})$, then presumably $Q\ (\boldsymbol{x})$' is valid, and $P\ (\boldsymbol{a})$ is the case, then presumably $Q\ (\boldsymbol{a})$" (as in the model proposed by Hage 1997). In the following, I will speak of argument warranted by a norm, to cover both models of norm-based reasoning.

If the coexistence of the three conditions above is impossible, then the norm N at issue can be said to be *strict* or *indefeasible* and can be modelled as a material conditional, having the form: "for all $\boldsymbol{x}$, if $P\ (x)$, then $Q\ (x)$," i.e. $\forall x\ (P\ (x)\ \rightarrow\ Q\ (x))$. More plausibly an indefeasible norm N could be represented through a universal strict conditional "for all x, if $P(x)$, then necessarily $Q\ (x)$," i.e. $\forall x\ (P\ (x)\ \twoheadrightarrow\ Q\ (x))$ (or possibly as a strict rule, which may not allowing for contraposition, see Prakken 2010). The strict conditional, which is here denoted with the arrow $\twoheadrightarrow$, expresses the idea that the correlation between $P\ (x)$ and $Q\ (x)$ does not depend on the present factual situation (on the actual world), but would rather hold in every possible factual situation (the norm being unchanged). If we accept the indefeasible norm N and also accept that its antecedent holds in any possible context, we must accept that also the norm's consequent holds in that context. N could not be the object of exceptions in a strict sense, namely, of provisions stating that the unmodified norm does not apply when an impeding circumstance $E(\boldsymbol{a})$ is established. To avoid the effect of N to be triggered in circumstance $E(\boldsymbol{a})$, we would have to substitute it with the new norm $\forall x(P(x)\ \wedge \neg E(x)) \twoheadrightarrow Q(x)$. Because of this change, the norm's effect could be established in a case only when both predicates P and $\neg E$ are established in that case. Rather than E being an impeditive fact capable of blocking the application of the norm, $\neg E$ would become a negative constitutive fact that must be established, for that effect to be triggered.

Let us consider for instance a norm linking the causation of harm to the obligation to compensate the victim. If the norm were defeasible, it would mean that if any individual x culpably harms another individual y, then presumably x must compensate y, and it could be modelled in the logical form: $N(x, y)\ :\ CulpablyHarms(x, y)\ \Rightarrow\ MustCompensate(x, y)$. If the norm were indefeasible, it would rather mean that for all individuals x and y, if x harms y, then necessarily x has to compensate y, and it could be modelled in the logical form: $\forall x, y(CulpablyHarms(x, y) \twoheadrightarrow MustCompensate(x, y))$.

Thus, from the perspective here developed, the defeasibility of a norm pertains to its content, as expressible in its logical form, and therefore is not affected by the fact that the norm may be declared invalid: this may happen, under appropriate conditions, for both defeasible and indefeasible norms. Similarly, the defeasibility of a norm is not affected by the fact that the norm may be modified or substituted through judicial interpretation or through legislation. Both defeasible and indefeasible norms can be modified by new legislation or case law. The difference rather pertains to the necessity of a modification to introduce an exception:

- If exceptions to a norm can be introduced without changing the norm (without affecting its content or meaning), then the norm is defeasible, regardless of whether exceptions are expressed or implicit and whether they closed or open, and regardless of what authority and procedure that is needed for introducing exceptions (for an analysis of different kinds of exceptions, see Celano 2012). In particular, it is irrelevant to the defeasibility of a norm, whether exceptions to it can be introduced through judicial interpretation, or only through legislation (or through new constitutional norms).
- If the only way to legitimately exclude the application of a norm to cases having feature E consists in changing that norm, extending in its antecedent with the negation of E ($\neg E$), then the norm is indefeasible.

In their analysis of the notion of defeasibility in the law, Ferrer Beltran and Ratti (2012b, 36) distinguish three cases: (1) the norm's validity is defeasible, in the sense that it depends on defeasible criteria, (2) the norm is externally defeasible, in the sense that the "conditions of applications contain implicit exceptions whose scope has not been determined," and (3) the norm's normative content is defeasible in the sense that it specifies operative facts that are contributory conditions for the production of the norm's legal effect. In particular, they say that in the third case "the norm's antecedent contains implicit exceptions which may not be exhaustively identified." They also affirm that in cases (1) and (2), it is not the norm itself which is defeasible, but rather the criteria for its validity or application, while the norm should be represented as a material conditional.

As Ferrer Beltran and Ratti (2012b, 36) rightly observe, only their third case of defeasibility is really significant: the first two cases depend on metanorms on validity or application that are defeasible in the third sense, namely, according to their content. However, the way in which they describe defeasibility by content differs from the approach here adopted in three regards.

Firstly, it makes the implicitness of the exceptions to a norm a necessary condition for the defeasibility of the norm. On the contrary, I have argued that even explicit exceptions to a norm presuppose the defeasibility of that norm, since they give rise to the pattern that characterises defeasible argumentation: the absence of exceptions is not needed to construct an argument warranted by the norm, though that argument can be defeated by arguments warranted by the exception.

Secondly, the characterisation of the antecedent of a defeasible norm as providing contributory conditions for the norm's conclusion fails to capture the idea of defeasible connection between the antecedent and the conclusion of that norm. As I observed in Sect. 3, a contributory condition for a conclusion may fail to provide any presumptive support for that conclusion. This is the case when a norm has a conjunctive antecedent so that its application results in a linked argument (see Fig. 2). On the other hand, the antecedent of defeasible norm can be described as a contributory reason for the norm's conclusion. A genuine contributory reason for a legal conclusion should indeed provide on its own sufficient presumptive support to that conclusion; i.e. it should match the antecedent of a legal default and so enable the construction of a separate defeasible argument. Separate defeasible arguments sharing the same conclusion may contribute to a stronger convergent argument supporting the same conclusion (see Figs. 3 and 4).

Finally, it is not clear to me why the issue of determining whether a norm is defeasible or not should be specifically addressed as a "matter of interpretation." It is a matter that pertains to the determination of the logical structure of the norm at issue, an issue that could pertain to interpretation or not depending on how one understands the notion of interpretation, i.e. as concerning every ascription of meaning to a text, or only the ascription of meaning meant to address some doubts (see Dascal and Wroblewki 1986). Interpretation in the first sense obviously covers the determination of every aspect of the content of a norm having a textual source, and therefore, it also covers the determination of the norm's logical structure. In fact, to determine whether a norm is defeasible or not, we have to consider—depending on whether we are approaching the norm from a socio-legal or from doctrinal perspective—(*a*) the way in which the norm is comprehended and used by the community of its users (those who endorse/follow/apply it) or (*b*) the way in which the norm should correctly be comprehended and used by the same community. This determination would involve an empirical assessment according to (*a*) or a normative assessment according to (*b*). This assessment would be no less (and no more) dependent on interpretation than the determination of any other aspects of the norm's content, such as the structure and the components of its antecedent and its consequent. Finally, the issue of determining whether a legal norm is defeasible would be utterly trivial if we were to adopt—both empirically and normatively—the view that all legal norms are defeasible in this abstract sense—i.e. the assumption that no strict legal norm exists, as a matter of fact—following Leibniz's suggestion.

It seems to me that we need to distinguish clearly two aspects concerning a norm's defeasibility. The first aspect, to which we have referred in this contribution by using the term "defeasible," pertains the intrinsic logical structure of a norm: is the norm meant to establish a presumptive or a conclusive link between its antecedent and its conclusion? Defeasibility so understood is a counterfactual property: to say that a norm *N,* having antecedent *P* and conclusion *Q* (I leave the variables implicit, for simplicity's sake), is defeasible just means that it is in principle possible to reject an argument for *Q* warranted by *N*, while accepting both *N* and *P*: we can imagine a system $\mathbb{L}$ and a factual constellation $\mathbb{F}$, such that with regard to the argumentation basis $\mathbb{L} \cup \mathbb{F}$ both *N* (unchanged) and *P* are justified (e.g. being unchallenged), but the argument that delivers Q on the basis of N and P is rebutted or undercut, by arguments constructible from $\mathbb{L} \cup \mathbb{F}$.

Once we have determined that N is intrinsically defeasible, we can address further issues. One issue concerns determining whether N-warranted arguments can be defeated in the normative system $\mathbb{L}$ containing N, i.e. whether a rebutting or undercutting counterargument to an N-warranted argument can be mounted by using only norms in $\mathbb{L}$, plus appropriate operative facts (see Sect. 11). A different issue concerns determining whether N could be defeated given the possible (permitted or empowered) judicial modifications of the current legal system $\mathbb{L}$, namely whether judicial construction/interpretation could introduce in $\mathbb{L}$ new norms that enable the construction of defeaters against the application of N. Obviously answering either of these issues may or will require the interpretation of the legal system $\mathbb{L}$ under consideration (on the connection between the possibility that a norm is defeated and interpretation, see Duarte 2011, 135).

Finally, the idea of a norm's defeasibility as pertaining to the logical structure of that norm leads me to address one further claim by Ferrer and Ratti, namely the view that whenever a norm's validity or application is determined by defeasible metarules, the norm itself must be defeasible. Since a norm's defeasibility only concerns the logical structure of that norm, the fact that a norm is defeasible does not exclude (nor require) that its validity as well as the domain of its intended application are governed by defeasible criteria. Inapplicability rules, however, may presuppose the defeasibility of the norm that they address: a rule stating that norm N is inapplicable under exceptional circumstances E is usually meant enable the construction of undercutters to N-warranted arguments, namely arguments having the following form: E is the case, if E than N does not apply (does not warrant its conclusion), therefore N does not apply: E, $E \Rightarrow \neg N$, therefore $\neg N$ (see argument C in Fig. 8).

In conclusion, I think that, notwithstanding the multifarious creative ways in which legal theorists have framed the idea of defeasibility, it would better to stick to the more limited and precise concept on which other disciplines—such as logic, philosophy, and computing—converge, namely, the view that defeasible reasoning is nonmonotonic and that the antecedent of a defeasible norm provides only presumptive support to the norm's conclusion. The considerations that have been presented as alternative analyses of the concept of defeasibility should rather be rephrased are pertaining to the ways in which (a) arguments applying defeasible legal norms can be rebutted or undercut, or (b) existing legal norms, both defeasible or indefeasible ones, can be abrogated or modified.

20 Conclusion

Defeasible reasoning is a key aspect of legal reasoning and problem-solving. Therefore, theories and logics of defeasible can greatly contribute to the study of legal argumentation and legal justification.

Recognising the strength of the connection between defeasibility and the law does not require abandoning logical rigour. On the contrary, it favours adopting logical models that precisely match certain important structures of legal knowledge, certain frequent patterns of legal reasoning, and of the dialectics of legal interaction. Argument-based theories of defeasible reasoning provide the most advantageous approach to address defeasibility in legal contexts.

I have argued that legal theory should address defeasibility using a shared conceptual framework and focus with the other disciplines—in particular, logic and computing—which have addressed defeasible reasoning. This does not exclude that legal theory can provide useful contribution to the study of defeasibility. In fact, the law provides a rich set of structures and patterns for defeasible reasoning. Therefore, the analysis of patterns of defeasibility in the law can contribute not only to legal theory and (computable) legal logic, but also to the development of general theories and logical models of defeasibility.

References

Alchourrón, C.E. 1996a. Detachment and defeasibility in deontic logic. *Studia Logica* 57: 5–18.

Alchourrón, C.E. 1996b. On law and logic. *Ratio Juris* 9: 331–348.

Alchourrón, C.E., and E. Bulygin. 1971. *Normative systems*. Dordrecht: Springer.

Alchourrón, C.E., and D. Makinson. 1981. Hierarchies of regulations and their logic. In *New studies on deontic logic*, ed. R. Hilpinen, 123–148. Dordrecht: Reidel.

Alchourrón, C.E., P. Gärdenfors, and D. Makinson. 1985. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic* 50: 510–530.

Alexy, R. 2002. *A theory of constitutional rights*. Oxford: Oxford University Press.

Aquinas, T. 1947. *Summa Theologiae*. Ed. and trans. Fathers of the English Dominican Province. Allen, TX: Benzinger Bros.

Aristotle. 1954. *Nicomachean ethics*. Ed. and trans. W.D. Ross. Oxford: Oxford University Press.

Baroni, P., M. Caminada, and M. Giacomin. 2011. An introduction to argumentation semantics. *The Knowledge Engineering Review* 26: 365–410.

Bench-Capon, T.J.M., and H. Prakken. 2006. Justifying Actions by Accruing Arguments. In *Computational models of argument. Proceedings of COMMA 2006*, ed. P.E. Dunne, and T.J.M. Bench-Capon, 247–258. Amsterdam: IOS Press.

Bench-Capon, T.J.M. 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation* 13: 429–448.

Bench-Capon, T.J.M., and G. Sartor. 2003. A model of legal reasoning with cases incorporating theories and values. *Artificial Intelligence* 150: 97–142.

Blair, J.A. 2012. *Groundwork in the theory of argumentation*. Dordrecht: Springer.

Brewka, G. 1991. *Nonmonotonic reasoning: Logical foundations of commonsense*. Cambridge: Cambridge University Press.

Brewer, S. 1996. Exemplary reasoning: Semantics, pragmatics and the rational force of legal argument by analogy. *Harvard Law Review* 109: 923–1028.

Brewer, S. 2011. Logocratic method and the analysis of arguments in evidence. *Law, Probability and Risk* 10: 175–202.

Brożek, B. 2004. *Defeasibility of legal reasoning*. Kraków: Zakamycze.

Brożek, B. 2008. Revisability vs. Defeasibility. *Northern Ireland Legal Quarterly* 59: 139–147.

Brożek, B. 2014. Law and defeasibility: A few comments on the logic of legal requirements. *Revus* 23: 165–170.

Celano, B. 2012. True exceptions: Defeasibility and particularism. In *The logic of legal requirements*, ed. J. Ferrer Beltran, and G.B. Ratti, 268–287. Oxford: Oxford University Press.

Chisholm, R.M. 1957. *Perceiving: A philosophical study*. Ithaca, NY: Cornell University Press.

Cicero. 1965. De inventione (Rhetorici libri duo qui vocantur de inventione). Stutgardiae: In aedibus Teubneri.

Clark, K.L. 1978. Negation as failure. In *Logic and data bases*, ed. H. Gallaire, and J. Minker, 293–332. New York: Plenum.

Dancy, J. 2004. *Ethics without principles*. Oxford: Oxford University Press.

Dascal, M., and J. Wróblewski. 1988. Transparency and Doubt: Understanding and interpretation in pragmatics and in law. *Law and Philosophy* 7: 203–224.

Duarte, D. 2011. Linguistic objectivity in norm sentences: Alternatives in literal meaning. *Ratio Juris* 24: 112–139.

Duarte de Almeida, L. 2013. A proof-based account of legal exceptions. *Oxford Journal of Legal Studies* 1: 133–168.

Dung, P.M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and n–person games. *Artificial Intelligence* 77: 321–357.

Fenton, N., M. Neil, and D. Berger. 2016. Bayes and the law. *Annual Review of Statistics and Its Application* 3: 51–77.

Ferrer Beltran, J., and G.B. Ratti (eds.). 2012a. The logic of legal requirements: Essays on defeasibility. Oxford University Press.

Ferrer Beltran, J., and G.B. Ratti (eds.). 2012b. Defeasibility and legality: A survey. In *The logic of legal requirements: Essays on defeasibility*, eds. J. Ferrer Beltran, and G.B. Ratti, 11–38. Oxford: Oxford University Press.

Gazzo Castañeda, L.E., and M. Knauff. 2016. Defeasible reasoning with legal conditionals. *Memory and Cognition* 44: 499–517.

Gärdenfors, P. 1987. *Knowledge in flux*. Cambridge, MA: MIT Press.

Ginzberg, M.L. (ed.). 1987. *Readings in nonmonotonic reasoning*. Burlington, MA: Morgan Kaufmann.

Gordon, T.F. 1988. The importance of nonmonotonicity for legal reasoning. In *Expert systems in law: Impacts on legal theory and computer law*, ed. H. Fiedler, F. Haft, and R. Traunmüller, 111–126. Tübingen: Attempto.

Gordon, T.F. 1995. *The pleadings game. An artificial intelligence model of procedural justice*. Dordrecht: Kluwer.

Gordon, T.F., H. Prakken, and D.N. Walton. 2007. The Carneades model of argument and burden of proof. *Artificial Intelligence* 171: 875–896.

Governatori, G., M.J. Maher, D. Billington, and G. Antoniou. 2004. Argumentation semantics for defeasible logics. *Journal of Logic and Computation* 14: 675–702.

Governatori, G., A. Rotolo, and G. Sartor. 2005. Temporalised normative positions in defeasible logic. In *Proceedings of the tenth international conference on artificial intelligence and law (ICAIL 2005)*, 25–34. New York: ACM.

Governatori, G., and A. Rotolo. 2010. Changing legal systems: Legal abrogations and annulments in defeasible logic. *Logic Journal of IGPL* 18: 157–194.

Guastini, R. 2012. Defeasibility, axiological gaps, and interpretation. In *The logic of legal requirements*, ed. J. Ferrer Beltran, and G.B. Ratti, 182–192. Oxford: Oxford University Press.

Hage, J.C. 1997. *Reasoning with rules: An essay on legal reasoning and its underlying logic*. Dordrecht: Kluwer.

Hage, J.C. 2005. *Studies in legal logics*. Dordrecht: Springer.

Hage, J.C., and A. Peczenik. 2000. Law, morals and defeasibility. *Ratio Juris* 13: 305–325.

Hart, H.L.A. 1951. The ascription of responsibility and rights. In *Logic and language*, ed. A. Flew, 145–166. Oxford: Basil Blackwell.

Hitchcock, D. 2017. *On reasoning and argument: Essays in informal logic and on critical thinking*. Dordrecht: Springer.

Holland, J. 2012. *Signals and boundaries building blocks for complex adaptive systems*. Cambridge, MA: MIT Press.

Holland, J., K.J. Holyoak, R.E. Nisbett, and P.R. Thagard. 1989. *Induction. Processes of inference, learning and discovery*. Cambridge, MA: MIT Press.

Holyoak, K., and P. Thagard. 1996. *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press.

Horty, J. 2001. Nonmonotonic logic. In *The Blackwell guide to philosophical logic*, ed. L. Goble, 336–361. Oxford: Blackwell.

Horty, J.F. 2007. Defaults with priorities. *Journal of Philosophical Logic* 36: 367–413.

Horty, J.F. 2011. Rules and reasons in the theory of precedent. *Legal Theory* 10: 1–33.

Horty, J.F. 2012. *Reasons as defaults*. Oxford: Oxford University Press.

Hunter, A. 2013. A probabilistic approach to modelling uncertain logical arguments. *International Journal of Approximate Reasoning* 54: 47–81.

Kant, I. 1949. On a suppposed right to lie from altruistic motives. In *Critique of Practical Reason and Other Writings in Moral Philosophy*, ed. L., White Beck, 346–350. Chicago: University of Chicago Press.

Koons, R. 2017. Defeasible reasoning. In *The stanford encyclopedia of philosophy*, ed. E.N. Zalta. https://plato.stanford.edu/entries/reasoning-defeasible/.

Kraus, S., D. Lehmann, and M. Magidor 1990. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence* 44: 167–207.

Leibniz, G.W. 1923. De legum interpretatione, rationibus, applicatione, systemate. In *Sämtliche Schriften und Briefe*. Edited by the Academy of Sciences of Berlin. Series VI, vol. iv. Darmstadt, Leipzig, Berlin.

Loui, R.P., and J. Norman. 1995. Rationales and argument moves. *Artificial Intelligence and Law* 3: 159–189.

MacCormick, D.N. 1995. Defeasibility in law and logic. In *Informatics and the foundations of legal reasoning*, ed. Z. Bankowski, I. White, and U. Hahn, 99–117. Dordrecht: Kluwer.

Maranhao, J.S.A. 2013. Defeasibility, contributory conditionals, and refinement of legal systems. In *The logic of legal requirements*, ed. J. Ferrer Beltran, and G.B. Ratti, 53–76. Oxford: Oxford University Press.

McCarthy, J. 1980. Circumscription: A form of non-monotonic reasoning. *Artificial Intelligence* 13: 27–39.

Modgil, S., and H. Prakken. 2010. Reasoning about preferences in structured extended argumentation frameworks. In *Computational models of argument. Proceedings of COMMA 2010*, ed. P. Baroni, F. Cerutti, M. Giacomin, and G. Simari, 347–358. Amsterdam: IOS Press.

Nute, D. 1994. Defeasible logic. In *Handbook of logic in artificial intelligence and logic programming. Vol. 3: Nonmonotonic reasoning and uncertain reasoning*, ed. D.M. Gabbay, C.J. Hogger, and J.A. Robinson, 353–395. Oxford: Oxford University Press.

Peczenik, A. 2005. *Scientia Juris: Legal doctrine as knowledge of law and as a source of law*. Dordrecht: Springer.

Perelman, C., and L. Olbrechts-Tyteca. 1969. *The new rhetoric: A treatise on argumentation*. Notre Dame, IN: University of Notre Dame Press.

Pollock, J.L. 1995. *Cognitive carpentry: A blueprint for how to build a person*. Cambridge, MA: MIT Press.

Pollock, J.L. 1998. Perceiving and reasoning about a changing world. *Computational Intelligence* 14: 498–562.

Pollock, J. L. 2008. *Defeasible reasoning*. In J Reasoning: Studies of Human Inference and its Foundations, ed. E. Adler and L. J. Rips, 451–470. Cambridge University Press.

Pollock, J.L. 2010. Defeasible reasoning and degrees of justification. *Argument and Computation* 1: 7–22.

Popper, K.R. 1959. *The logic of scientific discovery*. London: Hutchinson.

Prakken, H. 1997. *Logical tools for modelling legal argument: A study of defeasible reasoning in law*. Dordrecht: Kluwer.

Prakken, H. 2005. A study of accrual of arguments, with applications to evidential reasoning. In *Proceedings of the tenth international conference on artificial intelligence and law (ICAIL 2005)*, 85–94. New York: ACM.

Prakken, H. 2010. An abstract framework for argumentation with structured arguments. *Argument and Computation* 1: 93–124.

Prakken, H., and G. Sartor. 1996. Rules about rules: Assessing conflicting arguments in legal reasoning. *Artificial Intelligence and Law* 4: 331–368.

Prakken, H., and G. Sartor. 1998. Modelling reasoning with precedents in a formal dialogue game. *Artificial Intelligence and Law* 6: 231–287.

Prakken, H., and G. Sartor. 2009. A logical analysis of burdens of proof. In *Legal evidence and proof: Statistics, stories, logic*, ed. H. Kaptein, H. Prakken, and B. Verheij, 223–253. Farnham: Ashgate.

Prakken, H., and G. Sartor. 2015. Law and logic: A review from an argumentation perspective. *Artificial Intelligence* 227: 214–245.

Prakken, H., and G.A.W. Vreeswijk. 2001. Logical systems for defeasible argumentation. In *Handbook of philosophical logic*, ed. D. Gabbay, and F. Günthner, 218–319. Dordrecht: Kluwer.

Rahwan, I., and G.R. Simari. 2009. *Argumentation in artificial intelligence*. Dordrecht: Springer.

Raz, J. 1985. Authority, law, and morality. *The Monist* 68: 295–323.

Rescher, N. 1977. *Dialectics: A controversy-oriented approach to the theory of knowledge.* Albany, NY: State University of New York Press.

Rescher, N. 2006. *Presumption and the practices of tentative cognition*. Cambridge: Cambridge University Press.

Reiter, R. 1980. Logic for default reasoning. *Artificial Intelligence* 13: 81–132.

Riveret, R., H. Prakken, A. Rotolo, and G. Sartor. 2008. Heuristics in argumentation: A game-theoretical investigation. In *Computational Models of Argument. Proceedings of COMMA-08*, 324–335. Amsterdam: IOS Press.

Riveret, R., A. Rotolo, and G. Sartor. 2012. Probabilistic rule-based argumentation for norm-governed learning agents. *Artificial intelligence and Law* 20: 383–420.

Rodriguez, J. 2012. Against defeasibility of legal rules. In *The logic of legal requirements*, ed. J. Ferrer Beltran, and G.B. Ratti, 89–107. Oxford: Oxford University Press.

Ross, W.D. 1930. *The right and the good*. Oxford: Clarendon.

Ross, W.D. 1939. *Foundations of ethics*. Oxford: Clarendon.

Russell, S.J., and P. Norvig 2010. *Artificial intelligence. A Modern approach*. Upper Saddle River, NJ: Prentice Hall.

Sartor, G. 1993. Defeasibility in legal reasoning. *Rechtstheorie* 24:281–316.

Sartor, G. 1994. A formal model of legal argumentation. *Ratio Juris* 7: 212–226.

Sartor, G. 2005. *Legal reasoning: A cognitive approach to the law*. Springer.

Sartor, G. 2013. The logic of proportionality: Reasoning with non-numerical magnitudes. *German Law Journal* 14: 1419–1457.

Sergot, M.J., F. Sadri, R.A. Kowalski, F. Kriwaczek, P. Hammond, and H. Cory. 1986. The British Nationality Act as a logic program. *Communications of the ACM* 29: 370–386.

Schauer, F.F. 2012. Is defeasibility an essential property of law? In *The logic of legal requirements*, ed. J. Ferrer Beltran, and G.B. Ratti, 77–88. Oxford: Oxford University Press.

Stone Sweet, A. 2004. *The judicial construction of Europe*. Oxford: Oxford University Press.

Toulmin, S. 1958. *The uses of argument*. Cambridge: Cambridge University Press.

Verheij, B., F. Bex, S. Timmer, C. Vlek, J.-J. Meyer, S. Renooij, and H. Prakken. 2016. Arguments, scenarios and probabilities: Connections between three normative frameworks for evidential reasoning. *Law, Probability and Risk* 15: 35–70.

Viehweg, T. 1965. *Topik und Jurisprudenz. Ein Beitrag zur rechtswissenschaflichen Grundlagenforschung*. Munich: Beck.

Walton, D.N. 1996. *Argumentation Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates.

Walton, D.N. 2006. *Fundamentals of critical argumentation*. Cambridge: Cambridge University Press.

Walton, D.N. 2008. *Informal Logic: A Pragmatic Approach*. Cambridge University Press.

Walton, D.N. 2013. *Methods of argumentation*. Cambridge University Press.

Walton, D.N., C. Reed, and F. Macagno. 2008. *Argumentation schemes*. Cambridge: Cambridge University Press.

Walton, D., and G. Sartor. 2013. Teleological justification of argumentation schemes. *Argumentation* 2: 111–142.

Walton, D., G. Sartor, and F. Macagno. 2016. An argumentation framework for contested cases of statutory interpretation. *Artificial Intelligence and Law* 24: 51–91.

Analogical Arguments

Bartosz Brożek

In this chapter, I inquire into the structure of analogical arguments. I begin by considering several historical meanings of *analogy*, understood as a semantic relation, an act of cognition, and a kind of argument. I then proceed to provide a general characterization of arguments from analogy and identify four essential aspects thereof: the problem situation, *prima facie* similarity, relevant similarity, and the solution. In the remaining part of the chapter, I analyze these aspects in more detail.

1 The Many Faces of Analogy

In ancient thought, two different phenomena were identified which would hugely influence the subsequent discussion of analogy (cf. Hochschild 2010, 4–10). The first was what Aristotle called *pros hen legetai*, i.e., the use of the same word to refer to distinct, but somehow related, things. His famous example is *healthy*, an adjective which may be predicated of an animal, its complexion, and its urine.[1] The second comes from Greek mathematics and is designated with the Greek word *analogia*; it is a comparison of ratios—A:B::C:D (A is to B as C is to D), and while it was initially used to capture quantitative relations, its application had soon extended to non-numerical comparisons. For example, in the *Historia Animalium,* Aristotle claims that feathers are to birds as scales are to reptiles.[2]

In the Middle Ages, the basic distinction between those two ancient concepts was preserved, but—characteristically—with some terminological shifts. Boethius, who in his commentary to Aristotle's *Categories* and in *De Arithmetica* considers both

[1] Cf. Aristotle (1960), *Metaphysics*, 1003a34ff.

[2] Aristotle (1991), *History of Animals*, 486b.

B. Brożek (✉)
Department for the Philosophy of Law and Legal Ethics, Jagiellonian University, Kraków, Poland
e-mail: bartosz.brozek@uj.edu.pl

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_13

pros hen legatai and *analogia*, translates the latter term as *proportionalitas*. In the subsequent development of medieval theories, the term *analogia* was also used, but with a different meaning: It signified a way of speaking which is not univocal, but also not purely equivocal; a paradigmatic example of such a mode of speaking is the Aristotelian *pros hen legatai*. In other words, the Greek *analogia* became the Latin *proportionalitas*, while the Greek *pros hen legatai* became the Latin *analogia* (cf. Hochschild 2010, 8).

The development of the medieval theories of analogy culminated in Cajetan's *De Nominum Analogia*, written in the beginning of the sixteenth century. Cajetan distinguishes between three forms (*modi*) of analogy: of inequality, of attribution, and of proportionality. The first type of analogy "occurs when things are called by a common name and concept, but the concept is shared or participated in unequally" (ibid., p. 11). An example is the term *body*, which is predicated (unequivocally) of inanimate objects, plants, animals, and humans, while in fact there is an order of superiority among bodies. The analogy of attribution, in turn, is what Aristotle called *pros hen legetai*, i.e., "the common name is used with different relations to some one term" (ibid.). Finally, the analogy of proportionality is what the Greeks called *analogia*—and Boethius *proportionalitas*. An example Cajetan provides is that of *seeing* as predicated of the eye and of the intellect: "just as understanding exhibits a thing to the soul, so seeing [exhibits a thing] to an animated body" (ibid., 12). Cajetan further claims that the analogy of proportionality takes either the form of a metaphor ("when the transferred word does not properly belong in the new context") or constitutes proportionality sensu stricto ("when the name properly belongs not only in its original context but also in that context to which it has been transferred") (ibid.).

The above short outline of the ancient and medieval theories (cf. Ashworth 2013) reveals that Aristotle and his followers concentrated on the *semantic* dimensions of analogy: All the three kinds of analogy identified by Cajetan determine when one can use the same word to refer to distinct, yet somehow similar, things. We call stones and humans "beings," because—as much as stones are much lower on the hierarchy of being than are humans—both kinds of entities do belong to the hierarchy. We are justified in calling a man, his urine, and a medicine he takes "healthy," since all those things remain in some relation to health: A man is a subject of health; his urine, a sign of health; and medicine, a cause of health. We can say that the intellect *sees* something, because there is a kind of structural similarity—or proportionality—between seeing through one's eyes and understanding something through the operations of the intellect. It is not to say (as we shall—*nomen omen*—see below) that the semantic approach to analogy was the only way ancient and medieval thinkers understood the concept. However, there is little doubt that analogy seen as a semantic relation was their primary subject of interest.

A different, and much later, tradition sees analogy as a special kind of cognition (cf. Holyoak 2012, 234–259). The recent history of this way of thinking dates back to the seminal works of Ch. Spearman; J. C. Raven; and Spearman's student, R. Cattell. Anyone familiar with IQ scales and tests (such as Raven's progressive matrices or Cattell's Culture Fair Intelligence Test) will quickly guess what kind of analogy they investigated. They postulated that the ability to reason analogically (in particular using proportional analogies) is a good indicator of one's fluid intelligence (cf. ibid., 236). Another line of research on analogical cognition is connected to linguistics and cognitive science.[3] It is primarily concerned with metaphors, but also touches upon analogical reasoning, especially when metaphors are understood as a kind of analogy. George Lakoff and his collaborators go as far as to suggest that metaphors are at the very center of human cognitive activities. They believe that concrete concepts—such as "to grasp" or "a tree"—are formed in the interactions of the body with the environment. On their basis, by means of metaphors, more abstract concepts are formed. Of course, metaphor is not understood as a poetic device here; it is understood as a mechanism for "understanding and experiencing one kind of thing in terms of another" (Lakoff and Núñez 2000, 5). For example, importance is understood in terms of size ("This is a big issue," "It's a small matter"), while difficulties are conceptualized as burdens ("I've got some light housework," "He's overburdened") (ibid., 41).

> Each such conceptual metaphor has the same structure. Each is a unidirectional mapping from entities in one conceptual domain to corresponding entities in another conceptual domain. As such, conceptual metaphors are part of our system of thought. Their primary function is to allow us to reason about relatively abstract domains using the inferential structure of relatively concrete domains (Ibid., 42).

In turn, Mary Hesse's work in the philosophy of science (Hesse 1966), as well as the contributions of Winston (1980), Gentner(1983), Kolodner (1993), and Holyoak and Thagard (Holyoak and Thagard 1995), have been very influential in the development of theories of analogy within the paradigm of knowledge representation, particularly important in computer science. They have led to a number of different lines of research, including so-called case-based reasoning and analogical problem-solving in various disciplines: the sciences, law, morality, etc. (cf. Pal and Shiu 2004). This approach treats analogy not only as a mode of direct cognition, but also as a way of reasoning. In this, it is strictly connected to another understanding of analogy construed as a kind of *argument*.

Analogical arguments were widely used, and reflected upon, in ancient philosophy. Aristotle did not address the problem of analogical argumentation in any systematic way. He did, however, identify a few argument forms which are connected to analogy. The first is *paradeigma*, or the argument from example (in *Rhetoric* he says: "Enthymemes based upon example are those which proceed from one or more similar cases, arrive at a general proposition, and then argue deductively to a particular inference"). The second is *homoiotes*, or the argument from likeness (in *Topics* he says: "Try to secure admissions by means of likeness; for such admissions are plausible, and the universal involved is less patent"). The third is *epagoge*, or induction. Fourth, and finally, Aristotle speaks also of *parabole* (parable); parables were used by orators in inductive or indirect proof as a means of demonstration and illustration (Bartha 2013). Those Aristotelian distinctions proved highly influential. In his little

[3] This line of research was hugely influenced by the work of George Lakoff. See Lakoff and Johnson (2003).

treatise *Topica*, for example, Cicero (1949) distinguishes among three types of analogical arguments: *inductio* (which is equivalent to *epagoge*), *exemplum* (which is the Aristotelian *paradeigma*), and *conlatio* (equivalent to *parabole*). However, there were also developments with regard to analogical arguments with no direct link to the Aristotelian approach. In particular, the tradition of Roman law, from its early days up to the nineteenth century, although sometimes influenced by philosophical theories, remained a relatively independent source of the reflection on analogical argumentation (see Ando 2015).

In what follows I will leave aside the semantic understanding of analogy, as well as analogical cognition. I will concentrate on analogical arguments. However, instead of inquiring into the history of the problem and reconstructing the positions of particular philosophers, I will try to provide a general framework for analyzing and utilizing such arguments.

2 The Architecture of Analogical Arguments

Let us begin our inquiry into the structure of analogical arguments with three different examples. My first example comes from everyday experience. Let's imagine you are visiting Munich for the first time. You are hungry, it is quite late, and you wonder whether any restaurants are still open. You recall that during your visit to Bamberg all restaurants closed at 10 P.M., and you conclude that the same must be true in Munich. Instead of looking for a restaurant, you go to bed hungry.

The second example is an actual case from American law, *Adams v. New Jersey Steamboat Co.* Adams, a passenger on a boat operated by the New Jersey Steamboat Co., had some money stolen from his stateroom, despite having his door locked and windows fastened. The question the court had to answer was whether the defendant was liable as an insurer, i.e., without proof of negligence. There was no explicit rule stating the criteria for the responsibility of steamboat's operators. There were, however, other cases pertaining to the liability of service providers. In such cases as *Pinkerton v. Woodward* it was assumed that innkeepers were liable as insurers for their guests' losses. On the other hand, in cases such as *Carpenter v. N.Y., N.H. H.R.R. Co.*, it was established that the operators of a berth in a sleeping car of a railroad company are liable only if negligent. There are analogies between steamboats and inns, as well as between steamboats and sleeping cars. The court considered both analogies and decided that the first one was more relevant, stating that the steamboat's operator is liable as an insurer (cf. Weinreb 2005, 41–45).

My third and final example comes from science. In 1932, James Chadwick suggested the existence of the neutron. This discovery completely changed the views pertaining to the composition of atomic nuclei. Until then, it was assumed that nuclei consist of protons and electrons; after Chadwick's discovery, it became clear that they are composed of protons and neutrons. Initially, the nucleus was modeled as a rigid body that exerts a fixed overall average force on impinging neutrons (the one-body model). However, this assumption soon turned out to be mistaken, especially in the face of the phenomenon of selective absorption:

> [When] a continuous energy spectrum of neutrons first impinges on a layer of element X, some neutrons are absorbed, some pass through. The latter hit a second layer, of element Y, in which further absorption can take place. This second absorption is much smaller if X and Y are the same substance than when they are different. [...] [T]hese observations were in conflict with the predictions of the one-body model according to which absorption should exhibit a smooth behavior, inverse proportionality to the neutron velocity, for any kind of absorbing nucleus (Pais 1991, 337)

In 1935, Niels Bohr resolved this problem by assuming that the nucleus is compound, and its behavior resembles that of a drop of liquid. Although this analogy was far from perfect, since "the dynamics of a true liquid drop is vastly different from that of nuclei" (ibid., 339), it was useful enough to further research in nuclear physics (e.g., it later served as a framework for explaining nuclear fission).[4]

The examples provided above suggest the following general structure of analogical arguments. What ignites analogical reasoning is a situation in which something is unknown. You are in Munich and do not know whether restaurants are open late into the night; you are a judge on the *Adams v. New Jersey Steamboat Co.* case and there is no rule directly applicable to the case; or you are a physicist in the mid-1930s and realize that the one-body model of the nucleus does not account for the observational and experimental data. Thus, in all those cases an analogical (or some other) argument is called for, since there is a *problem situation*, i.e., the current knowledge is insufficient to provide a direct answer to a practical or theoretical question.

The second element of an analogical argument is the identification of other cases (situations) which are *prima facie similar* to the problem situation. Importantly, these similar cases must have a definite solution. In our first example, the previous visit to Bamberg and restaurants' opening hours there constitute similar circumstances to those you find yourself in while visiting Munich. In *Adams v. New Jersey Steamboat Co.* the situation resembles both that of *Pinkerton v. Woodward* and that of *Carpenter v. N.Y., N.H. H.R.R. Co.* Finally, in 1935 Bohr saw some similarity between atomic nuclei and drops of liquid.

The third building block of an analogical argument is the establishment of *relevant similarity* between the problem situation and (one of) the *prima facie* similar case(s). Presumably, the fact that in Bamberg restaurants are not open late into the night constitutes a sufficient reason to conclude that the same holds for Munich. In *Adams v. New Jersey Steamboat Co.* the court decided that the situation of a steamboat operator is relevantly similar to that of an innkeeper. Finally, Bohr's observation that the nucleus may be modeled as a drop of liquid proved useful—and the similarity involved turned out to be relevant—since the new theory of the nucleus not only accounted for some unexplained phenomena, but also became a framework for further research in nuclear physics.

[4] My presentation of Bohr's model is somewhat simplified. For more details see ibid., 335–341.

The final element of any analogical argument is the formulation of a *solution* to the problem situation: to go to sleep instead of looking for an open restaurant, to treat *New Jersey Steamboat Co*. in the same way as *Pinkerton v. Woodward*, or to construct a model of the nucleus as if it was a drop of liquid.

Thus, I believe it is useful to distinguish four different aspects of any analogical argument:

(1) the problem situation
(2) *prima facie* similarity
(3) relevant similarity
(4) the solution

Below I will consider these elements in some more detail. While I will try to characterize analogical arguments in the most general, domain-independent way, I will do so against the backdrop of some theories of analogical argumentation developed in legal theory. This choice is justified by two facts. First, inquiries into the nature and structure of legal analogy have a very long and rich history, with no match in other disciplines. Second, the accounts of analogy developed in legal theory often have different philosophical assumptions, motivations, and goals. In other words, they constitute a diverse, multifaceted source of theoretical reflection on analogy. In order to take advantage of this diversity, I will consider below three conceptions of analogical reasoning in the law—Robert Alexy's (Alexy 2010; see also Brożek 2007, 154–157), Scott Brewer's (Brewer 1996), and Bram Roth and Bart Verheij's (Roth and Verheij 2004)—which belong to different legal-theoretic paradigms.

3 The Problem Situation

It is surprising that existing theories of analogical reasoning—at least in legal philosophy, which serves here as an illustration—pay very little attention to describing or reconstructing the situation which gives rise to analogical arguments. Usually, it is simply stated that one needs to make recourse to analogical arguments when there is no legal rule directly applicable to the case at hand (This is exactly what happens in *Adams v. New Jersey Steamboat Co*., where there is no rule or prior precedent governing the liability of steamboat operators). A slightly different description of the problem situation is given by Scott Brewer, who speaks of the "context of doubt," when "reasoners are […] faced with questions about the scope and applicability of a norm or set of norms, whether it be texts written in canonical form […] or norms not tied to any particular form of words. Typically, but not always, these questions arise because of vagueness in some of the terms or central concepts used to express the norms" (Brewer 1996, 980). Thus, existing accounts of the problem situation in analogical (legal) argumentation underscore the psychological and pragmatic aspects of such situations. It seems that a more precise characterization is called for.

What is a problem situation, then? Arguably, this is a case when there exists a question for which our current knowledge provides no answer, or the available answers are—in one way or another—deficient. Let us begin with the simplest scenario.[5] The question "Is it the case that A?" may be rendered logically as:

(Q1) $\mathbf{?}\{A, \neg A\}$

where **?** is the logical constant in the erotetic logic under consideration. This is a close-ended question, i.e., a question for which there exists a finite number of direct answers (a direct answer is "directly and precisely responsive to the question, giving neither more nor less information than what is called for": Belnap 1969, 124). Importantly, at least for some class of questions, it is possible to speak of questions being *evoked* by some sets of sentences.[6] Let $\mathbf{C_0}$ stand for a set of sentences representing one's knowledge. When a question Q is evoked by $\mathbf{C_0}$, then the following conditions are fulfilled (Wiśniewski 1995, 12):

(1) No direct answer to Q belongs to $\mathbf{C_0}$.
(2) No direct answer to Q is entailed by $\mathbf{C_0}$.
(3) If all the formulas in $\mathbf{C_0}$ are true, question Q must have a true direct answer.
(4) Each presupposition of Q is entailed by $\mathbf{C_0}$.

Let us go back to our first example: You are in Munich and wonder whether any restaurants are open late into the night. Thus, when OR stands for "some restaurants are open late into the night" and *m* is a constant (proper name) representing Munich, our question may be formalized as

(Q2) $\mathbf{?}\{\mathrm{OR}m, \neg\mathrm{OR}m\}$

To say that this question is evoked by one's knowledge ($\mathbf{C_0}$) is to say that (1) $\mathbf{C_0}$ contains neither OR*m* nor ¬OR*m*; (2) neither OR*m* nor ¬OR*m* are entailed by $\mathbf{C_0}$; (3) if all the sentences in $\mathbf{C_0}$ are true, then either OR*m* is true or ¬OR*m* is true (given that OR*m* and ¬OR*m* are contradictory, this condition is trivially fulfilled given that condition (4) is also fulfilled); and (4) $\mathbf{C_0}$ contains or entails all the presuppositions of OR*m*, e.g., the existential presupposition that there exists Munich or that there exist restaurants in Munich (Similar to this situation is the one in our second case, *Adams v. New Jersey Steamboat Co.*, since the question of that case was: Should New Jersey Steamboat Co. be liable as an insurer?).

Things become more complicated when the problem situation gives rise to an open-ended question, i.e., a question for which there is no finite list of determinate direct answers. Our third example may be reconstructed as involving such a question: What is the structure of the atomic nucleus if it exhibits such behavior as selective absorption? Formalized, the question looks as follows:

(Q3) $\mathbf{?}\{\mathrm{SSA}x\}$

where SSA*x* is a propositional function (a formal counterpart of the predicate "is a structure, which exhibits selective absorption").[7] A direct answer to such a question

[5] I am using Andrzej Wiśniewski's (1995) theory of questions.

[6] In addition to evocation, Wiśniewski (1995, 12–13) speaks also of the *generation* of questions by sets of sentences.

[7] I considerably simplify Wiśniewski's account of such questions. For a detailed presentation of his view, see Wiśniewski (1995, 70–101).

consists in identifying an object which satisfies the propositional function SSAx. It is therefore clear that—at least in the general case—open-ended questions have a potentially infinite number of direct answers. Let us observe, however, that Q3 is a question evoked by the knowledge ($\mathbf{C_0}$) of physicists in the mid-1930s, since (1) $\mathbf{C_0}$ does not contain a sentence SSAa or SSAb or SSAc… where a, b, c… are proper names of certain structures; (2) SSAa or SSAb or SSAc… are not entailed by $\mathbf{C_0}$; (3) if all sentences in $\mathbf{C_0}$ (such as the claim that there exist neutrons, that the phenomenon of selective absorption is a genuine one, etc.) are true, then Q3 must have a direct answer; and (4) $\mathbf{C_0}$ contains or entails all the presuppositions of Q3, e.g., the existential presupposition that there exist atomic nuclei. By contrast, the question of whether Niels Bohr suffers from hay fever,[8] just like the question whether the continuum hypothesis is true, is not evoked by $\mathbf{C_0}$, because even if all sentences belonging to $\mathbf{C_0}$ are true, it does not have to be true that Bohr suffers (or does not suffer) from hay fever, nor does the continuum hypothesis have to be true (or false).

4 *Prima Facie* Similarity

Once the problem situation has been identified, an analogical argumentation proceeds by pinpointing a case or set of cases which have definite solutions and which are *prima facie* similar to the problem situation. In the legal-theoretic literature, the usual approach is to say that two cases are similar if they share some features. However, the point is that—from a logical perspective—any two cases share infinitely many features. Therefore, to compare cases and establish that they are *prima facie* similar, one needs to take into account only *some* of the features they share. The features which enter such comparisons are usually called *factors*. For example, in *Adams v. New Jersey Steamboat Co.*, the factors may include the steamboat operator providing accommodation for its guests, but not the color of the boat or quality of the food served at the messdeck. Roth and Verheij (2004, 640) underscore that "in general it is disputable which [features] are factors and which are not. In other words, in the law it depends on a contingent choice which factors are taken into account when comparing cases. Among other things this choice will depend on the legal domain under consideration, such as dismissal or trade secret law." Thus, according to their view, it is the background knowledge that determines the set of factors, and there seems to be no universal algorithm to detect them.

Robert Alexy also believes that the comparison of two cases is realized by considering the features they share. He introduces the following scheme of analogical reasoning:

A1: In every case c_i, each case c_j may be adduced with the argument that c_i shares with c_j the features F_1, …, F_f, and that c_i, for that reason and because there are

[8] Werner Heisenberg discovered quantum mechanics when he was suffering from a severe form of hay fever. Cf. Pais (1991, 275).

reasons for the rule $F_1, \ldots, F_f \rightarrow Q$, ought to be treated, as c_j, to the effect that Q (Alexy 2010, 10).

Thus, in *Adams v. New Jersey Steamboat Co.* (case c_i) one may identify another case (c_j), e.g., *Pinkerton v. Woodward*, in which an innkeeper was held liable as an insurer for the guest's loss. Both cases have some features in common ($F_1, \ldots, F_f$), e.g., innkeepers and steamboat operators provide accommodation services, their guests have their own rooms, etc. Now, Alexy says that *when* there are reasons for a rule "If *x* provides accommodation services and offers guests their own rooms, then *x* is liable as an insurer," *then* our case c_i should be treated as the analogical case c_j to the effect that steamboat operators are liable in the same way innkeepers are, i.e., as insurers (let us observe that Alexy does not claim that the rule "If *x* provides accommodation services and offers guests their own rooms, then *x* is liable as an insurer" is valid, only that there "are reasons" for it). It must further be stressed that Alexy does not suggest that one should take into account all the features of the two cases being compared, but only those which constitute *reasons* for resolving the case in one way or another. In other words, Alexy also claims that *prima facie* similar cases are those which share *some* features with the problem situation, and his answer to the question Which features are to be considered? is: those which have normative significance, i.e., which may serve as reasons for arriving at a practical decision.

Scott Brewer, in turn, takes a different approach. He does not directly address the problem of *prima facie* similarity; however, his views with regard to this issue may be reconstructed once one takes into account what he says about the process of analogical reasoning. "In the context of doubt," i.e., when it is not clear what extension some predicate has or what the meaning of some text is, "having found or been confronted with several examples, the reasoner (say, a judge) seeks to 'discover' a rule-like sorting of these examples; I refer to the rule thus discovered as the 'analogy-warranting rule'" (Brewer 1996, 962), which is initially treated as provisory or hypothetical. Thus, Brewer effectively says that an argument by analogy proceeds by considering cases similar to the unresolved case at hand, and these are the cases which may serve to abduct (discover) a rule that may solve the problematic case. To put it differently: *Prima facie* similarity between two cases is a situation in which the cases are seemingly governed by the same rule.

The foregoing succinct review of how *prima facie* similarity is dealt with in legal-theoretic literature on analogy reveals that the problem is difficult to address in purely formal or algorithmic terms. While it is noted that only some solved cases are *prima facie* similar to the case at hand, the method of identifying them is usually not suggested; it is rather assumed that one's background knowledge may easily serve as the criterion in such a selection process (although, of course, some building blocks of the criterion are suggested, as when Alexy claims that only those features of the case should be taken into account which may constitute reasons for an appropriate action, or when Brewer claims that the cases that should interest us are those which enable abducting a rule governing the case at hand).

I do not claim to have a comprehensive account of *prima facie* similarity to offer; however, I believe that once one realizes what the starting point of the argument from analogy is—i.e., the problem situation characterized as above—it is possible to say more about *prima facie* similarity. Let us recall that the problem situation may be described as a set $\mathbf{C_0}$ (the knowledge one has in the given situation) which evokes a question Q. Thus, in the case of our first example, the question is whether there are some restaurants in Munich which are open late into the night:

(Q2) $\mathbf{?}\{\mathrm{OR}m, \neg\mathrm{OR}m\}$

while the knowledge $\mathbf{C_0}$ in this situation includes a number of statements such as that you are in Munich, there are restaurants in Munich, Munich is a German city, etc. Now, the question is what cases are *prima facie* similar to the Munich case. A natural answer is that these are the cases which involve the same (kind of) question. Your previous visit to Bamberg constitutes such a case, since you know that in Bamberg no restaurants are open late into the night, i.e., that $\neg\mathrm{OR}b$ holds, where b is the constant standing for the proper name "Bamberg." Similarly, in *Adams v. New Jersey Steamboat Co.*, the question is whether the company is liable as an insurer:

(Q4) $\mathbf{?}\{\mathrm{LI}c, \neg\mathrm{LI}c\}$

and it turns out that in American law there are two sets of cases which answer similar questions: those which concern the liability of an innkeeper held liable as an insurer ($\mathrm{LI}x$), and those pertaining to the liability of railroad companies ($\neg\mathrm{LI}x$). Finally, in the example concerning the atomic nuclei the question asks what is the structure of the atomic nucleus if it exhibits such behavior as selective absorption?—a question we have rendered formally as

(Q3) $\mathbf{?}\{\mathrm{SSA}x\}$

It turned out that one such structure is a drop of liquid.[9]

Thus, the following definition of *prima facie* similarity between cases may be coined: Given two cases, $\mathbf{C_x}$ and $\mathbf{C_y}$, of which $\mathbf{C_x}$ evokes an unanswered question $\mathrm{Q_x}$ of the form$\mathbf{?}\{\mathrm{A}a, \neg\mathrm{A}a\}$, $\mathbf{C_y}$ is *prima facie* similar to $\mathbf{C_x}$ if $\mathbf{C_y}$ contains the expression $\mathrm{A}x/u_i$ or the expression $\neg\mathrm{A}x/u_j$. This definition works only for problem situations which involve simple yes-or-no questions, but can easily be extended so that it includes more complex situations as well: Given two cases, $\mathbf{C_x}$ and $\mathbf{C_y}$, of which $\mathbf{C_x}$ evokes an unanswered question $\mathrm{Q_x}$ of the form$\mathbf{?}\{\mathrm{A_1}x_1, \ldots, x_i, \ldots, \mathrm{A_n}\, x_1, \ldots, x_i\}$, $\mathbf{C_y}$ is *prima facie* similar to $\mathbf{C_x}$ if $\mathbf{C_y}$ contains $\mathrm{A_1}x_1/u_1, \ldots, x_i/u_i$ or … or $\mathrm{A_2}\ \mathrm{A_1}x_1/u_1, \ldots, x_i/u_i$. Still, this new definition has some limitations. Imagine that during your visit to Berlin you realized that in the German capital some restaurants are open 24 h a day, which may be formalized as $\mathrm{O24}r$. According to our current definition, the case of Berlin's restaurants is not *prima facie* similar to the Munich case, since the former does not involve any answer to the question$\mathbf{?}\{\mathrm{OR}x, \neg\mathrm{OR}x\}$. However, if in some city, restaurants are open 24 h a day, they are open late into the night: $\forall x(\mathrm{O24}x \rightarrow \mathrm{OR}x)$, and hence—given $\mathrm{O24}r$—we also have $\mathrm{OR}r$. The question you answered in the case of Berlin is simply *stronger* (i.e., it conveys more information or eliminates more possibilities) than the question you need to answer in the case of Munich. Finally,

[9] It should once more be stressed that this is a grossly simplified account of Bohr's discovery.

let us consider still another situation: During your stay in Würzburg you came to the conclusion that the restaurants there are open late into the night on some days only: OSDw. Of course, it is true that if the restaurants in a city are open late into the night, they are open late into night on some days: $\forall x(\mathrm{OR}x \rightarrow \mathrm{OSD}x)$. It may be argued that the Würzburg case is also similar to the Munich case, although the question in the former case is *weaker* (i.e., the answer to it conveys less information or eliminates fewer possibilities) than in the latter. Both these insights may be encapsulated in the definition of *prima facie* similarity between cases in the following way:

Given two cases, $\mathbf{C_x}$ and $\mathbf{C_y}$, of which $\mathbf{C_x}$ evokes an unanswered question $\mathrm{Q_x}$ of the form**?**$\{\mathrm{A}_1x_1, \ldots, x_i, \ldots, \mathrm{A_n}\, x_1, \ldots, x_i)$, $\mathbf{C_y}$ is *prima facie* similar to $\mathbf{C_x}$ if:

(1) $\mathbf{C_y}$ contains the expression $\mathrm{A}_1x_1/u_1, \ldots, x_i/u_i$ or … or $\mathrm{A}_1x_1/u_1,\ldots, x_i/u_i$; or
(2) $\mathbf{C_y}$ contains the expression $\mathrm{B}x_1/u_1,\ldots, x_i/u_i$ such that $\forall x_1, \ldots, x_i(\mathrm{B}x_1, \ldots, x_i \rightarrow \mathrm{A}_1x_1 \ldots x_i)$ or … or $\forall x_1, \ldots, x_i\ (\mathrm{B}x_1 \ldots x_i \rightarrow \mathrm{A_n}x_1 \ldots x_i)$; or
(3) $\mathbf{C_y}$ contains the expression $\mathrm{D}x_1/u_1 \ldots x_i/u_i$ such that $\forall x_1, \ldots, x_i(\mathrm{A}_1x_1 \ldots x_i \rightarrow \mathrm{D}x_1, \ldots, x_i)$ or … or $\forall x_1, \ldots, x_i\ (\mathrm{A_n}x_1, \ldots, x_i \rightarrow \mathrm{D}x_1, \ldots, x_i)$.

This definition says, in effect, that some case is *prima facie* similar to the case at hand if it answers an equivalent question (1), a stronger question (2), or a weaker question (3).

It should also be stressed that—in the general case—there is more than one case *prima facie* similar to the problem situation. In our first example—that of restaurants in Munich—the *prima facie* similar cases include one's previous visits to Bamberg, Berlin, and Würzburg. In *Adams v. New Jersey Steamboat Co.*, the *prima facie* similar cases include those in which an entrepreneur was held liable as an insurer (e.g., *Pinkerton v. Woodward*), and those in which liability was decided in a different way (e.g., cases, such as *Carpenter v. N.Y.*, *N.H. H.R.R. Co.*, pertaining to the liability of sleeping-car companies). Our third example, concerning the model of the atomic nucleus, might also have involved a number of *prima facie* similar cases. Instead of concentrating on the drop of liquid analogy, Bohr could have considered modeling the nucleus as a gas cloud (which turned out to be very fruitful when entire atoms were investigated: It was assumed that electrons in an atom form a "cloud" around the nucleus).

The fact that there usually are numerous cases that are *prima facie* similar to the problem situation is often neglected in existing accounts of analogical argument; yet, as we shall see in the next section, it is an essential aspect of this mode of argumentation.

5 Relevant Similarity

5.1 Relevant Similarity in Legal Analogical Arguments

After identifying the cases that are *prima facie* similar to the problem situation, the next step is to decide which of them is relevantly similar, i.e., which provides the solution to the case at hand. Again, let us illustrate this point with some conceptions developed in legal theory. Robert Alexy realizes that—given a problematic case—one can usually formulate opposite, and *prima facie* justified, solutions to it. In addition to the already cited rule of analogy, namely,

A1: In every case c_i, each case c_j may be adduced with the argument that c_i shares with c_j features $F_1, \ldots, F_f$, and that c_i, for that reason and because there are reasons for rule $F_1, \ldots, F_f \rightarrow Q$, ought to be treated, as c_j, to the effect that Q,

Alexy introduces a second rule as follows:

A2: In each case in which an argument of the form A1 is put forward, two counterclaims may be raised:
A2.1: It may be claimed that c_i is distinguished by the features $F_g, \ldots, F_m$ from c_j, and that c_i, for that reason and because there are reasons for the rule $F_g, \ldots, F_m \rightarrow \neg Q$, ought to be treated, in contradistinction to c_j, to the effect that $\neg Q$.
A2.2: It may be claimed that c_i shares with c_k the features $F_n, \ldots, F_z$, and that c_i, for that reason and because there are reasons for the rule $F_n, \ldots, F_z \rightarrow \neg Q$, ought to be treated, as c_k, to the effect that $\neg Q$ (Alexy 2010, 10).

Therefore, Alexy claims that—in the general case—the problem situation is *prima facie* similar to different cases leading to incompatible conclusions. He further posits that under such circumstances the decision as to which of the outcomes—Q or $\neg Q$—is to be preferred is reached through the utilization of the so-called weight formula (Alexy 2007). The formula is strictly connected to Alexy's view of the legal system. He believes that the law consists of rules (i.e., norms applicable in an all-or-nothing fashion) and principles (i.e., optimization criteria). The logical difference between rules and principles is that, while rules either apply or not in the given circumstances, whether principles apply will depend on whether they outweigh some other principles leading to an incompatible conclusion. It is the weight formula that serves to decide such conflicts between principles. In *Adams v. New Jersey Steamboat Co.*, for example, the case of the innkeeper's liability is arguably governed by a principle which posits that when a legal bond is connected to a relationship of extraordinary trust between the parties, the highest standard of responsibility is called for. On the other hand, the case of the sleeping-car company is governed by a different principle, saying that it is the essence of a contract, and not an additional service, which determines the level of liability: What sleeping-car companies provide is first and foremost transportation, not accommodation, and hence they are not liable as insurers. Our problem situation—i.e., the question of whether steamboat operators are liable as insurers or on the proof of negligence—is *prima facie* similar both to the case of innkeepers and to the case of sleeping-car companies, since—on the one hand—the service provided by steamboat operators arguably involves a relationship of extraordinary trust, yet—on the other hand—the accommodation provided for passengers does not constitute the essence of the contract. Thus, according to Alexy, the analogical decision in such a case should be reached through a balancing of the two principles. In *Adams v. New Jersey Steamboat Co.*, the court came to the conclusion that the principle "When a legal bond is connected to a relationship of extraordinary trust between the parties, the highest standard of responsibility is called for" outweighs the principle "It is the essence of a contract, and not an additional service, which determines the level of liability."

A different approach to determining the relevant similarity between cases is suggested by Scott Brewer. Let us recall that on his view analogy proceeds, first, by abducting an analogy-warranting rule that can serve to solve the case at hand; the abduction is based on a case (or any number of cases) similar to the problematic one. The second step, which essentially is the establishment of relevant similarity, boils down to the confirmation or disconfirmation of the analogy-warranting rule. This is achieved by recourse to so-called analogy-warranting rationales, i.e., various arguments which serve as the basis for "reflective adjustment" of the rule under consideration. One works back and forth between these arguments and the case at hand to establish whether the abducted rule is relevant to the case (the arguments, include *inter alia*, the requirement that the rule be consistent with other valid rules of law) (see Brewer 1996, 962ff.). To illustrate, in *Adams v. New Jersey Steamboat Co.*, the analogy-warranting rule is that a service-provider entrusted with extraordinary trust by his clients is liable as an insurer. In turn, the analogy-warranting rationales include all the arguments that support the rule, e.g., considerations of public policy, the pointing out of important differences between the legal position of the steamboat operators and that of the sleeping-car companies.

Finally, in accounts of analogical argument which are based on direct case comparison, the establishment of relevant similarity is decided in three steps. First, the relevant facts of the case, the so-called factors, are selected. Second, the analogy between cases is established by comparing them by reference to those factors. Third, the solution of a similar (decided) case is applied in the case at hand when the cases are sufficiently similar, or else the two cases are distinguished, when the similarity between them is not sufficient (cf. Roth and Verheij 2004, 635). One particular way of formalizing this general argument scheme is suggested by Roth and Verheij. They analyze legal cases in terms of the theory of dialectical argumentation. In the theory, each case is represented as a set of interconnected statements expressing some states of affairs. The statements may support or attack other statements. The conclusion of a case (an abstract legal statement) is the one that has enough support. Now, cases represented in this way may be compared. The comparison is carried out by considering factors, statements corresponding to relevant states of affairs. In any previously settled legal case, the conclusion has a certain level of support determined by its constellation of factors. Roth and Verheij claim that an undecided case may be solved analogically to the settled case if the factor analysis shows that it has at least as strong support for the conclusion as the settled case (cf. ibid., 642–643).

It is clear from the conceptions presented above that there is no single, commonly accepted way of determining the relevant similarity between two cases. However, even the small sample outlined above suffices for the following generalization. There are two methods for establishing whether the problem situation is relevantly similar to some other case: One may be called *theory-based* analogy,[10] the other *factor-based* analogy. According to the first, two cases are deemed relevantly similar if the theory used to solve one of them is justifiably applicable to the other. The second method, for its part, assumes that two cases are relevantly similar if they are identical with regard to (some established number of) relevant factors.

5.2 *Theory-Based Analogy*

Let us begin with a more detailed analysis of theory-based analogy. Let us recall that the problem situation is characterized by the question evoked by what one knows (the set $\mathbf{C_0}$):

(Q) $?\{\mathrm{A}_1 x_1, \ldots, x_i, \ldots, \mathrm{A}_\mathrm{n}\, x_1, \ldots, x_i\}$

We also know how to identify cases ($\mathbf{C_1}, \mathbf{C_2}, \mathbf{C_3}, \ldots$) which are *prima facie* similar to $\mathbf{C_0}$ (for the sake of readability I confine the discussion to situations in which *prima facie* similarity is established when some cases involve questions which are *equivalent* to the question evoked by the problem situation; I leave aside *prima facie* similarity involving stronger and weaker questions). A case $\mathbf{C_x}$ is *prima facie* similar to the problem situation if it answers an equivalent question, i.e., if it includes a statement of the form

(OA) $\mathrm{A}_1 x_1/u_1, \ldots, x_i/u_i$ or … or $\mathrm{A}_1 x_1/u_1, \ldots, x_i/u_i$

Let us call it the *original answer*. Thus, in the Munich example the original answer provided by the analogical case of Bamberg is that restaurants in Bamberg are not open late into the night, while in *Adams v. New Jersey Steamboat Co.*, one of the original answers is that some particular innkeeper is liable as an insurer, and the other original answer is that some particular sleeping-car company is not liable as an insurer. In the case of atomic nuclei, the original answer is that the structure which exhibits the phenomenon of selective absorption is a drop of liquid.

The idea behind theory-based analogy is that the analogical cases involve—or may serve as the basis for constructing—a theory which provides grounds for solving the problem situation. Therefore, it is essential—in relation to each *prima facie* similar case—to isolate the theory that give rise to the original answer. A straightforward way to do so is to identify the subset $\mathbf{OT_{Cx}}$ of the *prima facie* similar case $\mathbf{C_x}$ (let us call it the *original theory*) which deductively implies the original answer. Thus, the original theory includes OA, but also those sentences belonging to $\mathbf{C_x}$ which may jointly serve to derive OA. The $\mathbf{OT_{Cx}}$ may or may not include some general

[10] The identification of theory-based analogy is inspired by Bench-Capon and Sartor (2001).

statements. For example, in our first example, it presumably includes only a concrete statement that restaurants in Bamberg are not open late into the night. In *Adams v. New Jersey Steamboat Co.*, on the other hand, the statement that some particular innkeeper is liable as an insurer (e.g., in *Pinkerton v. Woodward*) is accompanied by some general rules which served to derive it (e.g., that all innkeepers are liable as insurers or that when a legal bond is connected to a relationship of extraordinary trust between the parties, the highest standard of responsibility is called for). The original answer that a drop of liquid exhibits selective absorption is also derived from a general theory, one that captures the physical characteristics of liquids.

When the original theory $\mathbf{OT_{Cx}}$ includes not only the original answer OA, but also some general statements, we can speak of a *proper original theory*. However, when $\mathbf{OT_{Cx}}$ consists only of OA, an additional step is necessary: the construction of a proper theory which would include OA, but also some general statements which imply it. There is no algorithm for how to do it. One option is to use the logic of induction. Let us imagine that—in addition to knowing that restaurants in Bamberg are not open late into the night—we also know that the same is the case in Kiel, Dortmund, and Hildesheim. We also know that all these cities are German. Therefore, by induction, we can arrive at a general conclusion that restaurants in German cities are not open late into the night and include it into $\mathbf{OT_{Cx}}$. This is the kind of procedure Brewer (1996, 962ff.) suggests in speaking of abducting an analogy-warranting rule on the basis of a number of cases similar to the case at hand.

The extracted—or constructed—original theory $\mathbf{OT_{Cx}}$ is not always directly applicable to the problem situation; we shall say that it does not always constitute the *target theory*. If that is the case, the theory must be modified in such a way that it provides an answer to Q (the question evoked by $\mathbf{C_0}$). Therefore, the modified theory $\mathbf{MOT_{Cx}}$ must meet the following criteria: (a) It must deductively imply an answer to Q, and (b) it must be obtained through some *structure-preserving transformation* of $\mathbf{OT_{Cx}}$. Again, there is no algorithm for how to achieve this, although some particular methods may be identified. For example, one can generalize $\mathbf{OT_{Cx}}$ so that it deductively implies both OA and the answer to Q. Such was the case when sound propagation was explained in terms of similarity with water waves, and the resulting theory was applicable to both sound waves and water waves (cf. Holyoak 2012, 234).

Having identified (i.e., extracted or constructed, and possibly modified) the original theory, one is in possession of the target theory, which provides a *possible solution* to the problem situation; but one still has to *decide* whether the target theory should be applied. In other words, it must be determined whether the *prima facie* similar case which gave rise to the formulation of the original theory is indeed relevantly similar to the problematic case at hand. To this end, one needs to employ some *normative criterion*, and the kind of criterion used depends on the field in which the analogical argument is employed (e.g., physics, everyday experience, law) and on one's axiological framework (e.g., utilitarianism vs. Kantianism in ethics), etc. (cf. Bartha 2013).

However, given that one works with theories, it is possible to identify a number of such criteria which are context-independent (cf. Bench-Capon and Sartor 2001, 16–17). First, one may ask whether the target theory is *consistent*: Given that one should avoid contradictions, an inconsistent theory should be rejected. Second, one may take into account the target theory's *internal coherence*: The more coherent it is, the better a candidate it will be for providing an answer to the problem situation.[11] Third, one may consider the target theory's *external coherence* with one's background knowledge: Again, the greater it is, the better. Importantly, given that one's background knowledge embodies one's theoretical as well as axiological frameworks, the criterion of external coherence seems pivotal in any analogical argumentation. Let us assume, for example, that a judge—let's call him RP—is considering *Adams v. New Jersey Steamboat Co*. RP has identified a case that is *prima facie* similar to the contemplated one: A case that gave rise to a target theory, which includes the rule that steamboat operators are liable as insurers. Let us further assume that RP adheres to the Economic Analysis of Law and believes that the law should promote social welfare. If the economic analysis shows that the rule "steamboat operators are liable as insurers" does not lead to an increase in social welfare, the target theory is not coherent with RP's background knowledge. Fourth, one may assess the target theory with regard to its *explanatory power*: The bigger the class of cases the theory solves, the better it is. Fifth, one may consider the target theory's *simplicity*: The simpler it is, the better. Simplicity may be defined in different ways, but—intuitively—it hangs together with the number of independent assumptions it includes. Sixth and finally, even the target theory's *heuristic power* may be taken into account. A theory's heuristic power is measured by the number of answers to *other* questions it provides, as well as by the number of new questions it evokes. For example, the drop of liquid model of atom nuclei proved useful not only in explaining the phenomenon of selective absorption, but also (a few years later) nuclear fission.

Moreover, it must be remembered that there are usually more than one case that is *prima facie* similar to the problem situation. It follows that one can usually extract or construct (and possibly modify) more than one target theory. Under such circumstances, the six criteria enumerated above may serve as *comparison criteria*. Thus, (1) a target theory TT^1 which is consistent should be preferred over a target theory TT^2 which is inconsistent; (2) a target theory TT^1 which is internally more coherent should be preferred over a target theory TT^2 which is internally less coherent; (3) a target theory TT^1 which is externally more coherent should be preferred over a target theory TT^2 which is externally less coherent; (4) a target theory TT^1 which has more explanatory power should be preferred over a target theory TT^2 which has less explanatory power; (5) a target theory TT^1 which is simpler (i.e., makes fewer independent assumptions) should be preferred over a target theory TT^2 which is more complicated; and (6) a target theory TT^1 which has more heuristic power should be preferred over a target theory TT^2 which has less heuristic power. It should be clear by now why the possibility of working simultaneously with several target theories (i.e., contemplating more than one *prima facie* similar cases) is reasonable: It enables a more thorough utilization of the normative criteria at play (e.g., instead of asking whether a given theory TT^1 is simple enough or coherent enough or has enough explanatory power, one checks whether its explanation is simpler, more coherent, or

[11] There are different measures of coherence. See Bonjour (1985) and Hage (2013).

more powerful than the one available under TT^2). This dialectical dimension of analogical arguments is crucial, especially in those fields where a decision concerning the problem situation *must be made* (e.g., in the law the court cannot abstain from issuing a ruling).

It should also be noted that in the literature—in particular in the legal-theoretic accounts of analogy—one often finds descriptions of *rule-based analogical reasoning* (cf. Hage 2005; Prakken and Sartor 1998). A rule-based analogy proceeds by identifying a case similar to the problem situation. In the second stage, the legal rule governing similar cases is extracted (or constructed). The third and final stage is the application of that rule (or its modification) to the problematic case. I have not presented rule-based analogical argumentation separately, as I believe it constitutes a special case of theory-based analogy.

5.3 Factor-Based Analogy

The second way of establishing relevant similarity is factor-based analogy. Again, we have a problem situation, i.e., the set $\mathbf{C_0}$, which evokes a question:

(Q) $\mathbf{?}\{\mathrm{A}_1 x_1, \ldots, x_i, \ldots, \mathrm{A}_\mathrm{n}\, x_1, \ldots, x_i\}$

This serves to identify *prima facie* similar cases $\mathbf{C_1}$, $\mathbf{C_2}$, $\mathbf{C_3}$, etc., that include an answer to a question which is equivalent to (or stronger or weaker than) Q. On the factor-based analogy, the next step is to compare $\mathbf{C_0}$ with $\mathbf{C_1}$, $\mathbf{C_2}$, $\mathbf{C_3}$, etc., with regard to factors, i.e., the relevant features of the case (statements expressing the relevant states of affairs). Factors are given (known from elsewhere), and so their determination is not a part of the procedure we analyze here. As an illustration, let us consider *Adams v. New Jersey Steamboat Co.*, which is compared both with decided cases (e.g., *Pinkerton v. Woodward*) in which the liability of an innkeeper was determined, as well as with cases (e.g., *Carpenter v. N.Y., N.H. H.R.R. Co.*) in which the liability of sleeping-car companies was at stake. The factors in those comparisons included, *inter alia*, the following statements:

(F1) The legal relationship between the service provider and the client involves a relationship of extraordinary trust.
(F2) The client is provided with a room of his own with a locking door.
(F3) The purpose of the contract is to provide transportation.
(F4) The client is accommodated in an area that anyone can easily access.

With the innkeeper case, *Adams v. New Jersey Steamboat Co.* shares factors (F1) and (F2), while with the case of the sleeping-car company it only shares (F3). The question is, which of the similarities is stronger. To establish this, one needs some criterion of comparison. For example—as Roth and Verheij suggest—one can make recourse to the concept of the strength of support for the conclusion (cf. Roth and Verheij 2004, 642–643). Arguably, in the innkeeper case the conclusion (liable as an insurer) is supported by both (F1) and (F2), while in the case of the sleeping-car company the conclusion (not liable as an insurer) is supported by (F3) and (F4). *Adams v. New Jersey Steamboat Co.* provides as strong a support for the "liable as an insurer" conclusion as the original case, while a weaker support for the "not liable as an insurer" conclusion. Thus, *Adams v. New Jersey Steamboat Co.* should follow the ruling in the innkeeper case.

As we can see, the main difference between theory-based and factor-based analogies lies in the way cases are compared. According to the former, a relevant similarity is established when the original theory extracted from (or constructed on the basis of) a *prima facie* similar case turns out to be applicable to the problematic case. The theory contains general statements and deductively entails the answer to the problem situation. Moreover, the theory is evaluated against a number of criteria, such as consistency, internal and external coherence, simplicity, explanatory power, and heuristic power. On the other hand, the factor-based analogy proceeds by establishing which of the (predetermined) factors are shared by the two cases being compared, and—if the constellation of such shared factors is sufficient—the conclusion of the *prima facie* similar case is applied to the problem situation. Thus, factor-based analogy involves no formulation of a general statement (rule), which would deductively entail the conclusion. Moreover, while the procedure helps to avoid inconsistencies, the outcome (i.e., the resulting shape of our knowledge concerning the given domain) is not checked for internal and external coherence, simplicity, and explanatory and heuristic power. In other words, factor-based analogy seems to produce ad hoc solutions, and hence may lead us to a local optimum, instead of a global one.

A defender of factor-based analogy, however, may argue that the procedure secures *some* level of external coherence (ultimately, different but similar cases are resolved in the same way, and so they are governed by the same implicit rule), simplicity (no complex system of independent assumptions is needed), and explanatory power (a number of cases, which display the same constellation of factors, are solved in the same manner). This is true and shows that factor-based analogy does not necessarily lead to outcomes much divergent from theory-based analogy. The difference between them lies in the manner of reconstructing one's domain knowledge (cf. Hage 2001; see also Brożek 2011). In the theory-based approach, one utilizes first-order logic (or some extension or modification thereof), which makes it possible to explicitly state all the assumptions of the analogical argument (the utilized premises, preferences, and comparison criteria). This also makes it a domain-neutral account of analogical argumentation. The factor-based analogy, on the other hand, takes advantage of a *dedicated mechanism* tailored to a particular domain. This enables one to sweep much of what is going on under the carpet: A number of crucial elements of analogical argumentation (e.g., the choice of factors, the criterion for determining the level of similarity between the cases, etc.) become encapsulated in the proposed formal mechanism. At the same time, factor-based analogy is computationally much easier to handle. For some applications—e.g., the development of legal databases or decision-aiding software—a dedicated, domain-dependent account of analogical argumentation may be the only viable option.

6 Solution

The general structure of analogical arguments as reconstructed above looks as follows:

(1) One encounters a problem situation ($\mathbf{C_0}$), which evokes a question Q.

(2) There are cases ($\mathbf{C_1}$, $\mathbf{C_2}$, $\mathbf{C_3}$, ...) that are *prima facie* similar to $\mathbf{C_0}$, i.e., cases which answer (a) an equivalent, or (b) a stronger, or (c) a weaker question than Q.

(3) A relevant similarity between $\mathbf{C_0}$ and some $\mathbf{C_x}$ is established by:

   a. extracting or constructing a theory based on $\mathbf{C_x}$ ($\mathbf{C_y}$, $\mathbf{C_z}$, ...) and deciding, whether—and which of them—is applicable to $\mathbf{C_0}$; or
   b. comparing cases $\mathbf{C_0}$ and $\mathbf{C_x}$ with regard to some selected factors and deciding whether the similarity is strong enough to apply the conclusion of $\mathbf{C_x}$ to $\mathbf{C_0}$.

After the relevant similarity is established, there remains the task of formulating *the solution* to the problem situation, i.e., the answer to question Q. According to the theory-based analogy the answer is deductively implied by the selected target theory, while on the factor-based approach it is the same answer which was given in the relevantly similar case (i.e., the case with which the problem situation shares enough factors).

In this context, an interesting problem is the status of the conclusion of an analogical argument. In the literature, pertaining to analogical reasoning much discussion revolves around the role of analogy: Is it a heuristic device only or can it provide justification (cf. Bartha 2013). In. the former case, the conclusion would be a mere suggestion as to how to solve the problem situation; in the latter, the answer supplied by analogical reasoning would be a justified solution to the case at hand.

This controversy—heuristics or justification—is perfectly legitimate when one speaks of analogy as a mode of cognition. However, once we move to analogical *arguments* it is clear that their outcomes should count as *justified* solutions to the problem situation. This is plainly visible when one considers theory-based analogy, where the answer one arrives at is deductively implied by the target theory, while the theory itself has some desirable features (consistency, internal and external coherence, explanatory power, and simplicity). At the same time, the establishment of *prima facie* similarity may be viewed as a heuristic stage of analogical argumentation. The situation is more troublesome with regard to factor-based analogy. In this case, the solution to the problem situation is not (at least at the surface level) a deductive conclusion from some set of premises; rather, it is formulated because the problem situation shares some deciding factors with another, previously decided case.

However, on closer inspection this theoretical trouble disappears. Let us go back to the theory of analogy developed by Roth and Verheij. They claim that an analogical solution to a case at hand can be established only if the support for the conclusion in the problem situation is at least as strong as in some (previously decided) case. They also show how it is possible to compare cases leading to incompatible conclusions (see Roth and Verheij 2004, 642–643; see also Prakken 1997). So, their formal framework encapsulates a certain view of justification which may be summarized as follows: (a) a conclusion is not justified if it has no support; (b) a conclusion is justified if its support is at least as strong as the support for the same conclusion in some other, previously decided case; and (c) when two contradictory conclusions are considered, the one which has stronger support is justified. Thus, given a plausible underlying argumentation framework, factor-based analogy may be viewed as a justification-generating kind of argument.

∗ ∗ ∗

I do not pretend to have covered everything that has been written about analogical arguments, or even the bulk of it. Such a comprehensive account would be very difficult—if not impossible—given the ever-growing number of studies pertaining to analogy, and the diversity of the proposed analogical schemes, which are often tailored to particular problems or domains, and which come with their own ontological and epistemological baggage. However, I do hope to have provided a plausible general characterization of analogical argumentation. Of course, the problem situation, *prima facie* similarity, and relevant similarity can arguably be reconstructed in a different way to what I have proposed, but I do not think that a significantly different account of the general framework of analogical arguments is possible.

At the same time, I believe to have stressed some structural features of analogical argumentation which are often neglected or not given due attention. First, I have provided a characterization of the problem situation in terms of erotetic logic. Second, I have distinguished between *prima facie* similarity and relevant similarity, and defined the former against the background of the theory of questions. Third, I have underlined the importance of the dialectical dimension of analogical arguments, i.e., that one usually works with many cases that are *prima facie* similar to the problem situation. Fourth, and finally, I have provided a general description of theory-based analogy.

References

Alexy, R. 2007. The Weight Formula. In *Studies in the philosophy of law 3: Frontiers of the economic analysis of law*, ed. J. Stelmach, B. Brożek, and W. Załuski, 9–27. Kraków: Jagiellonian University Press.

Alexy, R. 2010. Two or Three? In *On the Nature of Legal Principles*. ed. M. Borowski. *ARSP-Beiheft*, vol. 119, 9–18.

Ando, C. 2015. Exemplum, analogy, and precedent in roman law. In *Exemplarity and singularity*, ed. M. Lowrie, and S. Ludemanm. London-New York: Routledge.

Ashworth, J.E. Medieval theories of analogy. In *The Stanford encyclopedia of philosophy,* Winter 2013 Edition, ed. E.N. Zalta. http://plato.stanford.edu/archives/win2013/entries/analogy-medieval/.

Aristotle. *Metaphysics*, trans. R. Hope. Ann Arbor, MI: University of Michigan Press, 1960. (1st ed.).

Aristotle. 1991. *Historia animalium* (*History of animals*), trans. A.L. Peck. Harvard Cambridge, MA: University Press

Bartha, P. Analogy and analogical reasoning. In *The Stanford encyclopedia of philosophy*, Fall 2013 Edition, ed. E.N. Zalta. http://plato.stanford.edu/archives/fall2013/entries/reasoning-analogy/.

Belnap, N.D. 1969. Aqvist's aorrections-accumulating question sequences. In *Philosophical logic*, ed. J.W. Davis et al. Dordrecht: Springer.

Bench-Capon, T.J.M., and G. Sartor. 2001. Theory based explanation of case law domains. In *Proceedings of the eighth international conference on artificial intelligence and law*, 12–21. New York: ACM Press.

Bonjour, L. 1985. *The structure of empirical knowledge*. Cambridge, MA: Harvard University Press.

Brewer, S. 1996. Exemplary reasoning: semantics, pragmatics, and the rational force of legal argument by analogy. *Harvard Law Review* 109: 923–1028.

Brożek, B. 2007. *Rationality and discourse*. Warszawa: Wolters Kluwer Polska.

Brożek, B. 2011. Legal logic. myths and challenges. In *Theory of imperatives from different points of view*, ed. A. Brożek, J.J. Jadacki and B. Zarnic, 49–59. Warszawa: Semper.

Cicero (1949). *On invention*, *the best kind of orator*. *Topics*, trans. H.M. Hubbell. Cambridge, MA: Harvard University Press.

Gentner, D. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7: 155–170.

Hage, J. 2001. Legal logic: Its existence, nature and use. In *Pluralism and law*, ed. A. Soeteman, 347–373. Dordrecht: Kluwer Academic Publishers.

Hage, J. 2005. The logic of analogy in the law. *Argumentation* 19: 401–415.

Hage, J. 2013. Three Kinds of coherentism. In *Coherence: Insights from philosophy, jurisprudence and artificial intelligence*, ed. M. Araszkiewicz, and J. Savelka, 1–32. Dordrecht: Springer.

Hesse, M. 1966. *Models and analogies in science*. Notre Dame, IN: Notre Dame University Press.

Hochschild, J.P. 2010. *The semantics of analogy*. *Rereading Cajetan's 'De Nominum Analogia'*. Notre Dame, IN: Notre Dame University Press.

Holyoak, K.J., and P. Thagard. 1995. *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press.

Holyoak, K.J. 2012. Analogy and relational reasoning. In *The Oxford handbook of thinking and reasoning*, ed. K.J. Holyoak, and R.G. Morrison, 234–259. Oxford-New York: Oxford University Press.

Kolodner, J.L. 1993. *Case-based reasoning*. San Mateo, CA.: Morgan Kaufmann.

Lakoff, G., and R.E. Núñez. 2000. *Where mathematics comes from. How the embodied mind brings mathematics into being*. New York: Basic Books.

Lakoff, G., and M. Johnson. 2003. *Metaphors we live by*. Chicago, IL: The University of Chicago Press.

Pais, A. 1991. *Niels Bohr's times in physics, philosophy, and polity*. Oxford: Clarendon Press.

Pal, S.K., and S.C.K. Shiu. 2004. *Foundations of soft case-based reasoning*. Hoboken, NJ: Wiley.

Prakken, H. 1997. *Logical tools for modelling legal argument*. Dordrecht: Kluwer Academic Publishers.

Prakken, H., and G. Sartor. 1998. Modelling reasoning with precedents in a formal dialogue game. *Artificial Intelligence and Law* 6: 231–287.

Roth, B., and B. Verheij. 2004. Cases and dialectical arguments—An approach to case-based reasoning. In *OTM workshops 2004*, ed. R. Meersman, et al., 634–651. Dordrecht: Springer.

Weinreb, L.L. 2005. *Legal reason: The use of analogy in legal argument*. Cambridge, Mass.: Cambridge University Press.

Winston, P.H. 1980. Learning and reasoning by analogy. *Communications of the ACM* 23: 689–703.

Wiśniewski, A. 1995. *The posing of questions. logical foundations of erotetic inference*. Dordrecht: Springer.

Choosing Ends and Choosing Means: Teleological Reasoning in Law

Lewis A. Kornhauser

1 Introduction

Legal reasoning is often teleological though in a broader sense than used by John Rawls (1971) who distinguished between deontological reasoning and teleological reasoning. For him, teleological reasoning is goal-directed and hence consequentialist; it evaluates proposed actions, proposed policies, and proposed institutions exclusively on the basis of the achievement of the presupposed goals or, phrased differently, exclusively on the basis of the consequences that these actions, policies, or institutions would have.[1] Deontological reasoning, by contrast, evaluates actions, policies, or institutions at least in part on the basis of some non-consequentialist features of the action, policy or institution: something other than outcomes matters. Most deontological accounts, however, do not ignore consequences entirely; in some circumstances, deontologists may thus have to engage in teleological reasoning.

Teleological reasoning in law is, as for Rawls, goal-directed, but legal goals need not (always) be moral ones and the criterion of evaluation need not be consequentialist in the moral sense. Legislators everywhere have goals that they seek to promote through statutes; they must reason consequentially when they decide which legislation to enact. Administrative agencies, when rule-making, must adopt a similarly consequential stance. So, for example, the US Congress enacted the Clean Air Act to reduce the number of pollutants in the air. The Environmental Protection Agency promulgates regulations to further that goal. Similarly, Congress enacted Title VII, the Employment Discrimination Act, to promote a specific vision of labor market

[1] Some authors, e.g., Broome (1991) and Woodard (2008) distinguish between teleological and consequentialist views. On their accounts, a consequentialist view adopts an impartial or objective theory of value while a teleological one may have a subjective theory of value. I ignore this distinction here and use the terms "teleological" and "consequentialist" interchangeably.

L. A. Kornhauser (✉)
School of Law, New York University, New York, NY, USA
e-mail: lewis.kornhauser@nyu.edu

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_14

performance. The specific policies adopted are chosen at least in part on the basis of the consequences they have. Consequentialist reasoning is thus a central feature of legal reasoning. Its extent and role, however, may vary with the institutional setting. The extent to which adjudicators, moreover, must deploy consequential reasoning may vary with the structure of the legal system and the set of tasks allocated to the judiciary.

Teleological reasoning in law thus poses two questions. What legal goals do legal actors pursue? What is the process of teleological reasoning?

I address the second question in two stages as the structure of teleological reasoning, though apparently straightforward, contains some subtleties. The chapter thus begins by articulating the nature of teleological reasoning itself and its relation to rational choice in general and to rational choice theory in particular. I then turn to a set of questions that arise in a legal context. In the course of this latter discussion, I shall argue that much non-teleological reasoning in law elaborates the legal ends that teleological reasoners must pursue.

The difficulties that confront legal actors engaged in teleological reasoning are quite numerous though not unique to the legal setting. So, for example, legal ends are often unclear and incompletely specified; thus, as earlier chapters suggest, much legal reasoning is directed at the specification and elaboration of legal goals. In addition, legal ends are collective, rather than individual ends. The collective nature of legal ends poses two difficult problems of institutional design. Though legal ends are collective, they are interpreted and implemented by individuals. Each of these individuals has her own ends; legal institutions must insure that each individual reasons consequentially on the basis of the legal ends rather than on the basis of the individual's own ends. In addition, legal institutions exhibit an extensive division of labor. One institution sketches a goal; a second institution elaborates that goal; and a third agent implements it. Or, two distinct institutions grapple with different but related aspects of a policy problem.

Good institutional design would maximize the extent to which these different institutions coordinate their actions.

A third difficulty arises from the nature of the choices that legal decision-makers confront. Legal actors typically choose among policies or institutional frameworks.

Teleological reasoning requires that these actors evaluate their options in light of the legal goals. These legal goals are often consequentialist in a more ethical sense; achieving the goals requires that complex social behaviors change. The desired social states, however, are "distant" from the choices faced in two respects. Consider a choice of a simple policy such as the choice of a speed limit on a limited access highway. This choice only partially determines the outcomes relevant to policy assessment—travel times and accident rates, for instance; these policy-relevant outcomes depend on the decisions of drivers that the policy influences.[2] Teleological reasoning thus requires the policymaker to predict how agents will respond to each policy she might choose. The choice among institutional arrangements most clearly

[2] Indeed, the outcomes depend on many other factors as well—the extent to which law enforcement agencies police the speed limit; the level of the fine for violation, etc.

illustrates the second form of “distance.” Institutions are twice removed from outcomes. When a policymaker designs an institution, she creates an institution that will itself make policy which will structure but not fully determine the decisions of many other agents. Evaluating an institution thus requires not only predicting how other agents will respond to the policies formulated by the institution but also predicting what policies the institution will promulgate. As I will argue below, this additional level of predictions poses challenging barriers to assessment. Finally, and parallel to the last point, the relation between the options available to the agent and the criterion she seeks to maximize (or satisfy) are complex. Her choices may only indirectly advance her goals; typically, consequences depend not only on the agent’s actions but on chance and on the actions of others or even on additional choices of the agent. These features require refinement or modification of the notion of maximization.

The rest of the chapter proceeds as follows. Section 2 explicates the nature of teleological reasoning. The direction “attend to consequences” is deceptively simple; it conceals a number of complexities. One must determine (1) to what aspects of consequences one must attend, (2) the nature of the attention one must give to consequences, and (3) the relation between the agent’s choices and the consequences that arise. Some analysts suggest that an agent may, in answer to the second question, either maximize or satisfice. I define these terms and suggest that, generally, we may understand satisficing as constrained maximization.

Section 3 discusses various forms of legal instrumentalism, the view that law is an instrument for promoting collective ends. Legal instrumentalism thus dictates teleological reasoning but different forms dictate that the reasoning occurs at different levels and by different individuals.

Section 4 briefly discusses the choice of legal ends and the role of interpretation. Different interpretive theories prescribe distinct styles of reasoning when elaborating these ends. Purposive theories of interpretation dictate the use of teleological reasoning in the elaboration of ends. Section 5 briefly discusses the choice of legal means.

2 The Structure of Consequentialist Reasoning

Teleological reasoning requires the agent to adopt the available option that best advances her goals. This simple statement identifies three key elements of teleological reasoning: her ends, the set of feasible options, and the evaluation of each feasible option to identify the “best” one.

This section first discusses the structure of the agent's ends. It then considers the basic idea of rational choice among these ends. The discussion of rational choice initially assumes a transparent relationship between the agent's evaluation of consequences and her evaluation of the choices open to her; it assumes that the agent chooses consequences (or the extent to which she achieves her ends) directly and unequivocally. I consider more indirect choices later in the subsection on rational choice.

2.1 The Structure of Evaluation

We may understand an agent's goals as a set of evaluative criteria against which she assesses the actions available to her. Any actual piece of teleological reasoning thus depends crucially on the *content* of the agent's evaluative criteria. Occasions for teleological reasoning differ widely in the content of the evaluative criteria the agent uses, but the reasoning that the agent undertakes on these occasions nonetheless share a common structure; it is this common structure that I examine in this section. In considering teleological reasoning in law, however, it is important to remember that consequentialist theories require a theory of value; much legal reasoning formulates or elaborates the ends that legal institutions and agents ought to or do pursue.

An often ignored precondition of teleological reasoning concerns the nature and structure of the criteria used to evaluate the agent's success at achieving her goal. Typically, analysis begins with a *ranking*.[3] The existence of a ranking entails that the evaluations are comparative: given two consequences *A* and *B*, we say either that *A* is better than *B*, *B* is better than *A*, or *A* and *B* are equally good.[4] Our evaluation of actions, policies, institutions, characters or whatever must then rest on this underlying ranking of consequences.

[3] One might begin instead with a *choice function* that identifies, for each possible set of options available to the agent, the option that agent would choose. A large literature investigates the relation between choice functions and rankings. See, e.g., Hansson and Grune-Yanoff (2012) for a brief survey.

[4] The text understates the complexity of the comparative process. There are at least two other "answers" to the comparative process. It may be that the ranking is incomplete so that *A* and *B* are "not comparable." So, for example, the Pareto criterion provides a partial ranking for which any two Pareto optimal points are Pareto non-comparable. Ruth Chang suggests that "*A* may be on a par with *B*," a relation that undermines the construction of a ranking. Under this approach, the ranking is complete but it may not be transitive or even acyclic (Chang 2001).

We typically represent the goals pursued by an agent with a ranking.[5] A ranking permits the agent to compare some or all of her options. If a ranking is complete, the agent may compare any two options. If a ranking is incomplete or partial, then there may be pairs of options that the agent cannot rank; it will be the case that option *A* is not ranked more highly than option *B*; and option *B* is not ranked more highly than option *A*; and options *A* and *B* are not equally ranked. Both partial and complete rankings are transitive; i.e., if option *A* is ranked more highly than option *B*, and option *B* is ranked more highly than option *C*, then option *A* is ranked more highly than option *C*.

Consequentialist theories of value in ethics base the ranking of options on the consequences each option produces. Policy *P* is better than policy *Q* if and only if the consequences of policy *P* are better than the consequences of policy *Q*. We can identify a variety of consequentialist criteria that vary in their demandingness. Many legal goals also evaluate and rank options on the basis of their consequences, and I shall use consequentialist language generally, but the discussion applies equally to rankings that derive from other bases.

2.2 Rational Choice

To begin, then, consider a concrete problem of consumer choice under certainty. Suppose an agent Liza has gone to the market to purchase fruit, a task for which she has allotted a fixed budget m^*. For simplicity, assume there are only two types of fruit, apples and oranges. Consequences, or outcomes, in this framework, consist of fruit baskets containing apples and oranges. We may thus describe any fruit basket by an ordered pair (a, o) in which *a* represents the number of apples in the fruit basket and *o* represents the number of oranges.

Consequentialism requires that Liza base her choice solely on her evaluation of these consequences or baskets of fruit.

[5] Often, the analyst assumes that an agent who must choose an action has rankings over actions, that an agent who must choose a policy has a ranking over policies, or that an agent who must choose an institution has a ranking over institutions. The analyst often assumes, that is, that the agent's domain of preference corresponds to her domain of choice. But this elision of the two domains in some cases may mischaracterize the agent's decision problem. The agent requires only a ranking of consequences; she may be unable to rank the options in her domain of choice. Two examples will suffice. First, consider a simple game such as chicken in normal form. Each player has a ranking over each of the four outcomes. But she does not have a well-defined ranking over her actions. She ranks *swerve* over *straight* if the other player has chosen *straight,* but she ranks *straight* over *swerve* if the other player has chosen *swerve*. Second, consider a voter who must choose among several candidates (*A*, *B*, and *C*) to complete a legislature that will then enact a legislative program. The agent must have a ranking over legislative programs, but she need not have a ranking over candidates. She may rank *A* above *B* and *C* if the rest of the legislature consists of *X,* but she may rank *C* over *A* and *B* if it consists of *Y*. For further discussion see the text below at footnotes 21–25 and Kornhauser (2003).

We have not yet fully described the choice problem that Liza faces. At the outset, I shall describe the formal, rational choice framework within which Liza has a well-defined optimization problem that she solves.[6] This framework has several elements: (1) a set of outcomes or consequences; (2) a ranking of the possible outcomes (or a preference over the outcome set); (3) a set of actions or choices; and (4) the environment in which choice occurs.[7]

As we have seen, in the fruit basket example, Liza's domain of preference, the set of outcomes, corresponds to the set of all possible apple–orange pairs (a, o). The assumption of full rationality means that Liza's ranking is complete and transitive. Completeness means that for any pair of baskets (a, o) and (a', o'), Liza can say either than she prefers (a, o) to (a', o') or that she prefers (a', o') to (a, o) or she is indifferent between them.

The environment in which Liza chooses has three elements: the price of apples, the price of oranges, and Liza's wealth m^*. These three elements determine her budget constraint and hence the set of (a, o) pairs or fruit baskets that Liza can actually purchase. This set of purchasable fruit baskets constitutes Liza's choice set.

Teleological reasoning dictates that Liza bases her choice of fruit basket solely on the consequences of her choice. Teleological reasoning in rational choice theory may be summarized by the following principle:

C1: Choose action *A* if there is no action better than *A* available.[8]

In this simple framework, both the logic of teleological reasoning and the resolution of Liza's choice problem are transparent, indeed, almost self-evident. As Liza's choice set lies in the outcome set, she simply ranks her feasible choices as her preference ranking dictates.

Optimization entails that she simply chooses the feasible choice that ranks highest. (If there is more than one most highly ranked feasible choice, she is free to choose any one of them.)

This account of teleological reasoning has tied it directly to the structure of formal, rational choice theory. The framework outlined, however, allows us to see both how complex teleological reasoning might be, even in the rational choice context, and how teleological reasoning might work under less restricted conceptions of rationality.

[6] I have assumed that Liza exhausts her budget; holding cash has no value for her. If holding cash after purchasing fruit has value, then Liza must have preferences over triplets (m, a, o) where m represents the cash remaining after all purchases.

[7] This framework corresponds roughly to the framework introduced in Savage (1954). Savage, however, studies decision-making under uncertainty. He thus assumes that the agent has primitive preferences over actions (or choices), and then identifies conditions on these preferences so that the preferences over actions can be represented by a preference over outcomes (itself representable by a utility function) and a set of beliefs about states the world such that one action *A* is preferred to another action A' if and only if the expected utility of *A* exceeds the expected utility of A'.

[8] More precisely: "Choose an action *A* from among those actions for which no better action is available." When the ranking over actions is complete, the agent chooses from among those actions that are maximal; if there is more than one, she is indifferent among them. If the ranking is incomplete, the agent again chooses among actions that are maximal but now two actions available to her may be non-comparable (on the criterion used by the ranking).

We might weaken our conception of rationality in at least two ways. First, we might weaken the optimization requirement. In effect, we might substitute $C2$ for $C1$:

$C2$: Choose action A if action A is at least as good as some standard S.[9]

$C2$ is often described as a "satisficing" rather than a maximization criterion.[10] This simple modification, however, does not obviously greatly complicate the analysis. We can, after all, easily reformulate $C2$ as a maximization criterion by appropriate definition of the agent's ranking of alternatives.[11] Of course, $C2$ may leave the agent with extensive discretion; but so may $C1$.

$C2$ divides the options into at most two classes: acceptable and unacceptable classes. This dichotomous ranking will be complete and transitive. When an agent has a complete ranking, she can compare any two options and determine whether option A ranks above option B, whether option A ranks below option B, or whether A and B are equally ranked. Under $C2$, if two policies are at least as good as the standard S, they are equally ranked. Similarly, if two policies are not as good as the standard S, they are equally ranked. If option A meets standard S but policy B does not, then option A ranks more highly than option B. $C2$ requires the agent to choose from among the acceptable options.

Criterion $C1$ is generally understood as a maximization criterion. It requires the agent to choose an option from the set of best available options. Notice that this criterion might leave the decision-maker with much discretion. The extent of discretion depends on the degree of precision of the agent's ranking of alternatives.

A ranking may be imprecise or crude in at least two different ways. First, it may be incomplete or partial. In this instance, there may be several options, for each of which no more highly ranked option exists but that are all mutually non-comparable. We might reduce the agent's discretion by refining the ranking the agent deploys. Such refinement may be difficult because we must identify criteria that permit the comparison of non-comparable options but that do not introduce intransitivities among the already ordered alternatives.

Second, though a ranking is complete, there may be many options, all of which are equally good but, for none of which, there is a better option. We saw this most obviously in the transformation of $C2$ from a satisficing to a maximization criterion that assumes that the agent is indifferent among all those options that meet the standard S. But it may be true in many other instances as well. Suppose more than one

[9] Again, more precisely: Choose an action A from among those actions that satisfy standard S.

[10] There is a subtle distinction between optimization and maximization that the text ignores. Contrast the phrasing of $C1$—"Choose A if there is no better option available"—with the phrasing of $C1'$—"Choose A if A is better than every other option." (Or "Choose A if A is at least as good as every other option." When the agent's ranking is incomplete, no option may satisfy $C1'$ though many satisfy $C1$.

[11] Application of this reformulation does require some care however because it reformulates the agent's preference in a way that reduces the amount of information available. This reduction becomes significant when we weaken rationality by weakening the requirements imposed on the ranking. See below.

option is maximal so the agent has discretion. We might reduce the agent's discretion by refining our ranking. We might again add additional criteria that distinguish among these options.[12]

The second modification of the simple paradigm of consequentialist reasoning severs the direct link between the agent's domain of preference and her domain of choice. Phrased differently, in the simple, fruit basket example, when the agent chooses an act, she directly chooses an outcome. The ranking over outcomes thus translates directly into a ranking over choices. In many, if not most instances, by contrast, the agent's action does not fully determine the outcome. The agent's domain of choice thus differs from her domain of preference, the set of outcomes. Choice becomes more problematic either because we cannot directly derive the agent's ranking of choices from her ranking of outcomes or because she, in fact, has no well-defined ranking of choices at all.

The discrepancy between the domains of choice and preference occurs naturally in many environments. I consider three different environments: choice under uncertainty, strategic choice, and sequential choice.

Consider first sequential choice. Suppose for example that Liza lives for n periods; in each period j she chooses a fruit basket. (For simplicity assume that Liza may neither save nor borrow; in each period she must consume her entire endowment in that period.) Now Liza's domain of preference corresponds to *n-tuples* of fruit baskets, the first in the ordered pair consumed in period 1, and the second consumed in period 2, and so on.

Economists often assume that, in this setting, Liza's preferences over fruit baskets are separable in time; that is, they assume that Liza's evaluation of a fruit basket at time t is independent of the previously consumed fruit baskets and of the fruit baskets to be consumed in subsequent periods. When this assumption holds, Liza has a ranking over fruit baskets that is constant across time periods; consequently, we can easily represent Liza's preferences over consumption histories as a weighted sum of her consumption in each period.[13]

Liza's preferences over consumption histories, however, need not be separable. Her ranking of fruit baskets in period j might depend on her actual consumption during periods 1 through $j-1$. Liza might, for example, have developed a taste for durian or become addicted to coca leaves. This history dependence greatly complicates Liza's consequential analysis as, on traditional accounts of economic rationality, it excludes the choice of certain consumption histories.

Consider the case in which there are many possible fruits, not just apples and oranges, and, among this multitude, one available fruit—call it coca leaves—is extremely addictive. At time 0, Liza has "normal" preferences over fruit baskets and consumption histories of fruit baskets; she ranks most highly those histories

[12] Of course these additional criteria would apply to any pair, even when neither option was maximal. Adding a criterion that refines an existing ranking might thus be difficult as its application must not disrupt the ranking of non-maximal elements though it may also serve to reduce the size of equivalence classes.

[13] The weights represent her time preferences.

that provide her with balanced consumption of fruit types in each period. "Extreme addiction," however, means that, if Liza consumes some coca leaves at time t, her balanced preferences over fruit baskets will be transformed into lexical preferences; i.e., she will prefer most the subsequent history in which she consumes the most coca leaves possible. Thus, if in period 1, Liza consumes any coca leaves, in all subsequent periods she will consume *only* coca leaves.

On these assumptions, Liza's ideal consumption history consists of a mixed fruit basket in each period, one in which she consumes a few coca leaves each period as well as a mixture of apples, oranges, and other fruits. If Liza could commit in period 0 to a consumption plan, no problem arises.[14] She orders her n evenly balanced baskets in period 0; it arrives in each period, and she consumes it.

If, however, Liza is unable to precommit to this pattern, problems may arise for her. Consider two extreme assumptions about Liza's behavior. Suppose she has no foresight at all. In period 1, she then chooses her optimal single period basket; that basket contains coca leaves. Consequently, in every subsequent period, Liza purchases a market basket that consists only of coca leaves. If Liza has foresight, however, she will recognize that she cannot choose this evenly balanced basket in period 1 because she knows that, in each subsequent period, she will choose a basket only of coca leaves. If this is not her second most preferred consumption history, she would do better to eschew coca leaves in all but the last period.[15]

This example may seem contrived and esoteric with little relevance to legal reasoning.

The dilemma posed by sequential rationality, however, is endemic to legal decision. It embodies an inherent conflict between *ex ante* and *ex post* perspectives, or between preventive and acute interventions. Consider the following two, stylized examples.

Consider a family court judge who must first announce a rule governing the division of property and custodial responsibility upon divorce; he must then implement that rule. Suppose that the judge believes that equality demands an equal division of assets accumulated during marriage and an equal division of custodial responsibility. Divorcing spouses, of course, are allowed to bargain away from this rule (see Mnookin and Kornhauser 1979) . This Solomonic rule has many virtues, but it may not be sequentially rational. At some point, a divorcing couple that cannot tolerate each other will appear before the court. They may be unable to agree on any settlement; the rule dictates that the judge apply the Solomonic rule. The judge, however, might believe that though the Solomonic rule sets an appropriate disagreement point for bargaining, in the case before him, it is an extremely unfortunate allocation of

[14] Assuming that she cannot trade in any subsequent period.

[15] Some analysts think that Liza's conundrum undermines the standard economic account of rationality. These critics contend that it is rational for Liza to follow through on her plans even if it is not, at some time t, sequentially rational for her to do so. For different accounts, see McClennan (1990) and Gauthier (1990). The arguments offered against standard rationality in the sequential context are often extended to strategic situations such as the prisoner's dilemma. For further discussion of these issues from a somewhat different perspective see Zaluski, Chap. 6, part II, this volume, on "Interactive Decision Theory and Morality."

custodial responsibility to enforce as it maximizes the interaction between the warring parents and leaves the child in the middle.[16] The judge might thus conclude that she should award primary custody to one of the parents.

Healthcare policy presents similar issues. Preventive care is frequently more cost-effective than acute care. It is more effective to vaccinate against polio than to treat it.[17] A government might then rationally underwrite preventive treatment and bar acute care, or at least, certain, highly expensive forms of acute care. The comparable trade-off for the private individual is clearly not sequentially rational. An individual might face a choice between two healthcare policies: one that provides preventive care but bars acute care and one that provides only acute care. Suppose she purchases the preventive care policy because it is cheaper. If she contracts the disease nonetheless it remains rational for her to pursue the acute intervention and it seems hard to understand what possible reason she would have for adhering to her plan.

Government officials enforcing policy contracts, of course, do not face the same decision problem as the patient. They, too, nonetheless often face a sequential rationality dilemma.

When asked to enforce the bar on acute care, a judge might find it difficult to do; the more dire the disease, the more appealing it will be not to enforce the bar. That difficulty increases when acute care is available on a secondary market, largely to the wealthy while the petitioner before the court is poor. Nonetheless, the public official must take into account the systemic effects of an *ex post* revision of the terms of the policy because the funding arrangements for the plan, whether through premia or taxes, were set under assumptions about the covered procedures.[18]

Let us return to Liza's sequential decision problem. We might interpret Liza's fundamental preferences over consumption histories differently. In the sequential choice problem, Liza has a multi-criterial objective function; she cares about her consumption in each period and we must ask how she integrates the value of consumption in each period into an all-things-considered evaluation of complete consumption histories. The assumption that her preferences are separable across time implicitly adopts a very simple and naive process of integration. Decision-making under uncertainty may be interpreted similarly; moreover, it provides a better context in which to discuss the problems of integration.

[16] One might argue that courts do not first announce a rule and then apply it. But in a common law jurisdiction, we can imagine the announcing rule in a case in which both parents have the best interests of the child at heart and thus will achieve a reasonable allocation of custodial responsibility. In the original story, of course, Solomon announces the rule and then does not enforce it. See also the Chinese play *The Chalk Circle* and Brecht's variation on it, *The Caucasian Chalk Circle*."

[17] For some diseases, this is not true; it is more cost-effective to treat cholera than to vaccinate against it. This trade-off reflects the relative inefficacy of the vaccine and the low cost of treatment (in a well-functioning public healthcare system).

[18] The problem facing the government official may look different from the problem of sequential rationality because the legal institution that set the policy differs from the legal institution that enforces or implements it. In thinking about these issues, however, it is helpful to think of the public officials as a team that share an objective function. On this assumption, the problem of sequential rationality re-emerges.

In this context, the realized outcome depends both on the agent's choice of action and the realized state of the world. Suppose that, in the fruit basket example, Liza buys not apples and oranges but apple futures and orange futures; contracts, that is, for delivery of specified amounts of apples and oranges at a future date. The value of these contracts thus depends on the size of the harvests, i.e., the realized supply at the time of delivery. In this example, Liza chooses a *portfolio* but she has preferences over fruit baskets.

For simplicity, suppose the harvest of apples may be either High (*HA*) or Low (*LA*); similarly, the orange harvest may be either High (*HO*) or Low (*LO*). There are thus four possible realized states of the world (*HA*, *HO*), (*LA*, *HO*), (*HA*, *LA*), and (*LA*, *LO*).

Presumably, Liza's evaluation of a given bundle of apple futures and orange futures differs across these realized states of the world.

For Liza to reason consequentially, she must somehow evaluate each of her actions; this evaluation must depend on the four possible outcomes, each corresponding to the state of the world in which it would be realized, to which her action might lead. Liza thus has four distinct criteria against which to assess any given action, i.e, any given portfolio. Each criterion corresponds to the outcome realized (i.e., the realized fruit basket) in a given state of the world.

Liza might integrate these four criteria into a single ranking in many different ways. Some of these integrations yield partial rankings; others complete rankings. Liza, for example, might say that one portfolio P is better than another basket P' if and only if, fruit basket B in each state of the world in portfolio P, has at least as many fruits of each kind as the corresponding basket B' in portfolio P' and there is at least one state of the world in which the basket in P has more of at least one type of fruit than the corresponding basket in P'. This strategy of integration of the multiple criteria used to evaluate actions into an all-things-considered ranking of them is not complete; there will be some pairs of portfolios for which Liza will not be able to say that one portfolio is better than another or that they are equally good.

Liza might integrate these criteria into a *complete*, all-things-considered ranking in a number of different ways. She might use a maximin criterion that evaluates each portfolio in terms of the basket that she ranks lowest (of the four baskets held in the portfolio). She then prefers portfolio P to portfolio P' if and only if she prefers her least preferred basket in portfolio P to her least preferred basket in portfolio P'. Or Liza might weight each outcome in the portfolio by the probability that each outcome will occur; this weighting produces an *expected fruit basket*; she then prefers P to P' if and only if the expected fruit basket of portfolio P is better than the expected fruit basket of portfolio P'.

A third possibility exists; Liza may fail totally to integrate her multiple criteria into a ranking at all. She may not have well-defined preferences over the portfolios. This possibility is more easily understood in the context of strategic choice where, again, the agent's domain of choice differs from her domain of preference.

Game theory provides a formal, mathematical framework for the analysis of strategic choice.[19] Consider the simple game form presented in matrix 1:

Matrix 1: a simple game form

		Column		
		L	C	R
Row	T	a	b	c
	M	d	e	f
	B	g	h	j

Matrix 1 shows that each of two players has a choice set that consists of three strategies or actions. There are thus nine possible plays of the game; each possible strategy pair gives rise to some outcome, represented by the letter in the corresponding cell.

According to consequentialism, each agent's choice depends on her ranking of the outcomes. Game theorists agree; they derive the game matrix from the game form by reference to each player's ranking of the outcomes.

Suppose that Row ranks these outcomes as follows: $c > j > e > f > b > h > g > d > a$ while Column ranks the outcomes $g > j > e > h > f > d > b > c > a$. Assigning the number 8 to each player's most preferred alternative and then successively lower integers to successively less preferred outcomes yields the game matrix represented in matrix 2.[20]

Each player ranks the outcomes. But she does not unilaterally choose an outcome.

Rather she chooses one of three strategies. If each player had an unambiguous ranking of the three strategies, $C1$ would dictate that she choose the one that was best.

Matrix 2: a simple 3 × 3 game

		Column		
		L	C	R
Row	T	0, 0	4, 2	8, 1
	M	1, 3	6, 6	5, 4
	B	2, 8	3, 5	7, 7

[19] See Zaluski (Chap. 6, part II, this volume, on "Interactive Decision Theory and Morality") for a more extensive development of the theory of games. He does not address, however, the issue discussed in the text.

[20] Notice that players with different preferences over the outcomes in matrix 1 will play a different game. It may be helpful to restate the structure of the argument in terms of the prisoner's dilemma. That game form consists of the matrix that has, in each cell, the sentence that each prisoner receives given the pair of strategies that correspond to that cell. The standard prisoner's dilemma matrix assumes that each player ranks the outcomes solely on the basis of the sentence he receives and that he prefers shorter sentences to longer ones.

Does Row have a well-defined ranking of her choices T, M, and B? Row might treat her choice of strategy equivalently to a choice of an action under uncertainty. She thus treats Column's choice as an uncertain state of the world. The discussion above suggested that, on this interpretation, Row should recognize that she has a multi-criterial objective function and she must integrate the different criteria into an all-things-considered judgment.

This approach has some obstacles to overcome. Notice that Row prefers M to T if Column chooses C, but she prefers T to M if Column chooses R. Similarly, she prefers M to B if Column chooses C but B to M if Column chooses R.[21] Row does not obviously have a well-defined ranking over the set of options $\{T, M, B\}$. As none of her options dominate any others, Row would require a more complex way to integrate the different criteria she has for evaluating outcomes.

On the other hand, perhaps integration is not the appropriate approach in the strategic context because Column's choice is not a random event; it is a conscious, indeed rational, choice. Row might profit from an analysis of Column's reasoning. Game theory provides such analyses on the assumption that both players are rational.

Each solution concept provides a different resolution of the choice problem. Maximin requires Row to consider, for each strategy, the worst outcome that is possible and then to choose the strategy that has the best "worst" outcome. For Row, this strategy is B. For Column, maximin requires him to choose C. These choices would yield the outcome (B, C) with payoffs (3, 5)—i.e., the fourth worst for Row and the fourth best for Column.[22] Maximin, however, is not fully satisfying as a concept of rationality in this game because $C1$ suggests that Row should not be content with B conditional on Column having chosen C. Both Row and Column would do better if Row chose M.[23]

C1 thus suggests an alternative conception of rationality[24] that emphasizes that each player must, in equilibrium, conform to $C1$. This concept, Nash equilibrium, thus states that each player's choice of strategy must be a best response to the other player's strategy choice. In matrix 2, the pair (M, C) is the only Nash equilibrium. This concept of rationality thus does not yield a ranking of choices; it simply identifies a set of acceptable choices.[25]

[21] And she prefers T to B if Column chooses R but B to M if Column chooses L.

[22] Notice that maximin, as it does in the case of decision-making under uncertainty, does yield a ranking of each player's choice set. Row's ranking is $B > M > T$ while Column has the ranking $C > R > L$.

[23] Of course, Column under $C1$ would also not be content with C conditional on Row playing B. Column would prefer to play R. And as before, both Row and Column are better off at (B, R) than (B, C).

[24] Note that maximin is not inconsistent with $C1$; it just reflect an extreme degree of risk aversion. On an alternative account, under Maximin, Row assumes that Column has malevolent preferences and will act to maximize them. But if Column has the preferences he does have and if Row engages in strategic reasoning $C1$ suggests Nash equilibrium.

[25] One might think about the failure of integration of the agent's multiple criteria into a coherent all-things-considered ranking in another way. An agent with multiple criteria seeks to integrate them into a single ranking. As the prior discussion suggests, many different integrations are possible. A

Game theoretic analyses of the game represented by matrix 2 differ because different solution concepts rest on different conceptions of rationality. The maximin solution concept, for instance, parallels the precautionary principle as a method of integration of states of the world in decision-making under uncertainty.

Consequentialist reasoning rests on the agent's evaluation of certain outcomes. In any realistic setting, however, there are no certain outcomes; neither time nor uncertainty ends.

Some agent has further choices and some further uncertainty will be resolved. The decision-maker must end the regress; she does so by characterizing what Savage (1954) calls a "small world." A small world is a (relatively) self-contained and can be analyzed and understood in isolation from the unending choices and reverberations of choice that an agent undertakes. As Savage notes, however, it is not clear how to define the appropriate small world to guide decision-making.

The problem of small worlds is particularly pressing in legal contexts. Many policies decisions have far-ranging consequences. Global warming presents perhaps the starkest example. Emission of greenhouse gases now has consequences hundreds of years into the future. On a more mundane level, consider the budgeting process of the United States. Proposed bills are assessed in terms of their budgetary consequences. Under Senate rules, any bill that has significant budgetary impact after ten years may be blocked during the reconciliation process. In order to enact his tax cuts in 2001 and 2003, President Bush made them "temporary"; they "expired" after ten years; they thus had no "legal" budgetary consequences after ten years and were thus not subject to being blocked in the reconciliation process. Of course, the act had significant political consequences that had significant legislative consequences.

2.3 *Concluding Comments*

This discussion demonstrates that teleological reasoning is virtually co-extensive with economic or rational choice reasoning.[26] Teleological reasoning directs the agent, given her ends, to do the best she can. In very simple decisions, teleological reasoning leads to a simple choice of a maximal element of her ranking. More

rationality of ends thus requires not only a specification of rational ends or criteria against which to evaluate options but also a specification of rational conditions of integration. Little attention has yet been given to the identification of appropriate conditions of integration in policymaking. Which conditions are appropriate will depend on the nature of the decision problem. As we have seen, economists have intensively investigated the appropriate conditions for the state-by-state and period-by-period integration of a single decision-maker's criteria. The second area in which intensive investigation of criteria of integration has occurred is social choice theory. Social choice theory investigates the conditions under which individual rankings can be integrated into a social ranking. Arrow (1963) launched the discipline with a result that showed that four conditions of integration were logically incompatible.

[26] Of course the foundations of decision theory are themselves disputed. A number of different formal characterizations of consequentialist reasoning have been advanced. See for example Hammond (1988) and McClennan (1990). These authors study sequential decision-making to probe the underlying premises of rationality so understood.

generally, however, how well the agent can do and how she should best do it, depend on her options and the environment in which she chooses. Often the agent must choose among options that she does not value as ends but only as means to her ends. Equally often the outcomes she values depend not only on her own choices but on the choices of others. In these circumstances, the requirements of teleological reasoning become more complex and more controversial.

This essay does not pursue either the complications that lie at the foundations of decision theory or the role of teleological reasoning in non-consequentialist theories. Rather I turn to the particular implications and challenges thinking in and thinking about law pose to the general framework of consequentialist reasoning. This essay has three further parts. First, I develop the connection between teleological reasoning and a particular view of law that I shall call legal instrumentalism. Legal instrumentalism regards "law" as an instrument to achieve collective ends. Legal instrumentalism is a protean concept; its exact shape in any given legal system determines the scope of teleological reasoning in law.

Sections 4 and 5 address the two central aspects of teleological reasoning in law. Any teleological reasoner must first identify the ends to pursue and then identify the appropriate means for achieving those ends. In the context of law, the ends to be pursued are *collective* ends usually articulated in constitutions, statutes, regulations, and judicial opinions. The collective nature of these ends presents special problems to any policy maker but the nature of the complications may vary across legal institutions. Section 4 considers these problems choosing ends.

Section 5 investigates the choice of means.

3 Teleology and Instrumentalism

Teleological reasoning imputes a goal, embodied in a ranking, to "law." Individuals who deploy teleological reasoning about law thus view the law to some extent instrumentally. But instrumentalism bears a complex relation to teleological and consequentialist reasoning about and within law. This complexity has two sources. First, we must clarify what it means for the law "to serve a purpose." Second, we must refine our conception of the "law" that serves this purpose.

We may usefully distinguish at least three conceptions of the law "serving a purpose."

First, an external analyst might offer a *functional* account of law. A functional account of law identifies a social function that law serves; it has the structure "legal rule or institution L promotes socially valuable goal G." An adequate functional account must then identify the causal mechanism that explains how L came to promote G. The intentional action of public officials might be one such causal mechanism; the identification of this intentionality would be an instrumental account of L. More often, however, the analyst argues that L emerged not because public officials intended L to promote G but as an unintended consequence of other actions of legal and non-legal actors. An analyst who gives a teleological account of law would thus use teleological reasoning about the law but agents within the legal system would not use teleological reasoning understood appropriately.

The other two accounts of "law serving a purpose" assume that agents act intentionally but agents within the legal system of course might engage in teleological reasoning in at least two different ways. Each agent might use the law to promote her own ends. Litigants typically act in this fashion. Public officials might act in a similarly self-interested fashion. Such public officials would deploy teleological reasoning but this reasoning would be no different in character or aim than any other teleological reasoning in which they might engage to promote their ends. Call this form of legal instrumentalism, "weak" instrumentalism. Weak instrumentalism holds that individuals use the law to promote individual aims not collective aims.

Public officials might, however, endorse a stronger form of legal instrumentalism. An official might believe that the "law," or more accurately, the polity, has its own aims that she, acting in her official capacity, should pursue. The official would then use the law instrumentally to further these collective ends. She would reason teleologically taking the collective ends as the goals she must further.

Strong instrumentalism of this sort might take different forms that vary with the nature of the legal institution used as an instrument. We may imagine *systemic*, *institutional*, and *rule* instrumentalism. Under rule instrumentalism, the policymaker—a judge, a legislator, or an agency—treats the rule as a means to an end; she applies teleological reasoning to the aim of the legal rule to enact, interpret or apply it appropriately. According to (strong) rule instrumentalism, the purpose of the rule derives from the "intentions of the rulemaker"—a constitutional convention, a legislature, an administrative agency, or a court. As the next section indicates, however, diverse legal actors employ a diverse and conflicting panoply of methods to discern the "intention of the rulemaker" when they ascribe a purpose to the rule.

Rule instrumentalism may consequently impute different purposes to different rules either because different rulemakers promulgated each of the different rules, or because the same lawmaker undertook to promote a different aim with each rule enacted.

Under institutional instrumentalism, a designer treats the legal institution as a means to an end. Here the designer might be the drafter of a constitution, or a legislator creating a complex program that delegates significant rule-making power to an agency, the executive or a court. We might, for example, consider the constitutional creation of a bicameral legislature as instrumentally designed to make legislation slow and difficult. The constitutional drafters would have reasoned teleologically to consider the nature of legislation that would be enacted by a unicameral legislature as compared to a bicameral one. Notice that institutional instrumentalism may impute different purposes to different institutions.

It is important to distinguish between the intentions of the designer of the institution and the role and intentions of the agents *within* the institution. One might, for instance, have an institutionally instrumental conception of courts—one that attributes a purpose to judicial institutions that guide the development of that institution—but deny that judges, in deciding cases, should deploy teleological reasoning. One should thus reason teleologically about the structure of judicial institutions without thinking that judges act in a rule instrumental fashion. Phrased differently, the institutional designer reasons teleologically in its creation of the institution; but the designer might create an institution the agents of which are directed to reason deontically or formally. Adjudication might thus be institutionally instrumental but judges, within this instrumental institution might be directed to reason non-teleologically.

Finally, under systemic instrumentalism, the policymaker treats the entire legal system as a means to an end. The complex design of the system of separation of powers in the United States, for example, may be understood as a device created to curb the power of government by dividing governmental powers among the branches. None of the three branches, of course, is designed to curb governmental power, or even to curb its own power. The structure of the system accomplishes this goal and constitutional interpreters should, perhaps, elaborate the constitutional structure in light of this goal.

Most legal cultures have a (strong) rule instrumentalist view of legislation.[27] Citizens generally believe that the legislature enacts statutes to further some social goal and not merely to express approval or disapproval. When the legislature enacts a statute governing water quality, citizens understand that the legislature seeks to improve water quality; the legislature has not simply stated that high water quality is good thing and then hoped for the best. Similarly, legislation that prohibits employment discrimination seeks to create a labor market of a certain type. Legal culture in the United States is strongly rule instrumentalist all the way down: not only with respect to the legislature and the executive but also with respect to the judiciary.

Not all legal cultures share this strong rule instrumental view of adjudication. Civil law cultures, for example, do not typically require judges to be rule instrumental. These cultures, however, typically understand adjudicatory institutions instrumentally. Thus the legislature in designing the court system reasons teleologically but the judges within the court system are directed to deploy formalist modes of reasoning.

4 Choosing Ends

Consequentialist reasoning has focused on instrumental rationality, the logic of the appropriate choice of means to a given end. All, or at least many, of our ends, however, are not given; they are chosen or developed. Both the psychology and the rationality of the choice of individual ends is relatively undeveloped. Choice of (individual) ends has two aspects. The agents must determine which ends to pursue[28]; she must then determine how to integrate these ends into a ranking over the actions available

[27] Public choice theory denies that legislatures act in a strongly instrumental way. Rather scholars such as Buchanan adopt an institutionally instrumental stance toward legislatures but expect legislators within the institution to act self-interestedly.

[28] Frankfurt (1982), Schmitz (1996), and Richardson (2003) offer accounts of the problem of the identification of ends.

to her at any point. I shall largely put this second difficult problem to one side.[29] I shall focus instead on a related set of problems that arise when we shift attention from individual ends to legal, *collective* ends. Most legal activity and legal reasoning confronts this issue at the outset. We often speak of the "ends of the law" but the "law" on one account is simply a set of norms. On a second account, "law" is a set of institutions. Each official within each institution may have ends but it is unclear how each of these individual ends relates to the ends of the "law." A central question for teleological reasoning in law, therefore, asks: how do we determine our collective, legal ends?

How we determine our collective aims depends, in part, on *who* makes the determination.

The legislative determination of collective ends differs in important respects from the determination of legal ends in non-legislative contexts.

4.1 The Legislative Determination of Legal Ends

Legislatures enact statutes.[30] Statutes typically embody legislative ends as statutes are usually enacted to solve social problems or promote social goals. Call the ends embodied in the statute its "statutory purpose."[31] The statutory purpose of an enactment is important because it may dictate the legal ends that subsequent legal actors adopt when reasoning teleologically with that enactment.

If the legislature consisted of a single individual, our understanding of the statutory purpose would be relatively straightforward. It would correspond to the aim or intention of the unitary legislator. That legislator would, presumably, have reasoned teleologically in enacting the statute. After identification of the problem to be addressed, the legislator must identify potential legislative solutions and compare their efficacy. Sartor (2010) offers a normative account of the role of teleology in this legislator's reasoning.[32]

[29] We shall see, however, that the problem of integration of ends into a single ranking parallels an aspect of the second problem of determining collective ends. For further discussion of the problem of integration of ends, see Kornhauser (1998).

[30] In this essay, I put to one side the drafting of constitutions. As constitutions are typically drafted by assemblies—either constitutional conventions or legislatures—many of the same issues arise in the determination of constitutional ends as in the determination of legislative ends. Additional complications, however, may arise from the fact that constitutions are typically ratified by the citizenry.

[31] I refer to the collective end embodied in any legal norm as its "statutory purpose." Not only statutes and ordinances but also administrative regulations and judge-made rules thus have statutory purposes. As these norms are typically announced by collegial bodies, they present similar questions concerning the construction of the statutory purpose.

[32] Sartor, however, ignores two issues. First, he ignores the multiplicity of legislators. As the next paragraphs of the chapter indicate, legislators may disagree about the nature of the problem to be solved, the set of constraints the legislature faces, and the causal consequences of each alternative solution. A complete normative theory of legislator behavior would provide guidance to legislators

Legislatures, however, are collective bodies. Legislatures enact statutes together, not separately. Each legislator intended to vote for or against the enactment of the statute. A majority chose to enact the statute. The members of the majority need not have agreed on its meaning or purpose. Indeed, each legislator in the majority, moreover, may have had radically different intentions and understandings of the statute enacted. Which understanding constitutes the statutory purpose?[33]

Consider, for example, Hart's hypothetical ordinance that bans vehicles in public parks (Hart 1958). Legislator *A* may have understood the ordinance as primarily a safety measure. Legislator *B*, by contrast, may have understood the ordinance as primarily a measure to enhance the public's enjoyment of the park by maintaining quiet. Legislator *C* voted against the ordinance. Though each member of the majority of the city council enacting the ordinance ascribed a purpose to the enactment, it was not the same purpose. It is thus unclear what the statutory purpose of the ordinance is.

Public officials that subsequently apply the ordinance nonetheless may require an understanding of the ordinance's purpose. At the very least, they require a theory of interpretation to determine the meaning and applicability of the ordinance.

4.2 Interpretation and the Determination of Legal Ends

An enacted statute guides the conduct of public officials and citizens. No statute, however, anticipates all of the myriad situations to which it may apply. Consequently, those governed by the statute will often seek guidance in circumstances when the statute is silent, ambiguous, or contradictory. In these instances, the statute must be interpreted.

There are many different types of interpreters and many different theories of interpretation addressed by a vast literature (see, e.g., Marmor 1995, 2005). I shall distinguish two classes of interpreters and two classes of interpretive theories.

First, distinguish sophisticated from unsophisticated interpreters. Judges and high-level administrators are sophisticated interpreters. Citizens and low-level public officials are unsophisticated. Sophistication depends both on the amount of time the agent has to deliberate over the appropriate interpretation of a legal norm and the extent of her knowledge of or access to relevant interpretive materials.

Second, distinguish between purposive and non-purposive theories of interpretation. Purposive theories of interpretation direct the interpreter to identify the statutory purpose of the legal norm and to interpret it accordingly. Purposive interpretive theories thus direct the interpreter to reason teleologically. Non-purposive theories

on the resolution of these conflicts. For a parallel argument regarding normative theories of adjudication, see Kornhauser (2013). Second, he ignores the question of representation, i.e., the relation of the legislator's views to the views of her constituents.

[33] I put to one side questions raised by the literature on social choice theory and judgment aggregation. These theories start from the assumption that the statutory purpose should reflect the preferences, beliefs, and judgments of the legislators. Aggregation of these attitudes must confront various logical problems. For a discussion see List (2008).

interpret the legal norm without regard to its purpose. Rather the interpretation may rely simply on the text or “plain meaning” of the statute. Or it may refer to the intentions of the enacting legislator in which case the theory must articulate how such an intention is identified.

Different theories of interpretation might be appropriate for different interpreters. One cannot expect a policeman in the course of a traffic stop to deliberate in the same way about the reasonableness of a search and seizure as a judge contemplating the admissibility of the evidence discovered from such a search.[34] The right interpretive theory for the unsophisticated interpreter may thus be, at least in many contexts, non-purposive.

The arguments for the use of purposive theories by sophisticated interpreters are much stronger but will not be rehearsed here. Rather I shall discuss how purposive theories direct the interpreter, in interpreting the statute, to reason teleologically.

Return to Hart’s example introduced in the prior subsection. Recall that the prohibition on vehicles might have at least two purposes: to promote safety in the park and to promote quiet enjoyment of the park. Many instances may arise in which the presence of a vehicle would not impinge on these purposes or would impinge on one but not the other.

Hart introduced a hypothetical in which a monument, that includes a tank, is installed in the park. As the tank is not functioning, it neither creates a hazard nor produces noise. A bicycle, by contrast, may create a safety hazard, particularly for children, though it generates no noise. A child’s tricycle, on the other hand, is equally quiet but less dangerous. In interpreting the ordinance, a judge confronted with these cases must first articulate the purpose of the ordinance and then choose the interpretation that best promotes her understanding of the purpose of the ordinance. In these examples, the teleological reasoning is quite simple as the judge has only two options: permit the vehicle (or class of vehicles) in the park or prohibit it. She simply makes the decision that best promotes the statutory purpose.[35]

Thus, to interpret the ordinance, the judge reasons teleologically. She considers various possible interpretations in light of her understanding of the statutory purpose underlying the ordinance. She then chooses the interpretation that best advances that statutory purpose.

5 Choosing Means

Teleological reasoning has three stages. First, the agent identifies her aim. The prior section offered brief comments into this stage. Next, she must determine the set of available means to achieve this aim. Finally, the agent assesses each available option

[34] Rather we expect the policeman to decide on the basis of a departmental practice that perhaps was elaborated on the basis of a purposive interpretation of the legal norm. The officials promulgating the regulations governing police searches and seizures are sophisticated interpreters.

[35] Some cases are more complex. Consider a motorized wheelchair that might both pose a danger to children and be noisy. On the other hand, the judge presumably faces a set of background constraints about equal concern and respect for individuals with disabilities that might argue for permission.

against the criteria embedded in her aim and chooses the option that is (among the) best. This section offers some brief remarks about the latter two stages.

5.1 Identifying Options

To begin, consider the second stage of defining the available options. In economics textbooks, this stage is straightforward; the agent's set of *possible* options are self-evident; the agent simply needs to identify her budget constraint to identify her set of *feasible* options. She then proceeds to the third stage to choose a maximal, feasible option.

For agents faced with actual decision problems, however, the identification of options is difficult for at least three interrelated reasons: (1) our goals are often unclear; (2) our constraints are often unclear; and (3) the causal mechanisms linking actions and outcomes are often unclear.

To illustrate these difficulties, consider the USA in the early 1950s. It has highly segregated schools. Congress seeks to integrate those schools. What options are available to Congress?

The set of possible policies seems endless. Suppose Congress decides to use monetary incentives. Congress could subsidize schools that more closely approximate its ideal demographic composition. That of course requires Congress to specify this ideal sufficiently precisely to measure deviations. The difficulty in articulating this ideal demographic reflects uncertainty about our goals and causal mechanisms. Should the ideal demographic composition of a school reflect the demographic composition of the population in the jurisdiction? In the metropolitan area? In the state (or states) in which the metropolitan area lies? Should it reflect the school-age population generally? Or the school-age population in public schools?

The choice of baseline reflects different views about the importance to attach to choice or autonomy in characterizing the social aim. Similarly, the choice of baseline may reflect different understandings of the causes and extent of residential mobility in response to the legal rules. Or the choice of baseline might reflect an understanding of the constraints on Congress to impose remedies that cross-jurisdictional lines.

Notice that different legal actors may face different constraints and consequently different sets of available options. The United States Supreme Court struck down the laws sustaining segregated schools in 1954.[36] It then had to announce a remedy. Institutional constraints excluded the subsidy scheme outlined in the prior paragraphs from the Court's set of possible remedies. Similarly, in elaborating a remedial scheme that involved busing, the Court understood itself to be bound by the jurisdictional boundaries of existing systems; it refused to require busing across district lines though a remedy of that type would have been legislatively available.[37]

[36] Brown v. Board of Education 347 US 483 (1954).

[37] One could understand the Court's reluctance to require cross-jurisdictional busing as reflecting a different understanding of the relative weights of federalism and equality rather than as a different constraint.

Courts on occasion explicitly recognize these constraints on their set of remedial options by remitting the remedy to the legislature. The New Jersey Supreme Court, for example, held the state system of funding public education was unconstitutional and mandated that the state legislature remedy the situation. The Court then ruled on successive attempts by the state legislature to remedy the situation.[38] The Court's procedure thus recognized the limited set of options available to the Court for direct implementation.

Though administrative agencies often make rules in a legislative manner, they too face different constraints than legislatures and courts. Sometimes these constraints are written into the authorizing statute. But many constraints derive from the political context in which administrative agencies act.

5.2 *Choosing the Best Option*

After determining her ends and identifying her options, the agent must evaluate the identified options against the determined ends. The difficulty of this task depends on the substance of the ends.

Legal decision-makers typically choose policies or institutions.[39] When ends are procedural, the evaluation of the options may be relatively straightforward. A state might, for example, be creating an agency that has technical expertise but that is also responsive to the relevant interest groups. The options consist of various agency structures that specify both the appointment and tenure of the commissioners that will lead the agency and the staffing that will provide technical expertise. Ranking the options in this instance is relatively straightforward.

The statutory purpose of the legal decision-maker, however, is often substantive. Legislatures typically seek to solve social problems: to reduce air and water pollution, to insure the integrity of the food supply, to improve the functioning of the labor market, to improve the quality of primary and secondary education, to provide universal health care, and many others. To assess the available policy options, the policymaker must determine the consequences that each of the different policies will have.

In legal contexts, policymakers generally face great uncertainty about the causal relations between the instruments available to them and the consequences that flow from these means. This uncertainty concerns more than uncertainty about the state of the world. It also includes scientific uncertainty about the underlying causal mechanisms. As the global warming example indicates, we do not understand completely

[38] Abbott v. Burke 100 NJ 269, 495 A.2d 376 (1985), Abbott v. Burke 119 NJ 359, 575 A.2d 359 (1990), Abbot v. Burke 136 NJ 444, 643 A.2d 575 (1994), Abbott v. Burke 149 NJ 145, 693 A.2d 417 (1997), Abbot ex rel. Abbott v. Burke 199 NJ 140, 971 A.2d 989 (2009).

[39] Some legislative acts concern projects rather than programs. In the mid-1950s, for example, the United States Congress began a project to construct an interstate highway system. In subsequent years, Congress appropriated money to fund various stages of this project. The teleological reasoning underlying the appropriation bills was trivial.

the physical science underlying the processes that yield global warming. Our understanding of the underlying social phenomena is even thinner. How will individuals respond to the legal rules we adopt taxes, fines, tradeable permits, etc.—to control greenhouse gas emissions?[40]

One common attack on teleological reasoning in law rests on its extreme difficulty. Determination of the consequences of a policy is extremely difficult. Consider, for example, the straightforward policy change of installing parking meters in a downtown business district.[41] Installation of parking meters might have a large number of short-run and longer-run consequences. Predicting these consequences will depend on the exact specification of the policy and on a large number of features in the business district: how much off-street parking is available (and at what price); the exact extent—i.e., number of blocks—of the metered zone and its price; the availability of free on-street parking at the border of the metered zone; the nature of the businesses within the metered zone and the nature and location of their competitors. Assessment of the policy requires predictions of the change in traffic flows and the attendant effects on air pollution, of the change in business receipts in the short-term, and, over the long-term, changes in the land values and store rentals and the associated changes in the businesses that occupy these storefronts.[42] Over a longer run, we might expect to see more dramatic changes in land uses: if demand falls, land dedicated to off-street parking might be converted to commercial or regimental use; if demand rises, we might expect the opposite.

Predicting the consequences of many other policies present significantly greater challenges. The challenges are perhaps greatest in the context of constitutional design. Consider a polity choosing between a legislative structure that consists of a bicameral legislature subject to a presidential veto (and a potential override) and a unicameral parliamentary structure. (Call the first a "presidential system" and the second a "parliamentary system").

What predictions must the policymaker make?

To understand the complexity of the task of assessment, it may help to specify a criterion against which the systems are to be assessed. Suppose that the policymaker seeks to maximize social welfare. A fully consequentialist assessment would then rank the two systems on the basis of social welfare generated by each. This task, of

[40] A number of questions arise concerning how a policymaker should make these predictions. Some authors—most obviously, Brennan and Buchanan (1981, 1985) have argued that the policymaker should use rational choice theory to predict the consequences of different legal rules even if that theory is false. Kornhauser (2002) argues against their claim and suggests that, at least in some instances, prediction does not require an explanatory theory at all.

[41] For a discussion of the consequences of installing parking meters in the central business district of Pasadena, CA, see Shoup (2005). His entire seven hundred page book addresses the dramatic (negative) consequences of a range of policies that constitute "free parking."

[42] I have merely indicated some of the economic consequences. A policy change may have political consequences as well. The neighborhoods on the border of the metered zone may be upset by the increased demand for free parking in their neighborhoods. Similarly, if the business drops in the metered zone, the policymaker can expect political anger from the businesses there; if business increases, she may be targeted by those businesses outside the metered zone that have lost customers. Of course, if the losing businesses are outside her jurisdiction she may not care.

course, approaches the impossible. Suppose, for instance, that both the population in the polity and the environment in which that population acts are fixed and unchanging. The predictive task still requires the policymaker to first determine the legislation that each system will enact and then to predict the effects on social welfare of each piece of legislation. To make the first prediction, the policymaker must understand the nature of politics in the jurisdiction; she must know the preferences of each citizen and what institutional structures of politics will emerge.[43] To evaluate the institutions on the basis of consequences, however, requires that the policymaker extend the analysis further; she must also predict what consequences the enacted statutes will have.

One might circumvent the difficulty of predicting distant and complex consequences by adopting a different set of criteria against which to assess institutions or policies. One might, for example, adopt more procedural criteria against which to assess the policy or the institution. Or one might adopt criteria with shorter time horizons.[44]

Legal actors are differentially placed to make predictions of these types. Administrative agencies are generally designed to make technically complex decisions; they typically have staffs that have the relevant technical expertise and that generate the data relevant to assessing the uncertainties and technical complexities that many of these decisions entail. Legislatures too have a capacity to generate information and to call upon expertise in formulating policy. The legislature can hold hearings and deliberate for years prior to action. Courts, by contrast, are not well designed to make these technical decisions. Fact finding in adjudication tends to be retrospective rather than prospective. The court determines the facts underlying the dispute rather than the forwarding-looking facts needed in teleological reasoning. The court, moreover, must decide the case in a timely manner.

6 Concluding Remarks

Teleological reasoning requires an agent to choose wisely (or, at least, rationally) to promote her ends. This straightforward direction conceals a number of subtleties. It requires, for example, that the agent be able to rank, at least partially, potential consequences of her choices.

A legal actor engaged in teleological reasoning must first identify her ends, then identify feasible policies that promote those ends, and finally choose the means that best promote those ends. Each task is fraught with difficulty. In democratic polity, we might demand that the chosen ends satisfy two conditions: (a) the policies should reflect the normative and political judgments of the citizens and (b) its policymaking

[43] The Framers of the US Constitution did not anticipate the emergence of a party system.

[44] So, for example, the United States Congress uses a ten-year time horizon to assess the budgetary impact of legislation.

bodies should exhibit a high degree of coherence across policy decisions. Neither condition is easily met.

When legislators enact statutes, when administrators promulgate regulations, or when collegial courts announce legal rules, they typically disagree on the purpose of the promulgated norm. The legal actors that apply that norm as well as the citizens that must abide it must nevertheless interpret the promulgated norm and identify its statutory purpose. Purposive theories of interpretation direct interpreters to deploy teleological reasoning to perform this task.

The choice of means faces challenges equally great. Agents make policies that often have complex social consequences far into the future. In framing a policy, the agent faces great uncertainty that ranges from ordinary day-to-day variation in the weather to long-term variation in the patterns of weather variation. She is apt to have little understanding of physical and biological processes. The rate and nature of technological change is unknown as are the causal processes that determine the effects different legal policies will have on individual behavior.

The decision-maker, moreover, generally does not have complete control over the consequences her actions have. She may choose under uncertainty or in a strategic context in which the choices of other also determine the outcomes.

The challenges of teleological reasoning by legal agents do not argue for its abandonment. Legislation enacted without contemplation or concern for the consequences it engenders would be foolish indeed.

References

Arrow, K.J. 1963. *Social choice and individual values*. New Heaven: Yale university press.

Brennan, G., and J. Buchanan. 1981. The normative purpose of economic “science” rediscovery of an eighteenth century method. *International Review of Law and Economics* 1: 155–166.

Brennan, G., and J. Buchanan. 1985. *The reason of rules: constitutional political economy*. Cambridge: Cambridge University Press.

Broome, J., 1991. *Weighing Goods*. Oxford: Basil Blackwell.

Chang, R. 2001. *Making comparisons count*. Abingdon: Taylor and Francis.

Frankfurt, H. 1982. The importance of what we care about. *Synthese* 53: 257–272.

Gauthier, D. 1990. *Moral dealing: contract, ethics, and reason*. Ithaca, NY: Cornell University Press.

Hammond, P. 1988. Consequentialist foundations for expected utility. *Theory and Decision* 25: 25–78.

Hansson, S.O., and T. Grüne-Yanoff. 2012. Preferences. *The Stanford Encyclopedia of Philosophy*, ed. E. Zalta. http://plato.stanford.edu/archives/win2012/entries/preferences/.

Hart, H.L.A. 1958. Positivism and the separation of law and morals. *Harvard Law Review* 71: 593–629.

Kornhauser, L.A. 1998. No best answer? *University of Pennsylvania Law Review* 146: 1599–1637.

Kornhauser, L.A. 2002. Virtue and self-interest in the design of constitutional institutions. *Theoretical Inquiries in Law* 3: 21–47.

Kornhauser, L.A. 2003. The domain of preference. *University of Pennsylvania Law Review* 151: 717–746.

Kornhauser, L.A. 2013. Deciding together. *NYU Law and Economics Research Paper* 13–37. http://ssrn.com/abstract=2332236.

List, C. 2008. *Judgment Aggregation: A Short Introduction*. http://personal.lse.ac.uk/list/pdf-files/ja-intro.pdf.

Marmor, A. (ed.). 1995. *Law and interpretation*. Oxford: Clarendon Press.

Marmor, A. 2005. *Interpretation and legal theory*. Oxford: Hart Publishing.

McClennan, E. 1990. *Rationality and dynamic choice*. Cambridge: Cambridge University Press.

Mnookin, R., and L. Kornhauser. 1979. Bargaining in the shadow of the law: the case of divorce. *The Yale Law Journal* 5: 950–997.

Rawls, J. 1971. *A theory of justice*. Cambridge, Mass.: Harvard University Press.

Richardson, H. 2003. *Practical reasoning about final ends*. Cambridge: Cambridge University Press.

Sartor, G. 2010. Doing justice to rights and values: teleological reasoning and proportionality. *Artificial Intelligence and Law* 18: 175–215.

Savage, L.J. 1954. *The foundations of statistics*. New York, NY: Dover Publications.

Schmitz, D. 1996. *Rational choice and moral agency*. Princeton, NJ: Princeton University Press.

Shoup, D. 2005. *The high cost of free parking*. Chicago, Ill.: APA Planners Press.

Woodard, C. 2008.*Reasons, patterns, and cooperation*. NY: Routledge.

Interactive Decision-Making and Morality

Wojciech Załuski

1 Introduction

Interactive decision-making occurs when three conditions are met: There are at least two decision-makers; the effects of each agent's decision are co-determined by the decisions of other agents; what each agent does depends on her expectations as to what the other agents will do, and while forming these expectations, she knows that the other agents will form similar expectations regarding her own decision. This type of decision-making—also termed "strategic"—is studied in game theory, a branch of rational choice theory (which in addition to game theory embraces decision theory, concerned with nonstrategic choices, and social choice theory, concerned with the problems of group decision-making). In this essay, we shall try to distinguish and analyze different ways in which game theory—a mathematical theory of interactive decision-making—can contribute to moral philosophy. In our view, one can distinguish eight main ways in which game theory can be gainfully appealed to by a moral philosopher, that is, game theory can be viewed as a tool for better understanding a function of morality; determining the content of moral norms; criticizing certain moral conceptions; analyzing the problem of the validity of moral norms; analyzing the possibility of deriving morality from instrumental rationality; analyzing moral decision-making; analyzing the nature of moral dispositions; analyzing the functions of moral emotions; and analyzing the cultural evolution of moral norms. We shall present these contributions in Sects. 3 through 11. In the following section, we shall present some basic information about game theory.

W. Załuski (✉)
Department of Philosophy of Law and Legal Ethics, Jagiellonian University, Krakow, Poland
e-mail: zaluskiwojciech@gmail.com

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_15

2 Basic Information About Game Theory[1]

2.1 Tasks and Branches of Game Theory

The twofold task of game theory is to provide theoretical models of strategic interactions and to provide criteria of rational choice in them. Game theory is divided into the classical and the nonclassical. Classical game theory embraces noncooperative game theory and cooperative game theory, whereas nonclassical game theory is evolutionary game theory. Noncooperative game theory analyzes noncooperative games, i.e., games in which joint-action agreements between agents are *not* enforceable (binding), whereas cooperative game theory analyzes cooperative games, i.e., games in which joint-action agreements *are* enforceable (binding). A sub-branch of game theory is bargaining theory. Bargaining theory is aimed at solving the bargaining problem (i.e., the problem of distributing the surplus of goods between parties who contributed to bringing it about) by providing unique solutions to it. Bargaining theory can be constructed in two different ways: within cooperative game theory and within noncooperative game theory. One of the most plausible bargaining solutions provided within cooperative game theory (i.e., a solution satisfying a set of plausible axioms) is the Nash arbitration scheme, which prescribes the outcome that maximizes the product of the bargainer's utility increments in relation to their initial bargaining position (more information on this solution will be provided in Sect. 4).

2.2 The Concept of a Game

A game (in a game-theoretic sense) is an interaction between two or more agents which is determined by the rules specifying the list of players, the strategies available to each player, the sequence in which players make their moves, and each player's payoffs for all possible combinations of strategies pursued by the players. The above definition uses two concepts which need further clarification: strategies and payoffs. A strategy is a *complete plan of action*, i.e., a plan which determines what the agent is supposed to do at each possible stage of the game. A player's payoffs capture the values, i.e., utility, the player assigns to the various outcomes of the game. They may reflect various—and not only selfish—motivations of players. For example, if a player cares about satisfying her opponent's interests as much as she does about satisfying her own interests, then this "utilitarian" motivation will be reflected in her utility function (which is a technical tool for presenting a player's preferences over various outcomes of the game) and thereby in her payoffs.

[1] Some Portions of this section are taken from my book Załuski (2013), 19–73.

2.3 The Assumptions of Classical Game Theory

Classical game theory makes two basic assumptions regarding players: instrumental rationality and common knowledge of rationality. The former asserts that the players (i) have a set of rational preferences (i.e., preferences satisfying a number of axioms of rationality) over all possible pairs of lotteries defined over the outcomes of their choices and (ii) act as if they maximized the satisfaction of these preferences (i.e., as if they maximized their expected utility function), given their knowledge that the other players act likewise. In strategic contexts, the assumption that players maximize their expected utility function boils down to the assumption that they choose their strategies recommended by the solution concepts of game theory. Game theory also assumes that the players have common knowledge of the following proposition P: "All players are instrumentally rational"; thus, game-theoretical analyses make the assumption of common knowledge of (instrumental rationality). To clarify this notion, it is necessary to mention the distinction between individual knowledge, mutual knowledge, and common knowledge. Common knowledge implies mutual knowledge, and mutual knowledge implies individual knowledge. A proposition P is an object of *individual knowledge* in a group of agents if at least one agent knows P, but not every agent knows P. A proposition P is an object of *mutual knowledge* in a group of agents if each agent knows P, but at least one agent does not know that every other agent knows P. A proposition P is an object of *common knowledge* in a group of agents if every agent knows P and each agent knows that every agent knows P, and each agent knows that every agent knows that each agent knows P, etc. Mutual knowledge of P among a group of agents can become common knowledge of P among this group of agents if, e.g., P is announced publicly to them in such a way that each agent knows that the other agents must have heard that P.

In analyzing a game, one can make strong or weak assumptions regarding the players' knowledge (apart from the assumption of common knowledge of rationality, which is always made in classical game-theoretic analyses). The strong assumption says that players have complete knowledge of the rules of the game and thereby know the structure of the game; i.e., they know the list of players, the strategies available to each player, and each player's playoffs for all possible combinations of strategies pursued by the players. Games in which players have this kind of knowledge are called *games of complete information*. However, it is clear that in many real-life situations, the above assumption is not fulfilled, as one or more players may have incomplete information; i.e., they may fail to know one or more elements of the structure of the game. For this reason, game theory has had to face the challenge of analyzing games of *incomplete information*. Three more remarks seem in order here. First, the only games with incomplete information which have turned out to be amenable to formal analysis were games in which players may fail to know the other players' preferences (utility functions). Second, in games with incomplete information, agents still have complete knowledge of their own preferences; i.e., in such games, the assumption of instrumental rationality still holds. What agents may fail to know in such games are the preferences of other players. Third, two situations should be carefully distinguished: one in which a player is not certain as to what game she is actually playing, because she does not know her opponent's preferences, and one in which a player is not playing the game she thinks she is playing, because she has mistaken beliefs about the other players' preferences. The former situation, but not the latter, is a game with incomplete information. Game theory studies games with complete and incomplete information, but not games with mistaken information. Another relevant distinction is between *perfect information* and *imperfect information*. If each player at every stage of the game knows the game's entire previous history, the game is called a *game with perfect information*. If a player at any stage in the game does not know the game's entire previous history, then game is called a *game with imperfect information*.

2.4 Solution Concepts of Noncooperative Game Theory

Three basic solution concepts of game theory shall be presented: dominance, Nash equilibrium, and backward induction.

A strategy a_i of a given player *strongly dominates* her strategy a_j iff, for this player, a_i is a better response than a_j to each strategy of her opponent. A strategy a_i of a given player *weakly dominates* her strategy a_j iff, for this player, a_i does not strongly dominate a_j; a_i is a better response than a_j to at least one strategy of her opponent and is not a worse response than a_j to any of the strategies of her opponent. Now, *one of the most obvious criteria of rationality assumed in game theory is the ban on the choice of strongly dominated strategies.* Another criterion of rationality assumed in game theory, also very intuitive though less obvious than the previous one, is the ban on the choice of weakly dominated strategies. Let us illustrate the concept of dominance by means of a simple nonzero-sum game (unlike in zero-sum games, also called games of pure rivalry, in nonzero-sum games, players can benefit simultaneously; i.e., the gain of one player does not have to mean the loss of the other) (Fig. 1):

P1 denotes Player 1; *P2*, Player 2. Player 1's strategies are listed in the rows; Player 2's, in the columns. The payoffs in each cell of the matrix are determined by the choice of the players' respective strategies. The first number in each cell denotes *P1*'s payoff; the second number, *P2*'s payoff. The game's rational results (i.e., those determined by the solution concepts of game theory) are in parentheses. Since in this game a_2 is strongly dominated by a_1, we can expect that Player 1 *will* choose a_1.

Fig. 1 A game with a strongly dominant strategy

$P1/P2$	b_1	b_2
a_1	(4, 3)	1, 2
a_2	2, 2	0, 3

By saying "Player 1 will choose a_1," we assume a descriptive interpretation of game theory. Assuming a normative interpretation of game theory, we would say, "Player 1 *should* choose a_1." Clearly, we can assume a mixed interpretation of game theory, thereby giving to the word *will* both a descriptive and a normative sense. Player 2 has neither a strongly nor a weakly dominant strategy. However, the assumption of common knowledge of rationality enables us to say that since Player 2 knows that Player 1 is rational, and therefore Player 1 will choose a_1, Player 2 will choose b_1, which is her best response to a_1.

A *Nash equilibrium* is a steady state of a game, i.e., a state (outcome) such that no player has an incentive to unilaterally deviate from it. The stability of a Nash equilibrium results from the fact that it is generated by a combination of strategies which are best responses to one another; i.e., no player can improve her situation by switching to a different strategy. In other words, in a Nash equilibrium, each player's strategy is the utility-maximizing response to the other players' strategies. More technically, assume that Player 1 has the strategy set A: $\{a_1, a_2, \ldots, a_n\}$ and Player 2 the strategy set B: $\{b_1, b_2, \ldots, b_n\}$; assume also that a_{-i} will denote all strategies other than a_i and b_{-i} all strategies other than b_i. Now, a combination of strategies $\{a_i, b_i\}$ brings about a Nash equilibrium iff a_i is at least as good a response to b_i as a_{-i} and b_i is at least as good a response to a_i as b_{-i}. Now, the important theorem of game theory (the Nash theorem) says that if mixed strategies (probability distributions over pure strategies) are allowed, then there exists at least one n-tuple of strategies in the Nash equilibrium for every game. There are three main problems with the concept of Nash equilibrium—the fundamental solution concept of noncooperative game theory. First, there are many games in which there are no Nash equilibria in pure strategies. This problem, however, is resolved in the sense that each game must have a Nash equilibrium: if not in pure strategies, then in mixed ones. Second, many games have multiple Nash equilibria. Third, there are games in which Nash equilibria are not plausible as solutions of games, since they rely, for example, on incredible threats. In order to resolve the last two problems, game theorists impose additional requirements of rationality on the concept of Nash equilibrium. The solution concepts joining Nash equilibrium with some additional requirements of rationality are called "refinements of Nash equilibria." Thus, refinements eliminate one or more Nash equilibria from the set of all Nash equilibria. Examples of such refinements are rollback equilibrium, subgame-perfect equilibrium, sequential equilibrium, perfect equilibrium, proper equilibrium, and forward induction. There is no agreement among game theorists as to which ones are most plausible (cf. Bicchieri 2004). We shall present only one of them: the uncontroversial rollback equilibrium.

Rollback equilibrium is a Nash equilibrium generated by the reasoning referred to as backward induction. Rollback equilibrium is a refinement of Nash equilibrium: Each rollback equilibrium is a Nash equilibrium, but not each Nash equilibrium is a rollback equilibrium. Backward induction reasoning can be applied in extensive-form games with perfect information, i.e., to put it technically, games in which all information sets consist of single nodes. In this kind of reasoning, players assess what would happen at the last node if various game histories were realized and then use this knowledge to establish what would happen in the preceding nodes and thereby

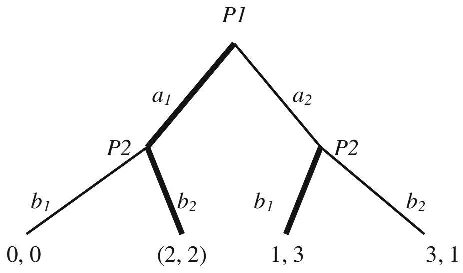


Fig. 2 Backward induction and rollback equilibrium

to choose their strategy. This procedure yields the unique outcome of a game—the rollback equilibrium. These concepts are illustrated in the following example (Fig. 2):

The first number in the final nodes refers to Player 1's payoff; the second one, to Player 2's payoff. The bold lines indicate the actions which will be chosen by players when they are to make their choices, whereas the bold numbers indicate the game's outcome.

2.5 Evolutionary Game Theory

Evolutionary game theory is a branch of game theory which seeks to define the concept of "evolutionarily efficient" strategies, i.e., those strategies that have proven to be successful in an evolutionary process in which many strategies compete with one another. Unlike classical game theory, evolutionary game theory dispenses with the assumption of the players' instrumental rationality; when used to model biological evolution, it assumes that agents play against each other without any prior knowledge of which strategy is fitness-enhancing, i.e., which one increases the representation of their offspring in successive generations. But evolutionary game theory can be used to model not only biological evolution but also cultural evolution. In the former case, payoffs are measured in terms of reproductive success; in the latter, in terms of some other value suitable for a given social context. When modeling cultural evolution, a different feedback mechanism than differentiated biological reproduction is also postulated: It is assumed that relatively less successful agents will tend to imitate the strategies of the more successful ones, and the relatively more successful players will tend to self-imitate (this is in fact the only rationality requirement imposed on players by evolutionary game theory in the context of cultural evolution). One more remark about evolutionary game theory may be in order here. As was mentioned before, one of the problems of classical game theory is the problem of selecting a Nash equilibrium in situations in which there are many Nash equilibria. Now, evolutionary game theory can be helpful in this regard, as one of its solution concepts (shortly to be presented in this section)—the evolutionary stable strategy—is in fact a refinement of a Nash equilibrium. It can *ipso facto* also be appealed to in order to justify the very concept of Nash equilibrium as a requirement of rationality: Because the concept of

Nash equilibrium is part of the definition of an evolutionary stable strategy, which in turn has proved to be a good model of biological adaptation, the concept of Nash equilibrium as a criterion of rationality can be said to be justified by appealing to "evolutionary rationality" (the strategies "picked out" by natural selection are bound to be equilibrium strategies).

There are three main theoretical approaches within evolutionary game theory: (1) the search for evolutionary stable strategies; (2) the replicator dynamics approach; and (3) agent-based modeling.

Ref. (1) (Axelrod 1984). An evolutionary stable strategy is a strategy that cannot be forced out of the population by any other strategy; it therefore cannot be invaded by any mutant strategy. As mentioned, the concept of an evolutionary stable strategy is a refinement of the concept of Nash equilibrium: All evolutionary stable strategies are Nash equilibria, but not all Nash equilibria are evolutionary stable strategies. For a strategy to be evolutionary stable, two conditions must hold (cf. Maynard Smith 1982, 14; Weibull 1995 , 32–46) :

(a) *The Nash equilibrium condition: Expected utility {I, I} ≥ Expected utility {J, I}, for all possible strategies J.* Expected utility $\{I, I\}$ is the expected utility derived by a player in a single round in which she plays strategy I, and her opponent also plays strategy I. *Expected utility* $\{J, I\}$ is the expected utility derived by a player in a single round in which she plays strategy J, whereas her opponent plays strategy I. Therefore, according to this condition, strategy I must be at least as good a response to itself as any other strategy J. It should be noted that this condition does not exclude a situation in which I is not the only best response to I. Therefore, when it holds as an equality, it admits of a situation in which a population playing I being invaded by an individual playing J is as good a response to I as I. Such an invasion is not possible if strategy I satisfies an additional condition—the stability condition.

(b) *The stability condition: Expected utility {I, I} > Expected utility {J, I} or Expected utility {I, J} > Expected utility {J, J}, for all possible strategies J, where $J \neq I$.* Therefore, if I is not the only best reply to itself, it must be a better reply to J than J to itself.

An important element of the evolutionary stable strategy approach is the *mutation mechanism* that generates random mutations, which add "noise" to the population system. This mechanism is absent in the second solution concept of evolutionary game theory, namely, replicator dynamics.

Ref. (2) (Axelrod and Hamilton 1981). The replicator dynamics approach consists in providing equations which describe the way a given population changes over time (Taylor and Jonker 1978). This approach is therefore dynamic, while the approach based on the concept of an evolutionary stable strategy is static (it does not explain how evolutionary stable strategies are reached). The replicator dynamics approach assumes that the propagation rate of each strategy in a given population is proportional to (a) the fraction of the population currently playing this strategy and (b) the difference between that strategy's mean payoff and the mean payoff of the population as a whole (evolution takes account of a strategy's fitness relative to the fitness of the population as a whole). Two concepts are important for the replicator dynamics approach: a stable steady state and a basin of attraction. A *stable steady state* of a dynamic system is a state *s* having the following two features: Once the system enters *s*, it never leaves it; and once the system approaches "close enough" to *s*, it remains near *s*. The *basin of attraction* of *s* is the set of initial states (population proportions) such that if the dynamic system begins in one of those states, then it will eventually converge toward *s*. In many cases, the best way to understand a dynamic system is to construct a "phase portrait" diagramming its steady states and their basins of attraction. The "state" of a dynamic system can be interpreted as the proportion of players choosing a strategy. Some additional remarks on evolutionary stable strategies and replicator dynamics may be helpful. First, all evolutionary stable strategies are resting points—stable steady states—of the replicator dynamics, but not all resting points of the replicator dynamics are evolutionary stable strategies. Second, as was already mentioned, replicator dynamics are a deterministic process; i.e., they do not take mutations into account, while the approach based on the concept of an evolutionary stable strategy does take mutations into account. Third, an evolutionary stable strategy or a steady state may be polymorphic. The notion of polymorphism is close to the notion of a mixed strategy, but there is an important difference between them: While a mixed strategy Nash equilibrium has to be interpreted as a situation in which individual players follow a mixed strategy, a polymorphism may be interpreted in two different ways: as a situation in which individual players follow a mixed strategy or as a situation in which the respective proportions of the population follow different pure strategies.

Ref. (3) (Bicchieri 2004). Replicator dynamics are an example of the aggregative model, which is based on the assumption of random interactions between members of a population. It therefore represents only the state of the population, thereby discounting the potential differences between various individuals playing the same strategies. The differences may consist in the following: that interactions between agents are not random (their frequency may be inversely proportional to the spatial distance between them); that various agents assume different learning rules; and that the population contains "key agents," whose change of strategy triggers of a large-scale shift in the proportion of strategies in the population. In short: Models based on replicator dynamics do not describe how the behavior of individual players changes over time; they describe aggregate trends in the change of the population as a whole. This feature of the replicator dynamics approach will not be a drawback if the aforementioned differences are inessential in the population under analysis. However, if they prove to be of vital importance (as they arguably are in many real-life interactions), then a different analytical approach—taking these differences into account—is necessary. An approach that allows for these differences is called agent-based modeling. Many different types of models have been proposed on this approach, viz. social network (local interaction) models, lattice models, small-world models, networks of bounded degree, dynamic models (cf. esp. McKenzie Alexander 2007, 33–53). What is common to replicator dynamics and agent-based modeling is that both model interactions between boundedly rational agents and can be interpreted in two different ways, viz. as models of biological evolution or as models of cultural evolution.

Having presented the basics of game theory, let us now move to the central part of this essay, viz. to a discussion of various ways in which game theory can be useful to a moral philosopher.

3 Game Theory as a Tool for Better Understanding a Function of Morality

It has been argued by many philosophers (e.g., Hardin 1965; Olson 1965; Ullmann-Margalit 1977; Mackie 1977; Taylor 1987; Bicchieri 2006) that morality is above all a means for fostering cooperation among agents, i.e., in economic jargon, for solving collective action problems. This claim can also be reformulated in the spirit of social contract theory: Morality is a means for transferring a population from a state of nature (a state of noncooperation) to a state of civilization (a state of cooperation). To avoid misunderstanding the claim that morality is a means for fostering cooperation, let us introduce some important distinctions. First, by morality-fostered cooperation, we shall mean joint actions that are beneficial to its participants and not detrimental to nonparticipants, i.e., in economic jargon, joint actions which do not bring about negative externalities. Joint actions which generate negative externalities (e.g., cooperation among criminal groups or among members of a cartel) should be counteracted by morality. Second, by joint actions, we shall also understand joint omissions. Accordingly, a joint action to be supported by morality will, for instance, be avoiding the use of violence in mutual relationships or refraining from stealing. By saying that the function of morality is to solve collective action problems, we therefore mean that the morality is to promote *positive* cooperation (undertaking joint actions) and *negative* cooperation (not undertaking actions which hinder other agents pursuing goals that do not harm other agents). There is an obvious difference between these two types of cooperation: The goal of the former is to enable agents to realize ends which they could not achieve by individual action, whereas the goal of the latter is to remove obstacles for realizing those ends by agents which they are capable of realizing individually. The question arises as to what role game theory can play in analyzing morality as a means of solving collective action problems. The role is twofold: Game theory enables a precise description of collective action problems that would arise in the absence of morality and makes it possible to show how morality can help solve these problems. In general, the claim that a function of morality is to solve collective action problems can be reformulated in game-theoretic parlance in the following way: A function of morality is to transform games with a noncooperative result into games with a cooperative result. Let us now analyze the concept of a collective action problem in greater detail. A collective action problem arises when people would benefit from acting together (acting referring to both commissions and omissions), but such acting encounters obstacles.[2] The benefit of undertaking a collective action may consist either in preventing some existing good from diminishing (e.g., in preventing overuse of a given resource: a body of water, a tract of land, a forest, etc.) or in creating some new good that individuals would not be able to create if they acted individually (e.g., in creating a new road joining two villages or in creating a "welfare surplus" through the very act of exchanging goods). It bears emphasizing that the good that may be created or prevented from deteriorating need not be material: It may also be an abstract good, such as security or freedom. Depending on the kinds of obstacles that hinder taking a collective action, one can distinguish two "pure types" of collective-action problems. The first type embraces those collective-action problems in which the main obstacle to taking a collective action lies in the fact that each agent is tempted to pursue her own egoistic interests, i.e., to pursue her individual gain at the expense of the gain of others. In this type of collective-action problem, each agent wants the other players to act cooperatively but is herself tempted to choose noncooperation. The most famous model for this type of collective-action problem is the Prisoner's Dilemma. In this game, players have a choice between two strategies: cooperating (C) and defecting (D). A game is a Prisoner's Dilemma if each player has the following preference ordering over the possible results of their choices (the ordering is presented from the viewpoint of Player 1): $T > R >$ P > S, where T is the Temptation result (arising if Player 1 defects and Player 2 cooperates), R is the Reward result (arising if both players cooperate), P is the Punishment result (arising if both players defect), and S is the Sucker's result (arising if Player 1 cooperates and Player 2 defects). Assume, e.g., that $T = 5$, $R = 3$, $P = 1$, and $S = 0$. The matrix will therefore be as follows (Fig. 3):

The problem is that the only steady state of the Prisoner's Dilemma, i.e., its Nash equilibrium, is $\{D; D\}$: No player is tempted to switch to strategy C, given strategy D of the other player. This result, however, is not Pareto-optimal, since each player's payoff would be increased if each of them chose C (the outcome of a game is *Pareto-optimal* if it cannot be improved upon, i.e., if no player can be made better off without making someone else worse off). Let us give an example of an interaction having the structure of the Prisoner's Dilemma. Tom and John are hot-tempered men who detest each other. A tradesman visits their small town and offers each of them a gun. Thus, they face the following choice: to buy (D) or not to buy (C) a gun. For each player, the best situation is that he plays D while his opponent plays C (the player

Fig. 3 Prisoner's dilemma

$P1/P2$	C	D
C	3, 3	0, 5
D	5, 0	(1, 1)

[2] We provide a very broad, rather nonstandard definition of collective-action problems. For instance, Hardin (2007, 59) narrows the notion of a collective-action problem to multiple-person interactions (involving more than two persons) in which the agents have a temptation to act self-interestedly.

buying a gun gains advantage over the other player); each prefers the situation "both play *C*" to the situation "both play *D*" (if both play *D*, there will arise the danger of a life-and-death gunfight); and each most fears the situation in which he plays *C* while his opponent plays *D*. The game therefore fits the Prisoner's Dilemma. The Prisoner's Dilemma serves also as a model for multiple other situations, especially those where there is some scarce good and agents have to restrain their consumption of it if they want to avoid the worst result, namely $\{D; D\}$—a fast depletion of the good. The collective-action problem would be solved if agents internalized a norm prescribing the choice of cooperative strategy (e.g., the norm "Thou shalt never be greedy"). The second type embraces those collective-action problems in which the main obstacle for taking a collective action is the fact that the agents involved encounter difficulties coordinating their actions (Schelling 1960; Lewis 1969; Postema 1998; Marmor 2009). Coordination problems are a manifestation of an indeterminacy of the results of a situation; i.e., there may be many possible ways of coordinating actions in a situation, such that the result of an interaction is indeterminate. In the first type of collective-action problem (modeled, e.g., by the Prisoner's Dilemma), there is a sharp conflict of interests among agents; in the second type (modeled by coordination games), there is no such conflict. It should be stressed, however, that coordination games are not a homogenous category. They embrace at least two different types of games, viz. pure coordination games and trust games. In pure coordination games, there are many Nash equilibria, and none of them is better for any of the players than the other Nash equilibria (all Nash equilibria are therefore cooperative). To give an example: Two persons are driving cars on the same road in opposite directions in a country in which there are no driving rules, so they face the dilemma of whether to drive on the right side of the road or on the left side. There are two Nash equilibria in this game, viz. one in which both players choose the right-hand driving rule and one in which both choose the left-hand driving rule. In trust games, there are many Nash equilibria: One of them is better for both players than the other Nash equilibria (not all Nash equilibria are therefore cooperative), and each player has a safe strategy to choose if she is not sure if coordination on the Pareto-optimal equilibrium will be successful. A classic example of this kind of coordination problem is the Stag Hunt game (based on a situation that Jean Jacques Rousseau described in his *Discours sur l'origine et les fondements de l'inégalité parmi les hommes*). In this game, each player must choose between the cooperation strategy "hunt a stag" (*C*), i.e., "cooperate," and the noncooperation strategy "hunt a hare" (*D*), i.e., "not cooperate." It is assumed that hunting a stag will be successful only if all the players take part in it; consequently, a player who hunts a stag, while the other player hunts a hare, gets nothing; that every player can hunt a hare alone (which guarantees a moderate return); that a share in a stag is valued more highly than a hare; and that the players choose their strategies simultaneously and cannot communicate with each other. Thus, each player has the following preference ordering $R > T = P > S$ (the names of the results, taken from the Prisoner's Dilemma, are not fully adequate here, but we will retain them for ease of exposition). Assume, e.g., that $R = 5$, $T = 4$, $P = 4$, and $S = 0$ (the numbers 5, 4, and 0 have been chosen arbitrarily; they can be replaced by any other set of numbers

Fig. 4 Stag Hunt

$P1/P2$	C	D
C	$(5, 5)$	$0, 4$
D	$4, 0$	$(4, 4)$

that preserves this ordering). The Stag Hunt game is set out in the following matrix (Fig. 4):

There are also some "mixed types" of collective-action problems, which in some form present both of the obstacles characteristic of the "pure types," viz. the temptation to act egoistically and the difficulty of coordinating actions. To sum up, collective action problems may result from the temptation to act egoistically and/or from the difficulty in coordinating the agents' actions.[3] Accordingly, one can distinguish three possible types of norms: those which solve collective-action problems stemming from egoism, those which solve collective action problems stemming from coordination difficulties, and those which solve collective-action problems stemming both from egoism and coordination difficulties. These norms may be moral (with one exception to be specified below) but also legal, for it is clear that law also fulfills the function of solving collective-action problems. Interestingly, one may argue that the specific character of law is that *norms solving pure coordination problems can be only legal, not moral*. This claim is based on the plausible assumption that morality deals with counteracting temptation to act egoistically, and in pure coordination problems, there is no such temptation (in trust games such a temptation does appear, though in a milder form than in the Prisoner's Dilemma). Thus, in addition to sharpening the claim that an important function of morality is to solve collective-action problems, game theory may offer some help in drawing a distinction between moral rules and legal rules. At the end of this section, it will be emphasized that game theory provides precious insights into *a* function of morality rather than *the* function of morality. For it is clear that the function of morality cannot be reduced to fostering cooperation, i.e., creating harmonious relationships between human beings. Morality has at least three parts (functions), so described by Clive Staples Lewis: "Morality […] seems to be concerned with three things. Firstly, with fair play and harmony between individuals. Secondly, with what might be called tidying up or harmonising the things inside each individual. Thirdly, with the general purpose of human life as a whole: what man was made for […]" (Lewis 1980, 72). Clearly, game theory in helpful in analyzing only the first of these three functions.

[3] A different classification of strategic situations, with a corresponding classification of norms, was proposed by E. Ullmann-Margalit (1977). Margalit assumed that every strategic situation can be classified as a combination of three paradigmatic cases: (a) the Prisoner's Dilemma; (b) coordination problems; and (c) inequality/partiality situations.

4 Game Theory as a Tool for Determining the Content of Moral Norms

Bargaining theory—one of the branches of game theory—has been used to clarify the content of the rules of justice. This way of applying game theory to moral philosophy was first developed in detail by R. B. Braithwaite. (1955). Bargaining theory solves so-called bargaining problems. A bargaining problem arises when two conditions are met: (1) The outcome of agents' noncooperative actions is not Pareto-optimal, such that a rational and voluntary agreement may yield additional benefits to each of them; (2) the set of Pareto-optimal improvements on the Status Quo point (*SQ*), i.e., the noncooperative result, contains at least two elements with respect to which the agents have opposing preferences; therefore, bargaining problems would not emerge if agents could distribute benefits produced by their mutual cooperation in only one way. Thus, in bargaining situations, the interests of agents in part *converge* (because they each want to reach a better outcome than their initial position) and in part *diverge* (because they have opposing preferences regarding which of at least two different Pareto-optimal outcomes should be reached). Given that the bargaining problem concerns the choice of distributing the surplus of goods generated by mutually advantageous cooperation from the set of available distributions, it can be regarded as a variant of the problem of distributive justice. It is clear a rational outcome of a bargain must meet the following two requirements. (1) It must be *individually rational*—it must afford the bargainers no less than they would receive from the outcome adopted as a starting point for bargaining, i.e., from *SQ*. In other words, for each player, the outcome must be a Pareto-improvement on her payoff in *SQ*. (2) It must be *collectively rational,* i.e., Pareto-optimal, and thereby situated on the upper bound of the utility space. The outcomes that meet these requirements form the *negotiation set* for a bargaining problem. The problem is that there is usually a large number of outcomes meeting these conditions. Accordingly, game theorists, such as von Neumann and Morgenstern, have claimed that the choice of a single point from among the points that satisfy these conditions is a matter of psychology and so cannot be assessed in terms of rationality. This belief was to prove to be unduly pessimistic, since several solutions were subsequently put forward that specify a unique outcome for each bargaining problem. Two such solutions shall be presented: the Nash solution and the Kalai–Smorodinsky solution. Let S be a two-dimensional utility space; i.e., $S \subset IR^n$ will be a compact and convex set of feasible utility vectors $s^i = (s_1, s_2)$, $i = 1, 2, \ldots, n$, and $N = (1, 2)$ be the set of bargainers. Let $d \in S$ denote the *Status Quo*, defined in terms of utility. Now, (S, d) is a bargaining situation if there are at least two different $s \in S$ such that $d < s$. A bargaining solution is a function f that assigns to each bargaining problem (S, d) a feasible utility vector, which is that problem's unique solution point. Thus, f is a mapping from the sets of the form (S, d) onto IR^n, with the qualification that $f(S, d) \in S$. Now, the Nash solution prescribes the choice of the vector which, in the feasible set, maximizes the product of individual utility gains from *SQ* (this product is called the "Nash product"). Formally: The Nash solution selects such $s^i = (s_1, s_2)$ that maximize the expression $[(s_1 - d_1)\ (s_2 - d_2)]$; more succinctly: The Nash solution is the function $f\ (S, d) = max\ [(s_1 - d_1)\ (s_2 - d_2)]$. The Kalai–Smorodinsky solution requires introducing an additional notion—that of an *ideal point*. This point is determined by the *highest utility payoffs* each player can obtain in the game. The Kalai–Smorodinsky bargaining solution selects $s^i = (s_1, s_2)$ as the maximal point in S, such that $[(s_1 - d_1)/(s_1{}' - d_1)] = [(s_2 - d_2)/(s_2{}' - d_2)]$. This solution is to be interpreted as follows: $(s_i{}'—d_i)$ is the maximal possible utility gain that a player i can obtain in relation to SQ; the ratio $(s_i - d_i)/(s_i{}' - d_i)$ is the "degree of success" this player had in trying to obtain the maximal possible utility gain. Accordingly, the Kalai–Smorodinsky solution selects the point x—utility vector—in which the ratio expressing the "degree of success" is equal for both players and maximal with respect to set S. Thus, the solution prescribes the equalization of the parties' sacrifice relative to the maximal gain they could expect in the available set of options.[4] These two (and other) solutions can be interpreted as rules of distributive justice. It should be stressed that bargaining theory is not the only branch of game theory that provides rules that can be interpreted as rules of distributive justice. In cooperative game theory, other solution concepts have also been proposed (e.g., the Shapley value) that can be so interpreted.[5]

5 Game Theory as a Tool for Criticizing Certain Moral Conceptions

The results of game theory can be invoked to critically analyze the ethical conception which can be referred to as "instrumental ethical egoism," defended, e.g., by Bernard Mandeville (in *The Fable of the Bees*) and Adam Smith (in *The Wealth of Nations*). The conception relies on the belief (called "Mandeville's law") that "private vices beget public virtues," i.e., on the belief in the existence of a causal connection between pursuing one's own interests and generating social utility. Accordingly, it prescribes egoism as a means for realizing social welfare. According to this conception, egoism is not good *in itself* (the view that egoism is good in itself can be termed "noninstrumental ethical egoism"), but is *instrumentally* good, i.e., good in that it contributes to the realization of an end—social welfare—that is intrinsically good. This view received some support from economics, which teaches that under conditions of perfect competition (i.e., conditions in which all agents in the market behave as price takers, all producers sell commodities that are identical with respect to all essential characteristics, producers and consumers possess perfect information regarding each commodity, and there is free entry into and exit from the market), the

[4] A similar solution was independently proposed by Gauthier (1986); Gauthier called his solution the "principle of minimizing the maximum relative concession."

[5] The Shapley solution asserts that assuming that each order in which the players join the grand coalition is equally probable, each player should receive her average contribution to this coalition. Excellent overviews of this and other solution concepts which can be interpreted as rules of distributive justice can be found in Brams (1990) and Peyton Young (1995).

actions of rational egoists generate a Pareto-optimal outcome. Thus, it turns out that under conditions of perfect competition the pursuit of private interests can indeed be claimed to lead to the maximization of social utility. The problem is that these conditions are seldom fulfilled in real life. A large part of human interactions has the form of the Prisoner's Dilemma. And in this type of condition, individual and social interests are not naturally aligned: Agents who *pursue* their own interests in strategic interactions find themselves in a worse situation than if they acted *against* their own interests. Thus, one can argue that game theory has provided a devastating critique of instrumental ethical egoism as a general ethical theory; instrumental ethical egoism can at best be defended as a "local theory," applicable only in very special and seldom realized types of circumstances.

6 Game Theory as a Tool for Analyzing the Problem of the Validity of Moral Norms

Game-theoretic, or strategic, thinking could arguably shed light on the problem of the validity of moral norms, or, more precisely, it could arguably help in developing one of two accounts of the validity of moral norms. For, as it seems, one can distinguish two basic accounts of the validity of moral norms: the nonstrategic and the strategic.[6] The nonstrategic account assumes that moral norms are valid (binding) irrespective of whether all (or almost all) agents follow them. Thus, this account does not regard the general (or almost general) compliance with moral norms as a necessary condition of their validity. The strategic account assumes that moral norms are valid only if they are generally (or almost generally) followed. Thus, according to this account, an agent is not bound to follow a moral norm if other agents with whom she interacts do not follow this norm. General (or almost general) compliance with a moral norm is therefore a necessary condition of its validity. The strategic theory was assumed, e.g., by Thomas Hobbes, who famously wrote:

> But because most men, by reason of their perverse desire of present profit, are very unapt to observe these Lawes, although acknowledg'd by them, if perhaps some others more humble than the rest should exercise that equity and usefulnesse which Reason dictates, those not practising the same, surely they would not follow Reason in so doing; nor would they hereby procure themselves peace, but a more certain quick destruction, and the keepers of the Law become a meer prey to the breakers of it. It is not therefore to be imagin'd, that by Nature, (that is, by Reason) men are oblig'd to the exercise of all these Lawes in that state of men wherein they are not practis'd by others. We are oblig'd yet in the interim to a readinesse of mind to observe them whensoever their observation shall seeme to conduce to the end for which they were ordain'd. (Hobbes 1998, chap. III, sect. XXVII)

[6] We have not found this distinction (which seems useful and important to us) in the relevant philosophical or game-theoretic literature. It is, however, contained *in nuce* in a small fragment of the book by the Polish philosopher Ossowska (1970, 235), in which she comments on a "disquieting" fragment of Hobbes's *De Cive* (to be cited below) which suggests that we are not obliged to behave morally toward those who do not behave morally toward us.

It should be stressed that the strategic account does not assume that the general (or almost general) following of moral norms is a sufficient condition of their validity: It says that it is a *necessary* condition. Thus, in order to be a complete account of the validity of moral norms, it should be enriched by indicating what conditions are necessary *and sufficient*. This enrichment may take various forms depending on one's views of the sources of the normativity of moral norms. It should also be noted that the strategic account of the validity of moral rules can be understood in a slightly different way than presented above: not as stating the necessary condition of the general validity (i.e., validity for all agents) of a given norm but as stating the necessary condition of the individual validity, i.e., validity for a given agent who finds herself in a situation to which the norm applies. According to this second variant of the strategic theory, an agent A is obliged to follow a moral norm, say, forbidding stealing, in her relation with an agent B only if the agent B follows this norm in relation to the agent A. The strategic account (in either of its two variants) may seem rather counterintuitive because it implies that moral norms have the structure of the conditional sentences "Only if other people do not steal from you (or slander you, and so on) shall you refrain from stealing from them (or slandering them)" (first variant) or "Only if A does not steal from you (or slander you, and so on) shall you refrain from stealing from A (slandering A)" (second variant), while most people do not consider these norms to be conditional on whether others follow them. Nonetheless, as was noticed by Ossowska (1970, 235), at least in one case (particularly stressed by Hobbes), this account appears convincing: We are not obliged to comply with the norm that forbids killing if others want to kill us. This case is explicitly recognized by the law in the form of the principle of self-defense.

7 Game Theory as a Tool for Analyzing the Possibility of Deriving Morality from Instrumental Rationality

One of the central problems of moral philosophy is that of the justification of moral norms. Two main views have been proposed to this problem. The first view defends the autonomy of morality, denying the possibility of its being grounded in some other realm. This view therefore denies the possibility of what is called the "fundamental justification" of morality, a fundamental justification of a realm being one that "does not appeal to any of the concepts of that realm" (Nozick 1974, 6).[7] The second view assumes that such a justification *is* possible. The most frequent version of this view says that such a justification should ground morality in nonmoral instrumental rationality. A fundamental justification may be foundational or nonfoundational: The first makes the assumption that what justifies morality is itself justified, whereas the latter does not make this assumption. The best known attempts at a fundamental justification of morality are those made by Gauthier (1986), McClennen (1990),

[7] According to Danielson (1992, 19) a better name than "fundamental justification" would be "reductive justification."

and Danielson (1992). In the remainder of this section, we shall focus on D. Gauthier's attempt. It should be noted that this attempt is not only fundamental but also foundational: Gauthier tries to show that the requirement of rational action follows from the structure of human agency. In Gauthier's view, the only way to solve conflicts between various desires (the conflicts that we clearly perceive owing to our capacity for semantic representation) is to maximize our expected fulfillment of desire. Accordingly, a maximization principle is the "only one plausible candidate for a principle of coherence" (Gauthier 1988, 1). Thus, Gauthier tries to ground the requirements of rationality in the very structure of human agency. Gauthier describes his main task—that of the fundamental justification of morality—in the following way: "We are committed to showing why an individual, reasoning from nonmoral premises, would accept the constraint of morality on his choices" (Gauthier 1986, 5). The question whether morality can be derived from instrumental rationality can be reformulated in order to make it amenable to a game-theoretic analysis: It can be reformulated as the question of whether it is possible to show that rational agents will choose the strategy of cooperation in the one-shot Prisoner's Dilemma, the game which is assumed to reveal a conflict between the rational and the moral point of view. The problem of the fundamental justification of morality can therefore be construed as the problem of justifying the choice of the strategy of cooperation in the one-shot Prisoner's Dilemma. This can also be called the compliance problem, given that the Prisoner's Dilemma can be used to model the problem of whether to comply with an agreement. As is well known, the cooperation strategy is commonly regarded as an irrational choice in the Prisoner's Dilemma. However, Gauthier argues that this view is wrong and so that rational and egoistic agents (*homines oeconomici*) will act cooperatively in this framing of the Prisoner's Dilemma. In practice, it means that, e.g., it is rational for Player 1 to keep her promise to Player 2 even if Player 1 knows that she will never meet Player 2 and no one will never know that Player 1 did not keep her promise. How does Gauthier argue this claim, that is, how does he construct his fundamental justification of morality?

It should be emphasized at the outset that Gauthier does not intend to defend the claim assumed by most game theorists that because of the value of reputation, it may be rational to cooperate in the indefinitely iterated Prisoner's Dilemma. Gauthier accepts this claim but finds it irrelevant for the problem of deriving morality from rationality because in the iterated games, no self-constraint imposed on an agent's preferences is necessary to achieve cooperation and, according to Gauthier, this sort of constraint is an essential element of moral choice. Gauthier begins his argument by claiming that the utility-maximization criterion should be applied not to an agent's particular strategies (as is assumed in standard rational choice theory) but to an agent's *dispositions* to choose particular strategies ("disposition" is Gauthier's term for rules of action). He writes: "A disposition is rational if and only if an actor holding it can expect his choices to yield no less utility than the choices he would make were he to hold any alternative disposition. We shall consider whether particular choices are rational if and only if they express a rational disposition to choose" (Gauthier 1986, 182–183). He then presents two dispositions to choose from: constrained maximization (CMD) and straightforward maximization (SMD). A constrained maximizer (a CM) is

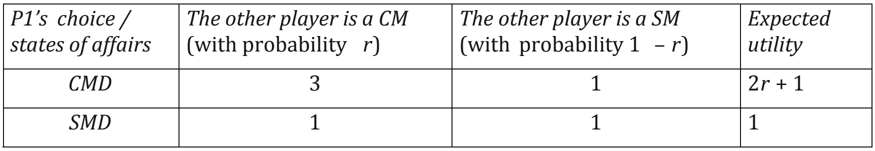
> (i)someone who is conditionally disposed to base her actions on a joint strategy or practice should the utility she expects were everyone so to base his action be no less than what she would expect were everyone to employ individual strategies […] (ii) someone who actually acts on this conditional disposition should her expected utility be greater than what she would expect were everyone to employ individual strategies. (Ibid., 167)

By contrast, a straightforward maximizer (a SM) is someone "who seeks to maximize his utility given the strategies of those with whom he interacts" (ibid., 167). An SM therefore acts in accordance with the requirement of standard rational choice theory. Accordingly, in the Prisoner's Dilemma, she will not choose strictly dominated strategies, i.e., will always play "defect"; a CM, in turn, will play "defect" only if her opponent is an SM; when interacting with another CM, she will play "cooperate." Thus, CMD (constrained maximization disposition), a conditional disposition to comply with the outcomes agreed upon in a bargain, can be interpreted as a moral disposition.[8] Let us not pass to the presentation of Gauthier's argumentation for the rationality of CMD and thereby for the possibility of the fundamental justification of morality. This argumentation rests on the intuition that agents who adopt CMD thereby gain an opportunity to cooperate, an opportunity that is not available to SMs. Consequently, CMs can expect to enjoy benefits that are not available to SMs. We shall now examine the way Gauthier elaborates this basic intuition.

Assume that a rational agent must choose between a disposition to act as an SM or as a CM. This agent's decision-making situation is characterized as parametric, not strategic; i.e., what codetermines the outcomes of the agent's choices is modeled by states of affairs, not by active players striving to predict the agent's choices. Accordingly, when faced with choosing a disposition, the agent should calculate the benefits to be drawn from respective dispositions and choose that disposition which yields at least as much expected utility for her as any other alternative disposition. In what follows, we propose a simplified version of Gauthier's defense of CMD and so of the choice of cooperation in the one-shot Prisoner's Dilemma. In the following matrices, the numbers represent different payoffs in the Prisoner's Dilemma: $R = 3$, and $P = 1$; r denotes the probability that a randomly selected opponent will be a CM. Thus, the situation of a rational agent faced with the problem of choosing a disposition can be set out in the following matrix (Fig. 5):

As we can see, provided the value of r exceeds 0, the rational agent will choose CMD. SMD is less beneficial than CMD because it does not have the opportunities to obtain the rewards of cooperation which present themselves to constrained maximizers. Thus, according to Gauthier, moral constraint can be rational (if conditional on the other player's cooperation). Does this argument provide a fundamental justification of morality? It does not seems to; it can be criticized on various grounds as follows.

[8] As Gauthier writes "a constrained maximizer is prepared in certain circumstances [i.e., playing against another CM—WZ] to base her actions on a joint strategy [determined by a bargaining solution—WZ], without considering whether some individual strategy would yield her greater expected utility" (ibid., 167).



P1's choice / states of affairs	*The other player is a CM* (with probability r)	*The other player is a SM* (with probability $1 - r$)	*Expected utility*
CMD	3	1	$2r + 1$
SMD	1	1	1

Fig. 5 Parametric choice of a disposition on an assumption of transparency

(1) The argument presupposes that CMs have the capacity to recognize the dispositions of other players, which protects them from being exploited by SMs and from failing to take advantage of potentially cooperative encounters with other CMs. In other words, it presupposes the transparency of all players to CMs. Should one, however, accept the assumption of transparency, the defense of the rationality of CMD, even if it were correct, would be deprived of practical import on account of its very narrow application. But Gauthier realized this defect. This is why he examined whether it is rational to choose CMD under more realistic conditions. Such conditions are reflected in the assumption of *translucency*, on which "persons are neither transparent nor opaque, so that their disposition to cooperate or not may be ascertained by others, not with certainty, but more than mere guesswork" (Gauthier 1986, 174) . He therefore admits that a CM may not succeed in recognizing the other players, thus either failing to act cooperatively (if her antagonist happens to be a CM) or failing to avoid being taken advantage of (if her antagonist happens to be an SM). Naturally, the assumption of translucency, as compared with the assumption of transparency, improves the position of SMs and worsens the position of CMs. But Gauthier has successfully shown that even on this assumption his argument works.

(2) Gauthier considers only CMD and SMD, but there are other dispositions to choose from, e.g., universal cooperation and reciprocal cooperation. The disposition of reciprocal cooperation is defined in the following way: "cooperate when and only when cooperation is necessary and sufficient for the other's cooperation" (Danielson 1992, 89) . Now, disturbingly, reciprocal cooperation, though arguably an immoral strategy, seems to fare better than CMD. This shows that even if one assumes Gauthier's idea that rationality should be applied to rules of action rather than to actions themselves, the project of the fundamental of justification seems to fail because the "fundamentally justified" strategy is not a moral one.[9]

(3) Gauthier's claim that CM is a rational disposition implies that a CM should play a strictly dominated "cooperate" strategy when playing with another CM. Such a requirement is very implausible given that the ban on playing strictly dominated strategies seems to be the most intuitive and least controversial prescription of strategic rationality.

[9] For an argument that reciprocal cooperation can be regarded as a moral strategy, that it is not a "moral monster," see Danielson 1992, Chap. 6.

(4) It is not clear why an agent who has chosen CMD, and is now having to make a decision in her interaction with another CM, can at this time be expected to resist the temptation of "switching" to the strategy of "defecting" with a view to maximizing her expected utility. One can imagine two possible justifications for such resistance. The first one says that it is irrational to "switch" to the strategy of defecting with a view to maximizing individual utility because this would be inconsistent with CMD, which is "truly" rational. But this justification is implausible, given the way in which Gauthier tries to prove the rationality of CMD. Gauthier tries to demonstrate that given a sufficiently high probability of mutual recognition between CMs, it is rational, i.e., utility-maximizing, to dispose oneself to be a CM. On the other hand, however, he refuses to admit that for a CM who knows that her opponent is disposed in the same way, it would be rational to defect even if this kind of behavior would maximize her utility. Therefore, considering that the grounds for defection are identical to those which Gauthier appeals to in order to justify the rationality of CMD, it is hard to understand why he questions the rationality of the "false" CM's behavior.[10] In sum, Gauthier does not provide a convincing answer to the question of why a CM should cooperate when she achieved mutual recognition with another CM. The second justification says that once an agent has selected a disposition, i.e., has formed an intention to be an SM or a CM, she cannot act in a way contrary to that disposition: She is causally determined to play "cooperation." Thus, according to this justification, CMD should be interpreted as a deterministic mechanism inducing cooperation that functions as a brake on preference expression. Yet this justification is implausible, for the simple reason that it assumes that a CM in the Prisoner's Dilemma lacks the freedom to choose when she plays with another CM. This consequence is unacceptable, since even if the agent adopted CMD, she must still remain capable of choosing to "defect" if the game she plays is to be treated as a Prisoner's Dilemma. And if she retains this capability, the considerations of rationality will direct her to choose "defect" rather than "cooperate," because, as was argued before, it is implausible to maintain that it is rational to play in accordance with CMD.

In summary, there are serious grounds for doubting that Gauthier has succeeded in providing a fundamental justification of morality.

At the end of this section, it may be instructive to compare Gauthier's account of the relation between morality and rationality with the account proposed by the renowned game theorist J. Harsanyi (1976a, b, 1983). Harsanyi claimed that ethics constitutes a part of the general theory of rational behavior (which in addition includes decision theory and game theory). What distinguishes ethics, in his view, is that it is the theory of rational behavior in the service of the common interest of society as a whole. According to Harsanyi, there are four reasons for treating decision theory, game

[10] It could, of course, be argued that a "true" CM can recognize a "false" CM, which would lead to the noncooperative outcome. This, however, fails to take into account the fact that a "false" CM could initially be a "genuine" CM who, just before making her decision, came to the realization that it is irrational to cooperate.

theory, and ethics as branches of the same, more general theory, viz. the theory of rational behavior: (1) Their method is essentially the same (they start by formulating a primary definition of rational behavior; this definition, which is based on a set of axioms, enables one to formulate a secondary definition of rational behavior, e.g., expected utility maximization in the case of decision theory, various solution concepts in the case of game theory, and, according to Harsanyi, the maximization of the average utility level of all individuals in society in the case of ethics); (2) the axioms of the three theories are mathematically closely related; (3) it has turned out that some basic problems of game theory and ethics can be reduced to decision-making problems (e.g., Bayesian models of games with incomplete information have been constructed); (4) the structure of ethical rules is essentially identical to that of the rules of decision theory and game theory, namely "if you want your action to satisfy axioms $A_1, A_2, \ldots, A_\mathrm{n}$, then do X"; this is therefore the structure of hypothetical imperatives (the character of these imperatives is noncausal, since they specify what one should do in order to satisfy certain criteria). Harsanyi stresses that the passage from the primary definition to the secondary one is a purely technical problem; the real problem—the conceptual (philosophical) one—is how to discover and justify the primary definition of rational behavior. Now, Harsanyi (1983, p. 50) shows that should one introduce into a set of axioms the notion of moral value judgment—the judgment passed by the agent with ethical preferences (i.e., acting impartially by attaching the same weight to the interests of all agents, thereby acting as if she herself had the same probability of being any of these agents), then there exists a unique utility function—a social welfare function—that a rational agent ought to maximize. This function is the utilitarian principle:

$$W_i = 1/n \sum_{j=1}^{n} u_j$$

In the formula, u_j refers the utility an agent j derives from occupying a given social position, $j = 1, 2, \ldots n$ to the set of members of society, and $1/n$ reflects the equi-probability assumption, i.e., the assumption that an agent could with equal probability be any of the agents (including herself) whose interests are at stake in a given decision-making problem. The right-hand side of the equation is the arithmetic mean of all individual utility levels in society. As can easily be seen, Harsanyi's ethical theory is a sophisticated version, expressed in the language of rational choice theory, of the idea of the impartial and sympathetic observer. The differences between Gauthier's and Harsanyi's views are easy to detect. Gauthier claims that in some situations, a rational agent must act morally, whereas Harsanyi claims only that there is a rational way of acting morally, not that rationality requires one to act morally. Thus, Harsanyi's utilitarian principle is not a necessary requirement of rationality; it is addressed to those agents who want to act morally (and thus accept the moral axiom from which, in conjunction with other axioms, the utilitarian principle can be derived). In Harsanyi's interpretation, the rationality of the moral principle (viz. the utilitarian principle) lies in the fact that it can be shown to result from a set of axioms which includes a moral axiom (the assumption of equal weight of the interests of all agents). Gauthier, in turn, claims to derive morality as a rational constraint from the set of nonmoral principles of rational choice. Thus, unlike Gauthier, Harsanyi did not intend to derive ethical rules from nonmoral premises: He did not intend, then, to offer a fundamental justification of morality.

8 Game Theory as a Tool for Analyzing Moral Decision-Making

We shall now address the question of whether rational choice theory (an essential part of which is game theory) is sufficiently general to model not only nonmoral but also moral choice. Contrary to the oft-made assertion, it shall be argued here that rational choice theory does not exhibit this sort of generality: It can only be used to model a certain range of moral choices, viz. those which are motivated by moral emotions and empathy, but not those which are principle-based or—more generally—not emotionally based. Let us first distinguish three accounts of the relation between rational choice theory and moral decision-making. According to the first account, standard rational choice theory is sufficiently general to model all types of moral choice. According to the second account, *revisionist* rational choice theory, but not the standard theory, is sufficiently general to model all types of moral choice. According to the third account, neither standard nor revisionist rational choice theory can model all types of moral choices. Let us analyze these three accounts at some length.

The first account proceeds from the assumption that a decision-maker's moral motivation can be reflected in her utility function. Let us assume that two players *P1* and *P2* play a game which has n possible outcomes (a_i, b_i), $i = 1, 2, \ldots, n$, where for each outcome i, a_i is P1's reward and b_i is P2's reward. Now, one can construct different utility functions that reflect different ways in which the players may value the outcomes. One can distinguish nonmoral utility functions and moral utility functions. Nonmoral utility functions are, for example, the following ones:

(a) If P1 is an *egoist*, then her utility function u will be $u(i) = a_i$ (this corresponds to the statement that for any two outcomes, say, 3 and 4, P1 prefers 3–4 only if $a_3 > a_4$).

(b) If P1 is *of competitive disposition*, then her utility function u will be $u(i) = a_i - b_i$ (this corresponds to the statement that for any two outcomes, say 3 and 4, P1 prefers 3–4 only if $a_3 - b_3 > a_4 - b_4$).

Moral utility functions are, for example, the following ones:

(c) If P1 is a *sympathetic player*, then her utility function u will be $u(i) = a_i + xb_i$, where $0 < x < 1$ (this corresponds to the statement that for any two outcomes, say 3 and 4, *P1* prefers 3–4 iff if $a_3 + xb_3 > a_4 + xb_4$).

(d) If P1 is a *utilitarian player*, then her utility function u will be $u(i) = a_i + b_i$ (this corresponds to the statement that for any two outcomes, say 3 and 4, P1 prefers 3–4 only if $a_3 + b_3 > a_4 + b_4$).

(e) If P1 is *altruistic*, then her utility function u will be $u(i) = a_i + xb_{\mathrm{i}}$, where $x > 1$ (this corresponds to the statement that for any two outcomes, say 3 and 4, P1 prefers 3–4 only if $a_3 + xb_3 > a_4 + xb_4$

(f) If P1 is *acting out of a sense of justice*, then her utility function u will arguably be like, e.g., $u(\mathrm{i}) = \mathrm{a_i}$ + *value accompanying the satisfied or unsatisfied sense of justice* (the value will be positive or negative, respectively).

To sum up, on this account, moral choice is a special case of rational choice, viz. it is a rational choice in a decision situation shaped by moral utility functions. Let us now pass to a critique of this account. First, to speak of moral utility functions (e.g., utilitarian or altruistic) other than those necessarily reflecting moral emotions (e.g., sympathetic) is to blur the distinction between moral agents with moral preferences (i.e., agents having a moral way of valuing outcomes) and moral agents who are deprived of moral preferences but make counter-preferential moral choices. For example, a "utilitarian" utility function may describe both the behavior of an agent who has amoral preferences but makes her choices by appeal to the utilitarian ethical criterion and the behavior of an agent who values outcomes in a utilitarian manner. Their choices are identical but reached in an essentially different decision-making process. Second, the above explication of justice-based behavior in terms of utility functions dissolves it into a set of emotions (e.g., moral anger). But justice-based behavior may be of a different kind: It may be an instance of a rule-guided behavior, i.e., where an action consistent with a rule is chosen for the reason that it is prescribed or prohibited by that rule. And being guided by moral rules is a way of moral decision-making which, as it seems, cannot be expressed by means of utility functions. An example of such a mode of moral decision-making is the categorical imperative. In view of this criticism, it seems that the first account seems suited only to moral choices motivated by moral emotions: Moral emotions have an impact on the valuation of various outcomes (the affective element of an emotion modifies the payoff structure of the decision-making situation) and so can be mirrored by utility functions. It is inadequate as a general characterization of moral decision-making because it does not adequately describe counter-preferential and/or rule-guided choices.

The second account has been defended by those who make the claim, discussed in the preceding section, that morality can be derived from instrumental rationality (e.g., by Gauthier). Let us recall that they assumed that for a choice to be moral, it must be *constrained*, i.e., counter-preferential, involving a *cooperative* strategy contrary to one's preferences. Moral choice, in other words, involves a constraint requiring that an agent act contrary to her own interests. On this account, moral choices are of two kinds: constrained and unconstrained (determined by moral emotions). The former are moral choices sensu stricto. Both can be modeled, within revisionist rational choice theory, as rational choices: The former are choices of a constrained maximizer, the latter are the choices of a straightforward maximizer. In contrast to the first account, this account implies not only the thesis that rational choice theory can model moral choices but also a much stronger thesis that moral principles can be derived from principles of instrumental rationality. Thus, according to the first account, each moral choice can be modeled as a rational choice, but rationality does not require making moral choices, whereas according to the second account, each moral choice can be modeled as a rational choice and rationality requires making moral choices. The criticisms against Gauthier's theory were formulated in the preceding section.

The third account assumes that neither standard nor revisionist rational choice theory is sufficiently general to also model moral choices. According to this account, from among two types of moral choices, viz. those determined by moral emotions and those not determined by moral emotions (counter-preferential and/or rule-guided), only the former can be plausibly modeled by means of the tools of rational choice theory. Given the criticisms of the two other accounts, this account seems most plausible. If this account is indeed correct, then one can say that the problem of moral decision-making reveals the limits of rational choice theory, since it turns out that rational choice theory does not have sufficient conceptual resources to model a certain type of moral choices (viz. counter-preferential and/or rule-guided ones).

9 Game Theory as a Tool for Analyzing the Nature of Moral Dispositions

There is a serious controversy among scholars as to whether there are any moral tendencies embedded in our biological nature, and—if so—what their exact content is. There seems to be wide agreement, however, that if moral tendencies are in fact embedded in our nature, central among them are the tendency to engage in reciprocally altruistic behavior and the tendency to manifest in some rudimentary form the sense of justice. Now, game theory can be used to hone the evolutionary arguments in favor of the existence of such tendencies.

Ref. (*reciprocal altruism*). We owe an exact characterization of human tendency to engage in reciprocal altruism to game theorists who have precisely described what kind of behavior is most effective for its user in generating beneficial outcomes in an indefinitely iterated Prisoner's Dilemma and so is likely to have been preserved by natural selection. According to evolutionary biologists and game theorists, human beings are neither universal co-operators ("suckers") nor universal defectors ("cheaters"): They are reciprocal altruists, i.e., Tit for Tat (*TFT*) players (Trivers 1971; Axelrod 1984) . *TFT* is a strategy that starts with a cooperative move and then imitates an opponent's last move. This strategy is therefore nice, reciprocating, forgiving, not envious, and clear; i.e., it is never first to defect, punishes defectors and therefore cannot be exploited by them, forgives after an opponent's period of cooperation, does not want to gain more than her opponent, and—being easily recognizable—is efficient in generating cooperation. The insight that in an indefinitely iterated Prisoner's Dilemma *TFT* is the strategy that fares best and so has the highest chance of reproductive success in reciprocal relationships is a very intuitive one: It accords with the common-sense observation that repeated interactions create the possibility of punishing defectors and so of discouraging them from defecting. This insight is supported by many common-sense observations. For instance, we tip higher in restaurants to which we are more likely to return; taxi drivers are more honest toward their clients in small cities than in big ones, since the likelihood of meeting the same taxi driver in a small city is higher than in a big one; people are more polite to each other in small cities than in large ones, often for the same reason; neighbors often help each other knowing that they will have repeated interactions. One more remark on this strategy is here in order. *TFT* can be seen to embody certain moral rules of a specific character: conditional and prudential. These rules can be presented in two equivalent ways. *TFT* can be said to embody the rule "Do unto others as you would have them do unto you only if others do unto you as they would have you do unto them," or to embrace the following rules (Axelrod 1984, 27–53; Singer 1995, 129–153) : (1) Start by cooperating; (2) do good to those who do good to you and do harm to those who do harm to you; (3) be forgiving; (4) do not be envious. Since *TFT* does not prescribe unconditional moral action, it cannot be regarded as embodying a moral norm sensu stricto: It is a prudential rule, i.e., one aimed at serving our own long-term interests. But, arguably, it can be said to be a good starting point for developing a more mature moral consciousness.

Ref. (*rudimentary sense of justice*). Another game-theoretic argument that can be used to justify the claim of evolutionary biologists that our moral dispositions have biological roots has been proposed by Skyrms (1996, 1–22) and developed by McKenzie (2007, 148–198). It should be pointed out, however, that both Skyrms and McKenzie are more inclined to give their argument (and the other arguments they provide) a cultural rather than biological interpretation; i.e., they are not inclined to maintain that moral dispositions are a product of natural/sexual selection, but rather claim that these dispositions have evolved in the process of cultural evolution (this interpretation will be analyzed in greater detail in Sect. 11). However, these models can also be invoked as supporting the claim that moral dispositions are (at least partly) a product of natural/sexual selection.[11] The gist of the argument can be summarized in the following way. Imagine a population composed of three types of players engaged in a "Division of a Cake" game in which each of two players asserts a claim (from 0 to 1) to a given good—cake or whatever: If the sum of their claims exceeds 1, then each receives nothing; if it is smaller than or equal to 1, then each receives what she claimed; fair players (F) always demand exactly 1/2 of the good; greedy players (G) always demand 2/3 of the good; and modest players (M) always demand 1/3 the good. Assume that the players meet regularly and have to bargain over the good. A Nash equilibrium in this game is any combination of claims whose sum amounts to 1 (all those Nash equilibria are strong; i.e., by unilaterally deviating from it, an agent will worsen her situation).[12] Thus, the Nash equilibria arise if two *Fs* meet

[11] The view that our moral dispositions have been shaped by natural selection is developed, e.g., in Gibbard 1990. I analyze this view at length in Zaluski 2009.

[12] A Nash equilibrium may be strong or weak. It is *strong* if for each player the condition holds that her unilateral deviation from this outcome would cause her utility loss. It is *weak* if for at

each other or if *G* meets *M*. What is important in the context of the analyses pursued here is that *F* is the only pure evolutionary stable strategy. As can be easily noticed, the population of *Fs* cannot be invaded either by *Gs* or by *Ms*. If *Fs* bargain with each other over the good, each of them will obtain 1/2 of it. If *G* plays against *F*, will *G* obtain nothing (nor, therefore, will *G* fare as well against *F* as *F* will against himself, much less will *G* fare better than F); likewise, if *M* plays against *F*, *M* will obtain 1/3 of the good (nor, therefore, will *M* fare as well against *F* as *F* will against himself, much less will *M* fare better than F). The analysis, however, is incomplete because it does not take into account the possibility of evolution preserving a polymorphic state of the population. For instance, consider the population state in which half the population plays *G* and the other half plays *M*. Now, this polymorphic state is also stable. However, as Skyrms and McKenzie Alexander note, it is less efficient than the monomorphic population of *F* players: The average payoff for *F* players is 1/2, and the average payoff in the aforementioned polymorphic population is 1/3. This difference has far-reaching implications for the prediction of the evolution of the population. The replicator dynamics analysis shows that it is more probable that the population will evolve toward the monomorphic state of *F* players than toward any polymorphic state; in technical jargon: The basin of attraction for the former is larger than the basin of attraction of any polymorphic state (let us recall that a basin of attraction of a given final state *s* represents the number of initial states from which evolution proceeds to *s*). The probability becomes even higher when one allows for the effects of correlation, i.e., if one replaces the assumption of random interactions with the assumption of correlated interaction (so that *Fs* can interact only with *Fs*). Skyrms and McKenzie Alexander have shown that even small correlations have very large effects. Thus, the general message of this analysis is optimistic: It is highly probable that some of our moral dispositions are the product of biological evolution, since *F* is the only evolutionary pure stable strategy in the game "Division of a Cake" and is also more likely to dominate the population than the other evolutionary stable strategies (the polymorphic ones). This message was strengthened by McKenzie Alexander's agent-based models (of the evolution of cooperation, trust, fairness, and retaliation), since the models have shown that in two-person games (though to a much lesser extent in multi-person games) structured interactions between agents are even more likely to lead to the emergence of "moral behavior" than nonstructured ones (based on the assumption that the probability of interactions between any two members of a population is constant). McKenzie Alexander summarizes his results in the following way:

> […] I have attempted to show how boundedly rational individuals who face decision problems in structured environments, and who make choices using rules like *Imitate the Best*, would learn to behave morally. I have argued that such individuals would learn to cooperate in the

least one player the condition holds that her unilateral deviation from this outcome would not cause her utility loss but would just not improve her situation, i.e., would yield her the same utility as her equilibrium strategy. Thus, in a weak Nash equilibrium, a player has no incentive to play her equilibrium strategy if she expects her opponent to play her own Nash strategy: She simply does not have an incentive *not* to play her equilibrium strategy in such a situation; she is therefore indifferent between playing and not playing her equilibrium strategy.

> prisoner's dilemma, trust in the Stag Hunt, share equally in resource-allocation problems, and even (in some limited cases) behave fairly while adopting punitive behavior for unfair offers in the ultimatum game. I have also argued that many of these tendencies persist when we move from considering two-player games to N-player games. These results are not conclusive, of course, but they are, I believe, better than merely suggestive. (McKenzie Alexander 2007, 290)

10 Game Theory as a Tool for Analyzing the Functions of Moral Emotions

Game theory can also be helpful in the analysis of the evolutionary function of moral emotions. Two main hypotheses regarding this function have been proposed in the literature: the hypothesis of emotions as mimicking the *TFT* strategy and the hypothesis of emotions as solving impulse-control and commitment problems.

According to the first hypothesis (e.g., Trivers 1971), certain moral emotions (e.g., benevolence, anger, guilt, forgiveness) evolved because they were effective in implementing the *TFT* strategy in an iterated version of the Prisoner's Dilemma. These emotions were preserved by natural selection because they "moved" the agents to take actions that would have been taken were the agents able to carry out rather complicated calculations that a fully rational player is expected to make. Thus, agents who act on these emotions act as if they played the *TFT* strategy. Therefore, these emotions are especially effective in supporting relationships of reciprocal altruism.

According to the second hypothesis (Schelling 1978; Hirshleifer 1987; Frank 1988) , moral emotions evolved because they helped overcome impulse-control problems—i.e., the temptation to choose options serving our short-term rather than our long-term interests—and commitment problems, i.e., problems arising "when it is in a person's interest to make a binding commitment to behave in a way that will later seem contrary to self-interest" (Frank 1988, 47).[13] As for impulse-control problems, moral emotions help solve them by providing an agent with additional incentives to choose options that serve her long-term self-interest; they therefore enhance the agent's ability to take account of long-term consequences by acting as current motivational proxies. Let us present this insight in a more precise way. In the case of an agent A with no moral emotions, at time t (the moment of making a choice), the expected utility of option x serving A's short-term interest is greater than the expected utility of option y serving A's long-term interest. But in the case of an agent A with moral emotions (e.g., an emotion of shame), at time t (the moment of making a choice), the sum of expected utility of option x serving A's short-term interest and the disutility caused by moral emotions activated by choosing x is smaller than the expected utility of option y serving A's long-term interest. Moral emotions can also provide an advantage in strategic interactions; i.e., they help solve commitment problems. For instance, a player's reputation for being inclined to experience moral

[13] It should be noticed, though, that *nonmoral* emotions are one of the main causes of impulsive, akratic behavior.

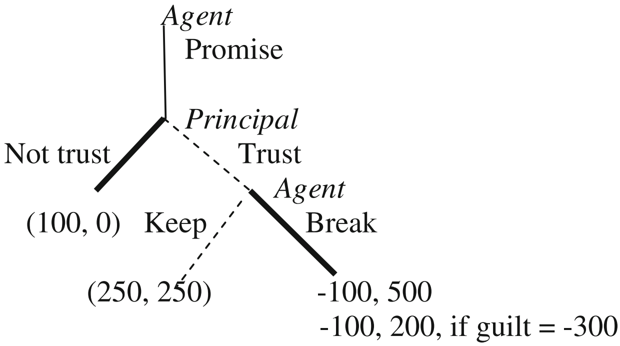


Fig. 6 Agency game

emotions (e.g., guilt, shame, moral indignation) is likely to make her promises and threats credible and hence to serve her long-term interests. Thus, e.g., the disposition of a player *A* to have feelings of guilt if she plays noncooperatively is likely to make her promise to player *P* to play cooperatively credible and as a consequence is likely to induce *P* to play cooperatively in a game with *A*. Let us illustrate this last point with the game of agency. In this game, there are two players: an agent and a principal. An agent can increase (say, by 500%) the amount of money held by the principal (say, $100). An agent promises to the principal that he will give him back one-half of the increased sum (i.e., $250). The principal may trust the agent (and give him the money) or not trust him (and not give him the money). Assuming that the agent is a *homo oeconomicus* (and so is not prone to experience feelings of guilt), the result of the game (marked by a bold line and established by means of reasoning that in game-theoretic parlance is termed "backward induction") will be (100, 0): The principal will not trust the agent's promise. But if the agent is prone to experience feelings of guilt and the principal is aware of this fact, then the result of the game (marked by a dotted line and established by means of "backward induction") will be (250, 250) (it is assumed that guilt generates costs to the agent equivalent to the loss of $300) (Fig. 6).

11 Game Theory as a Tool for Analyzing the Cultural Evolution of Moral Norms

Many game theorists (e.g., Axelrod 1986; Binmore 2005; Gintis 2000; Skyrms 1996, 2004; McKenzie Alexander 2007) maintain that the cultural evolution of moral norms can be gainfully modeled by means of game theory. It should be noted at the outset that the models presented in Sect. 9 can also be interpreted (and were indeed interpreted by their authors) as modeling cultural rather than biological evolution (whether one assumes that they can also be plausibly interpreted as models for the biological evolution of our dispositions to act morally will depend on whether one is ready to assent to the claim of evolutionary psychologists that these dispositions are, at least in their rudimentary forms, products of natural and sexual selection). In this section, the considerations pursued in Sect. 9 shall be complemented by presenting Ken Binmore's game-theoretic account of the cultural evolution of moral norms. According to Binmore, for a moral norm to be preserved by cultural evolution, it must satisfy three conditions: stability (it must be a Nash equilibrium in a game of life), efficiency (it must be Pareto-optimal), and fairness. Binmore makes use of the theory of indefinitely repeated games to model the emergence of moral norms. In particular, he invokes the so-called Folk Theorem of this theory, which says that an indefinitely repeated Prisoner's Dilemma yields equilibria of any mixture of cooperation and defection at least as good as mutual defection. This theorem has important implications for the problem of the emergence of moral norms: (a) It implies that the equilibrium condition is insufficient for explaining the emergence of a given norm, as it demonstrates that in indefinitely repeated games there is a great number of efficient equilibria; (b) it determines the class of moral norms that can survive as an equilibrium, i.e., are feasible; (c) it implies that every norm on which rational players might agree in the presence of external enforcement is available as an equilibrium outcome—sustained by reciprocity mechanisms—in the indefinitely repeated Prisoner's Dilemma. Now, according to Binmore, moral norms (especially fairness norms) evolved as an equilibrium-selection device in games of life—a device that enables societies to coordinate on one of these equilibria admitted by the Folk Theorem. Human beings therefore use culture to solve the equilibrium-selection problem. Binmore claims that societies which managed to coordinate on fairness norms have survived, and those which failed to reach such an coordination have perished. The question arises as to what fairness norms can be expected to be agreed upon in the course of cultural evolution as an equilibrium-selection device. Binmore's response is that in noniterated games, the parties might agree (as Harsanyi believed) upon the utilitarian solution, but in indefinitely iterated games, in which agents can renegotiate the terms of the contract, the only viable solution is egalitarian (in the version of the Nash bargaining scheme). The utilitarian solution is unstable because, to use Rawls's phrase, it imposes on the agents "strains of commitment" that are too high. Since this solution admits of a situation in which some agents, for the sake of maximizing social utility, must sustain much higher costs than other agents, it could be stable only in the presence of some enforcing institution. More generally, according to Binmore, the egalitarian solution (in the version of the Nash bargaining scheme) is the only one that is stable and thus puts an end to the process of renegotiation. This part of Binmore's analysis is controversial but rather original and insightful. It is controversial because it leads to two, by no means self-evident, predictions that every existing moral norm, i.e., every moral norms preserved by cultural evolution, will be Pareto-optimal and that every existing Pareto-optimal norm can, at least approximately, be interpreted as incorporating the Nash bargaining solution. Now, it seems that the first prediction is overly optimistic, and the second one is rather difficult to test. The part of Binmore's analysis just presented can be termed "social-scientific." Binmore supplements this part with philosophical comments, but this is a decidedly less worked-out part of his project. He seems to believe that morality is *nothing more* than the prevailing equilibrium in a given society, and so that his analysis of moral norms in terms of equilibrium selection provides an exhaustive account of morality. Accordingly, he appears to assume a naturalistic definition of morality (morality as a convention solving the problem of equilibrium selection) with relativistic implications. This definition implies that one cannot meaningfully speak of fairness in abstraction from fairness as realized in a concrete society. Accordingly, it leads to the conclusion that we cannot judge one moral norm (i.e., one equilibrium-selection device) as morally better than another because there do not exist independent criteria of right and wrong. Furthermore, Binmore arbitrarily assumes that purely moral motivation is nonexistent: He claims that we are egoistic species, with inclinations to lying and cheating whenever possible. This implies that altruism and the sense of duty are emergent phenomena that can be explained in terms of self-interest. But, fortunately, it is possible to separate the social-scientific (insightful and original) part of Binmore's analysis from the philosophical (rather superficial) one.

12 Conclusions

The foregoing considerations seem to justify the thesis that to use Braithwaite's phrase, game theory can be a useful "tool for the moral philosopher." It can deepen and sharpen an analysis of both descriptive and normative questions in moral philosophy. As for the latter, it helps us tackle in a novel way, e.g., the problem of a function of morality, the problem of the origins of the disposition to act morally, the problem of the function of moral emotions, and the problem of the cultural evolution of moral norms. As for the latter, it can be especially helpful in determining the content of the notoriously vague concept of justice, as well as in evaluating certain moral conceptions (especially instrumental ethical egoism). There are also other ways of applying game theory in moral philosophy which cannot be characterized as either descriptive or normative, e.g., applying game theory in analyzing the problem of the validity of moral norms and in analyzing moral decision-making. The problem of the validity of moral norms has not, to my knowledge, been analyzed at greater length as yet, but I think it deserves a thorough analysis. What was provided in Sect. 6 is only a rough sketch of the problem. As for the question of the possibility of modeling moral choices by means of rational choice theory, this assuredly also deserves a more in-depth analysis than the one provided in Sect. 8.

References

Axelrod, R. 1984. *The evolution of cooperation*. New York: Basic Books.

Axelrod, R. 1986. An evolutionary approach to norms. *American Political Review* 80: 1095–1111.

Axelrod, R., and W.D. Hamilton. 1981. The evolution of cooperation in biological systems. *Science* 211: 1390–1396.

Bicchieri, C. 2004. Rationality and game theory. In *The Oxford handbook of rationality*, ed. A.R. Mele, and P. Rawling, 182–205. Oxford: Oxford University Press.

Bicchieri, C. 2006. *The grammar of society. The nature and dynamics of social norms*. Cambridge, Mass: Cambridge University Press.

Binmore, K. 1994. *Playing fair*. *Game theory and the social contract*, vol. I. Cambridge, Mass: The MIT Press.

Binmore, K. 1998. *Just playing*. *Game theory and the social contract,* vol. II. Cambridge, Mass: The MIT Press.

Binmore, K. 2005. *Natural justice*. Oxford: Oxford University Press.

Braithwaite, R.B. 1955. *Theory of games as a tool for the moral philosopher*. Cambridge: Cambridge University Press.

Brams, S. 1990. *Negotiation games*. New York: Routledge.

Danielson, P. 1992. *Artificial morality*. London and New York: Routledge.

Frank, R. 1988. *Passions within reason: The strategic role of the emotions*. New York, London: W.W. Norton and Company.

Gauthier, D. 1986. *Morals by agreement*. Oxford: Clarendon Press.

Gauthier, D. 1988. Morality, rational choice, and semantic representation. *Social Philosophy and Policy* 5: 173–221.

Gibbard, A.F. 1990. *Wise choices, apt feelings. A theory of normative judgment*. Oxford: Clarendon Press.

Gintis, Herbert. 2000. *Game theory evolving*. Princeton: NJ: Princeton University Press.

Hardin, G. 1965. The tragedy of the commons. *Science* 162: 1243–1248.

Hardin, R. 1988. *Morality within the limits of reason*. Chicago, London: The University of Chicago Press.

Hardin, R. 2007. *David Hume. Moral and political theorist*. Oxford: Oxford University Press.

Harsanyi, J. 1976a. Ethics in terms of hypothetical imperatives. In *Essays on ethics, social behavior and scientific explanation*, ed. J. Harsanyi, 24–36. Dordrecht: Reidel Publishers.

Harsanyi, J. 1976b. Advances in understanding rational behavior. In *Essays on ethics, social behavior and scientific explanation*, ed. J. Harsanyi, 89–117. Dordrecht: Reidel Publishers.

Harsanyi, J. 1983. Morality and the theory of rational behavior. In *Utilitarianism and beyond*, ed. A. Sen, and B. Williams, 39–62. Cambridge: Cambridge University Press.

Hirshleifer, J. 1987. On the emotions as guarantors of threats and promises. In *Latest on the best: Essays on evolution and optimality*, ed. J. Dupré, 307–326. Cambridge, Mass: The MIT Press.

Hobbes, T. 1998. *On the citizen*, trans. Richard Tuck, Michael Silverthorne. Cambridge: Cambridge University Press (1st ed. in Latin in 1641.).

Hume, D. 1978. *A treatise of human nature*. Oxford: Clarendon Press (1st ed. in 1739.).

Hume, D. 1998. *An enquiry concerning the principles of morals.* Oxford, New York: Oxford University Press (1st ed. in 1751.).

Lewis, C.S. 1980. *Mere Christianity*. New York: HarperCollins (1st ed. in 1952).

Lewis, D. 1969. *Convention. A philosophical study*. Oxford: Basic Blackwell.

Mackie, J.L. 1977. *Ethics: Inventing right and wrong*. New York: Penguin Books.

Marmor, A. 2009. *Social conventions. From language to law*. Princeton and Oxford: Princeton University Press.

Maynard Smith, J. 1982. *Evolution and the theory of games*. Cambridge: Cambridge University Press.

McClennen, E. 1990. *Rationality and dynamic choice: Foundational explorations*. Cambridge: Cambridge University Press.

McKenzie Alexander, J. 2007. *The structural evolution of morality*. Cambridge: Cambridge University Press.

Nozick, R. 1974. *Anarchy, State, Utopia*. New York: Basic Books.

Olson, M. 1965. *The logic of collective action*. Cambridge, Mass.: Harvard University Press, Cambridge.

Ossowska, M. 1970. *Normy moralne. Próba systematyzacji*. Warsaw: PWN.

Peyton Young, H. 1995. *Equity. In theory and practice*. Princeton, NJ: Princeton University Press.

Postema, G.J. 1998. Rationality, conventions, and law: Introduction. *Law and Philosophy* 17: 347–350.

Schelling, T. 1960. *The strategy of conflict*. Cambridge, Mass.: Harvard University Press.

Schelling, T. 1978. Altruism, meanness, and other potentially strategic behaviours. *American Economic Review* 68: 229–230.

Singer, P. 1995. *How are we to live? Ethics in an age of self-interest*. New York: Prometheus Books.

Skyrms, B. 1996. *Evolution of the social contract*. Cambridge: Cambridge University Press.

Skyrms, B. 2004. *The stag hunt and the evolution of social structure*. Cambridge: Cambridge University Press.

Sugden, R. 1986. *The economics of rights, co-operation and welfare*. Oxford: Basic Blackwell.

Taylor, M. 1987. *The possibility of cooperation*. Cambridge: Cambridge University Press.

Taylor, P., and L. Jonker. 1978. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences* 16: 76–83.

Trivers, R.L. 1971. The evolution of reciprocal altruism. *Quarterly Review of Biology* 46: 35–57.

Ullmann-Margalit, E. 1977. *The emergence of norms*. Oxford: Clarendon Press.

Verbeek, B., M. Christopher. 2004. Game theory and ethics. In *The Stanford Encyclopedia of philosophy* (Winter 2004 Edition), ed. E.N. Zalta. http://plato.stanford.edu/archives/win2004/entries/game-ethics/.

Weibull, J.W. 1995. *Evolutionary game theory*. Cambridge, Mass: The MIT Press.

Załuski, W. 2009. *Evolutionary theory and legal philosophy*. Cheltenham: Edward Elgar.

Załuski, W. 2013. *Game theory in jurisprudence*. Krakow: Copernicus Center Press.

Part III
Special Kinds of Legal Reasoning

Evidential Reasoning

Marcello Di Bello and Bart Verheij

When a suspect appears in front of a criminal court, there is a high probability that he will be found guilty. In the USA, statistics for recent years show that the conviction rate in federal courts is roughly 90%, and in Japan reaches as high a rate as 99%.[1] In the UK, the numbers are slightly lower, with a conviction rate of roughly 80%, while in the Netherlands the conviction rate is around 90%.[2] This does not mean that the fact finders deciding about the facts of a case have an easy job. Whether laypeople, such as jury members selected from the general public, or professionals, often experienced judges having completed postgraduate education, all face the difficulties associated with handling the evidence that is presented in court. What to do with conflicting testimonies? Does an established DNA match outweigh the testimony that the suspect was not seen at the crime scene? How to coherently interpret a large body of evidence? When is there enough evidence to convict?

The primary aim of this chapter is to explain the nature of evidential reasoning, the characteristic difficulties encountered, and the tools to address these difficulties.

[1] On the conviction rate in US federal courts, see the statistical reports of the Offices of the United States Attorneys, available at www.justice.gov/usao/resources/annual-statistical-reports. Most of these convictions are guilty pleas, not convictions after trial. On Japan's conviction rate, see *White Paper on Crime 2014*, Part 2, Chap. 3, Sect. 1, available at http://hakusyo1.moj.go.jp/en/63/nfm/mokuji.html.

[2] On the UK conviction rate, see *Criminal Justice Statistics–March 2014*, available at www.gov.uk/government/statistics. As in the US case, the rate include mostly guilty pleas. For the Netherlands, see CBS, the Dutch central bureau of statistics, publishing its data at www.cbs.nl.

M. Di Bello (✉)
Lehman College - City University of New York, Bronx, USA
e-mail: marcello.dibello@lehman.cuny.edu

B. Verheij
Faculty of Science and Engineering, University of Groningen, Groningen, The Netherlands
e-mail: bart.verheij@rug.nl

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_16

Our focus is on evidential reasoning in criminal cases. There is an extensive scholarly literature on these topics, and it is a secondary aim of the chapter to provide readers the means to find their way in historical and ongoing debates.

This chapter does not aim to offer legal practitioners, lawyers, judges and expert witnesses, practical tools that can be immediately used to better litigate a case. But we hope that practitioners interested in the theoretical underpinnings of evidential reasoning in court will benefit from reading this chapter. And, more generally, philosophers, legal scholars, statisticians, logicians, and those scholars and practitioners interested in the theoretical aspects of reasoning with evidence will hopefully find this chapter an interesting resource and point of departure for further thinking on the matter.

1 Setting the Stage

We set the stage by using two important and often encountered kinds of evidence as an illustration: eyewitness testimony and DNA profiling. These two kinds of evidence will be used to establish a list of central questions that structure the exposition that follows.

1.1 Eyewitness Testimony

Eyewitness testimony has always been a central source of information in criminal proceedings. It typically takes the form of oral statements by the witness in court, in response to questions by the prosecution, the defense, the court, and sometimes, albeit rarely, the jury. Eyewitness testimony can also come in the form of reports of oral examinations written in the pre-court stages of the criminal investigation.

Eyewitness testimony can provide information about what happened on the scene of the crime. Here is an example.

> Q: Can you describe what happened that day?
>
> A: I was in the park and suddenly heard a lot of noise, close by. I saw two men quarreling, shouting. Suddenly one of them pulled a gun, and I heard a shot. The other man fell to the ground. The shooter looked around, looked me in the eye, and then started to run.
>
> Q: Can you describe the shooter?
>
> A: He was a young men, in his twenties, I think. Tall, blonde, with a white skin, and unusually blue eyes. He looked unhealthy, with bad teeth, like a drug addict. He was wearing a perfectly ironed shirt, which surprised me.

The information contained in the testimony can be more or less detailed, and on its basis, the fact finders can form a hypothesis about what happened. Still, it remains a hypothesis. There are many reasons why the hypothetical events reconstructed on the basis of the testimony might not be true. Typical reasons against the truth of the events reported by an eyewitness include that the witness wrongly interpreted what she saw, that time distorted her memories, or that the witness is lying.

1.2 DNA Evidence

DNA evidence has become very common in criminal cases. Perpetrators sometimes leave traces of themselves and their actions, such as pieces of hair, skin tissues, drops of blood, or other bodily fluids. By using forensic DNA technology, a genetic profile associated with the crime traces can be created, and if this profile matches with an individual's profile, this establishes a link, at least *prima facie*, between the matching individual and the crime.

What is a DNA profile? A DNA profile is determined by analyzing a number of specific locations, the so-called *loci*, of a DNA molecule. Different countries use different sets of *core loci* for their DNA profiles. For instance, the CODIS system in the USA uses 20 core loci.[3] At each specific locus, a different *allele* might occur. A core locus that is often used, called CSF1PO, has up to 16 allele types, depending on how often the molecular sequence AGAT is repeated at that location.[4] A DNA profile, then, consists in a list of allele types for a certain number of select core loci.

The evidential relevance of a DNA profile stems from the fact that although most of the structure of the DNA molecule is shared among all human beings (more than 99%), the select core loci used to construct DNA profiles are highly specific. To be sure, DNA profiles need not be unique, but their proportional frequency in a reference population is expected to be very low. So, how is this low number arrived at?

Many countries have created extensive reference databases that contain millions of DNA profiles, and these are used to assess the rarity of a profile. This is a two-step process. First, the number of occurrences of each allele at each core locus in the reference database is counted. This gives a measure of the proportional frequency of each allele at each core locus in the population. Second, the measured proportional frequencies for the alleles at the core loci are multiplied. This allows us to assess the overall proportional frequency of the DNA profile or to use a more common terminology, the *Random Match Probability*. More recently, the terminology *Conditional Genotype Probability* has also been introduced. The sets of core loci have been chosen such that Random Match Probabilities, or Conditional Genotype Probabilities, are typically small, for instance, in the order of 1 in 50 billion, amply exceeding the number of people on our planet.

There is discussion about whether we can reasonably affirm such low numbers and if it makes sense to report them. A key assumption underlying the model—used when multiplying the measured proportional frequencies of specific alleles—is that there are no dependencies among the alleles at different loci in the population considered. This assumption does not always hold, for instance, in a population with

[3] See www.fbi.gov/services/laboratory/biometric-analysis/codis.

[4] See www.cstl.nist.gov/strbase/str_CSF1PO.htm.

family relations. Scientists have also established certain dependencies among the profiles within ethnic groups. Testing the independence assumption can be hard and would require the assessment of more profiles than reasonably possible.

With this background in place, suppose now that a trace of blood—which allegedly came from the perpetrator—is found at the crime scene and that the DNA profile created from the trace matches the DNA profile of a suspect. The match lends support to the hypothesis that the suspect—as opposed to an unknown individual—is the source of the blood trace, and the Random Match Probability associated with the profile provides a measure of the evidential strength of the match. Importantly, the hypothesis that a DNA match can support is rather circumscribed. It is limited to the suspect being the *source* of the trace and should not be confused with the hypothesis that the suspect is *factually guilty*, at least absent other information about how the trace got there.[5] Further, given the declared match, the hypothesis itself that the suspect is the source need not be true. We should always be wary of possible laboratory errors and false positive matches. And even if no laboratory error occurs, the suspect and the perpetrator, while different individuals, might share the same DNA profile, either because they are identical twins or because, though unrelated, they happen to share the same profile by sheer coincidence.[6]

1.3 Central Questions

Using the two kinds of evidence as an illustration, we now provide a list of central questions about evidential reasoning in the law. These questions will structure the discussion that follows:

Question 1: How should we understand conflicts between pieces of evidence? Legal disputes often occur because the evidence provides conflicting perspectives on the crime. For instance, a witness claims that the criminal has blond hair, but the suspect whose DNA matches the traces at the crime scene has dark hair. How should we understand conflicts between pieces of evidence? What are the different ways in which such conflicts arise?

Question 2: How should we handle the strength of the evidence? Some evidence is stronger than other evidence. This is most obvious in the case of DNA evidence, where DNA profiles are associated with different Random Match Probabilities or Conditional Genotype Probabilities. But also some eyewitness testimonies are stronger than others. For instance, the description of a criminal by a witness who could only view the crime scene in bad lighting conditions is of lesser value. How to address the strength of evidence?

[5] We use the terms “factually guilty” or simply “guilty” to express factual guilt, which is not the same as the legal verdict of guilt.

[6] At a rate of a dozen or more twin births per 1000 live births, identical twins are not that rare. Source https://en.wikipedia.org/wiki/Twin#Statistics.

Question 3: How should we coherently interpret the available evidence? A DNA match can support the claim that the suspect is the source, and a witness can add information about how the crime was committed. In general, there is a lot of evidence that needs to be coherently combined in order to make sense of what has happened. How do we combine all information in a coherent whole?

Question 4: How should we decide about the facts given the evidence? When are we done? After a careful and exhaustive investigation in the pretrial and trial phases of the criminal proceedings, the question arises of when a decision can be made and what that decision is. When is the burden of proof met? What is "proof beyond a reasonable doubt"?

The plan is as follows. In the next section (Sect. 2), we discuss three normative frameworks that can help us understand how to correctly handle the evidence. In the remaining sections, we discuss the four questions we set out above in light of the three frameworks and their distinctive features (Sects. 3, 4, 5 and 6).

2 Three Normative Frameworks

In this section, we discuss in broad outline three normative frameworks for the assessment of the evidence presented in a case: arguments, probabilities, and scenarios. Those frameworks constitute systematic and well-regulated methods for examining, analyzing, and weighing the evidence. In this section, we shall only briefly emphasize their distinctive theoretical strengths. Arguments can naturally capture the dialogical dimension, by modeling relations of support and attack. These are the issues raised by the first question above about conflicting pieces of evidence. Probabilities are better suited to quantify the value of the evidence. This is the issue raised by the second question about different pieces of evidence having different strengths. Finally, scenarios are best in offering a coherent and holistic interpretation of large bodies of evidence. These are the issues raised by the third question above about combing different pieces of evidence in a coherent whole. Neither framework, by itself, is well suited to address the fourth question about reaching a decision on the basis of the evidence. As we shall see toward the end, each framework will have to be supplemented with a decision-theoretic component.

2.1 Arguments

The first normative framework that we discuss uses arguments as its primary tool. Arguments are best analyzed in a dialogical setting, for they contain reasons that *support* or *attack* a certain conclusion of interest. For instance, when a witness reports that she saw the suspect at the crime scene, this evidence constitutes a reason for the conclusion that the suspect was, in fact, at the crime scene. But if the DNA profile found at the crime scene does not match the suspect's DNA profile, this constitutes

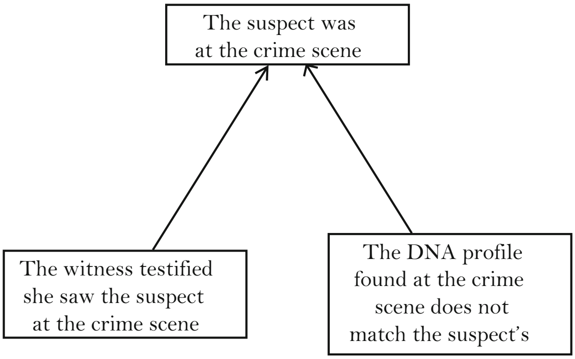

The suspect was at the crime scene
The witness testified she saw the suspect at the crime scene
The DNA profile found at the crime scene does not match the suspect's

Fig. 1 Arguments with supporting and attacking reasons

a reason attacking the conclusion. An argument with a supporting and an attacking reason is represented in Fig. 1.

The analysis of the structure of arguments goes back to the early twentieth century when John Henry Wigmore (1913) developed his famous evidence charts. The work by Anderson et al. (2005) continued from Wigmore's insights. Independently, and not focusing on evidence in criminal cases, the structure of arguments for and against conclusions was formalized and studied computationally by the philosopher John Pollock (1987, 1995). Pollock's work stimulated an extensive literature on the formal and computational study of arguments (van Eemeren et al. 2014a).

2.2 Probabilities

The second normative framework uses probabilities as its primary tool. In handling evidence in court, a crucial question from the probabilistic perspective is, how probable is a certain hypothesis H given a body of evidence E? This is the *conditional probability* of H given E, or in symbols, $\Pr(H|E)$. Another crucial question is, how does the probability of H change in light of evidence E? This *probability change* is expressed by the difference between the so-called posterior probability $\Pr(H|E)$ and prior probability $\Pr(H)$. Both questions can be addressed with Bayes' theorem:

$$\Pr(H|E) = \frac{\Pr(E|H)}{\Pr(E)} \cdot \Pr(H).$$

This formula—which can be easily proven from the probability axioms—shows how the posterior probability $\Pr(H|E)$ of hypothesis H given evidence E can be computed by the prior probability $\Pr(H)$ and the factor $\Pr(E|H)/\Pr(E)$.[7]

The interest in probabilistic calculations as a tool for the good handling of the evidence has recently been stimulated by the statistics related to DNA profiling and by some infamous miscarriages of justice that involved statistics, in particular the Lucia de Berk and Sally Clark cases (Dawid et al. 2011; Fenton 2011; Schneps and Colmez 2013). The interest is not new (Finkelstein and Fairley 1970; Tillers 2011) and can in fact be traced back to early developments of probability theory (Bernoulli 1713; Laplace 1814) and forensic science in the late nineteenth century (Taroni et al. 1998). To what extent probabilistic calculations have a place in courts has always been, and remains, the subject of debate (Fenton et al. 2016; Tribe 1971).

2.3 Scenarios

Finally, the third normative framework centers around scenario analysis. In a scenario, a coherent account of what may have happened in a case is made explicit. Scenario analysis proves helpful when considering a complex case and its evidence. For instance, the following brief scenario can help to make sense of a murder case:

> The robber killed the victim when caught during a robbery but lost a handkerchief.

This scenario can make sense of a number of facts, for example, that no one in the victim's circle of acquaintances is a possible suspect; that there are signs someone broke into the victim's apartment; and that a handkerchief was found on the floor although it does not belong to the victim. Such a unifying explanation in the form of a scenario can be regarded as a sense-making tool for handling cases with a large dossier.

Legal psychology has contributed to our knowledge about the role of scenarios in handling the evidence (Bennett and Feldman 1981; Pennington and Hastie 1993b). Scenario analysis is also connected with inference to the best explanation (Pardo and Allen 2008). Scenarios, however, can be misleading. Experiments have shown that a false scenario told in a sensible chronological order can be more persuasive than a true scenario whose events are told in a random order. Still, the legal psychologists Wagenaar et al. (1993) have emphasized the usefulness of scenario analysis for the rational handling of the evidence. In their work, they use scenario analysis for debunking dubious case decisions.

[7] Bayes' theorem can be derived using the definition of conditional probability. We have $\Pr(E|H) = \Pr(H \wedge E)/\Pr(H)$. Here, we use logical conjunction $\wedge$ to write the combined event H and E. Since $\Pr(H \wedge E) = \Pr(E|H) \cdot \Pr(H)$, it follows that $\Pr(H|E) = \Pr(H \wedge E)/\Pr(E) = \Pr(E|H) \cdot \Pr(H)/\Pr(E)$, proving Bayes' theorem. Note that the theorem holds generally for probability functions and does not assume a temporal ordering of taking evidence into account, as instead suggested by the terminology of "prior" and "posterior" probability. This terminology is standard in the context of Bayesian updating.

3 Conflicting Evidence

In many situations, it is clear what the facts are. In a standard case of tax evasion, for example, it will be easy to establish whether you filed for taxes on time and whether your employer paid you 100,000 dollars in 2015. Only in special circumstances, such as administrative errors, there will be something to dispute here. But cases that are litigated in court are typically more complicated. Disputes emerge because the two parties—who then become the defense and the prosecution in a criminal trial—introduce evidence that supports conflicting reconstructions of the facts. In this section, we illustrate how each of the three frameworks can represent and model conflicts between different pieces of evidence.

3.1 Arguments

In the argument-based framework, conflicting evidence is analyzed in terms of reasons for and against a certain conclusion. Consider a criminal case where a witness testifies she saw the suspect at the crime scene. The witness testimony constitutes a reason supporting the conclusion that the suspect indeed was at the crime scene. This can be understood as an argument *from* "a witness testified she saw the suspect at the crime scene" *to* "the suspect was in fact at the crime scene." This argument consists of three parts: the conclusion; the reason (also called the premise); and the connection between the reason and the conclusion. In what follows, we describe three ways this argument can be attacked and three symmetric ways the same argument can be further supported by additional reasons.

Three kinds of attack can be distinguished: rebutting, undercutting, and undermining. Consider the argument that the suspect was at the crime scene because the witness reports that she saw the suspect at the crime scene. First, the conclusion can be attacked. For example, suppose DNA testing shows that the suspect does not genetically match with the traces found at the crime scene. Such an attacking reason is called a *rebutting attack*. It supports the opposite conclusion, namely that the suspect was *not* at the crime scene. Second, the reason itself can be attacked. For instance, if the witness never actually testified that she saw the suspect at the crime scene, this attacks the existence of the supporting reason itself. This kind of attack is referred to as *undermining attack*. Third, the connection between the reason and the conclusion can be attacked. The fact that the lighting conditions were bad when the witness saw the crime is an example of such an attack, referred to as an *undercutting attack*. In contrast with a rebutting attack, an undercutting attack provides no support for the opposite conclusion. In the example, if the lighting conditions were bad, there would be no reason explicitly supporting that the suspect was not at the crime scene. The three examples of the different kinds of attack are shown in Fig. 2.

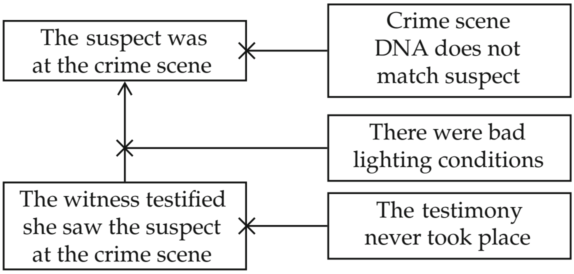

The suspect was at the crime scene
Crime scene DNA does not match suspect
There were bad lighting conditions
The witness testified she saw the suspect at the crime scene
The testimony never took place

Fig. 2 Three kinds of attack

Three kinds of support can be distinguished: multiple, subordinated, and coordinated. Just as attacking reasons can target the conclusion of an argument, its supporting reason, or the connection between the two, additional reasons can provide further support for each of these parts. Additional reasons can be seen as responses to attacking reasons or as reasons strengthening an existing argument.

Consider, once again, the argument that the suspect was at the crime scene because the witness reports that she saw the suspect at the crime scene. First, the conclusion can be further supported, for example, by a second witness testimony. If a conclusion is supported by more than one reason, this is referred to as *multiple support*. Second, the reason itself can be supported, for example, by a video recording of the witness testimony itself. Support of the reason itself is called *subordinating support*. Finally, the connection between the reason and the conclusion can be further supported, for example, by another testimony that the witness has always been trustworthy and reliable. Support for the connection between the reason and the conclusion does not have a standard name, but is closely related to a third named kind of support: *coordinated support*. In coordinated support, the support for the conclusion consists of at least two supporting reasons which, in their conjunctive combination, provide support for the conclusion. Coordinated support is distinguished from multiple support because in the latter each supporting reason provides support for the conclusion by itself.

Figure 3 shows the three kinds of (further) support. Multiple and subordinated support are graphically visualized with an arrow, whereas coordinated support is shown with a line. An arrow indicates the support of the connection between reason and conclusion.

Arguments can involve complex structures of supporting and attacking reasons. So far we have looked at an elementary argument, consisting of a reason and a conclusion, along with three types of attacking reasons and three types of symmetric supporting reasons. But an argument can also be more complex; for example, it can contain *chains of reasons*.

Consider, once again, the example of a witness who reports that she saw the suspect at the crime scene. The witness testimony constitutes a reason supporting the conclusion that the suspect was at the crime scene, and this conclusion—in turn—functions as a reason that supports the conclusion that the suspect committed the crime. This chain of supporting reasons is graphically depicted in Fig. 4, on the left.

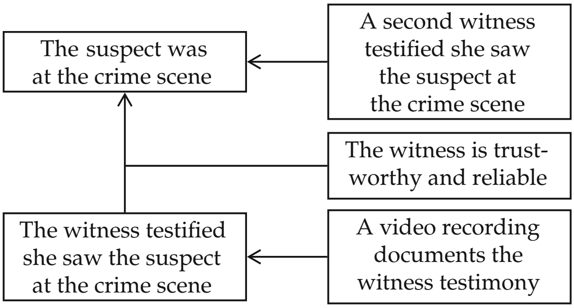

The suspect was at the crime scene
A second witness testified she saw the suspect at the crime scene
The witness is trustworthy and reliable
The witness testified she saw the suspect at the crime scene
A video recording documents the witness testimony

Fig. 3 Three kinds of (further) support

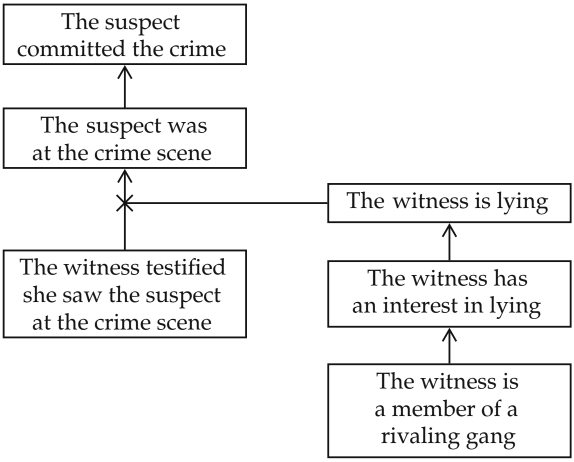

The suspect committed the crime
The suspect was at the crime scene
The witness is lying
The witness testified she saw the suspect at the crime scene
The witness has an interest in lying
The witness is a member of a rivaling gang

Fig. 4 Supporting and attacking reasons can be chained

Attacking reasons can also be chained. For example, when it is discovered that the witness is a member of a rivaling gang, this constitutes a reason for concluding that the witness has an interest in lying, and further, for concluding that the witness is in fact lying (Fig. 4, on the right). This conclusion attacks—undercuts, to be precise—the connection between the witness testimony and the conclusion the suspect was at the crime scene.

3.2 Scenarios

In the scenario-based framework, conflicts are analyzed by considering different scenarios about what may have happened. While in the previous framework, conflicts were modeled as conflicts between attacking and supporting reasons within arguments, here the perspective is more holistic, and conflicts are modeled as conflicts between scenarios.

There may be conflicting scenarios about what happened. The prosecution and the defense sometimes present different scenarios about what happened. In a murder case, for example, prosecution and defense may put forward the following conflicting scenarios:

S_1: The defendant killed the victim when caught during a robbery.

S_2: The victim's partner killed the victim after a violent fight between the two.

The two scenarios conflict insofar as they offer incompatible reconstructions of the killing and point to two different perpetrators.

At trial, however, while the prosecutor is expected to identify the perpetrator, the defense is not expected to identify another perpetrator. Two scenarios, then, can be conflicting even though they do not each point to a different perpetrator, such as the following:

S_1: The defendant killed the victim when caught during a robbery.

S_3: The defendant was at home with his wife.

Scenarios S_1 and S_3 are still clearly in conflict because they cannot be both true. Still, scenario S_3 does not say who killed the victim or how the crime occurred. It only asserts, in the form of an *alibi*, that the defendant did not do it.

Evidence can be explained by one scenario, but not by another. Conflicts between scenarios can also exist in relation to the evidence, for example, when one scenario can explain a piece of evidence but the other cannot. Two senses of "explanation" are relevant here. First, a scenario explains the evidence in the sense that it *predicts* the evidence. If the scenario is assumed to be true, the evidence must be (likely to be) there. There is another, albeit closely related, sense of explanation. A scenario explains the evidence in the sense that it exhibits the *causal process* by which the evidence was brought about.

To understand the difference between the two senses of explanation, consider the conflicting scenarios S_1 and S_2, one referring to the robber scenario and the other to the partner scenario described above. Suppose now that laboratory analyses find a genetic match between the DNA profile of a tissue trace found under the victim's fingernails and her partner, and it is clear that the skin tissue could not have gotten there unless there was a violent fight between the two. Scenario S_1, the robber scenario, cannot explain the presence of the trace matching the victim's partner. Scenario S_2, the partner scenario, can explain the presence of the matching trace. The explanation is that the victim's partner is the source of the trace, which was deposited during the violent fight between the two. The scenario can predict the presence of the trace, in the sense that if the scenario is assumed to be true, the matching trace must be there or likely to be there. The scenario also exhibits the causal process that brought about the trace, namely the violent fight, altercation and physical contact between the two.

However, suppose another piece of evidence is that the victim's house was in fact robbed in concomitance with the victim's death. Scenario S_1 can explain this evidence, both in terms of prediction and in terms of causal process. By contrast, scenario S_2 cannot offer the same explanation. All in all, scenarios S_1 and S_2 are not only inconsistent on their face and they also diverge in terms of the evidence that they can or cannot explain.

Scenarios can be contradicted by evidence. So far we considered scenarios that are inconsistent with one another because they cannot be both true, and also scenarios that diverge in terms of the evidence they can or cannot explain. There is another type of conflict worth discussing. This takes the form of a quasi-inconsistency between scenarios and evidence. The quasi-inconsistency occurs when the evidence taken at face value—typically testimonial, not physical evidence—asserts that such-and-such an event occurred, while the scenario denies precisely that.

Suppose a video recording shows the defendant breaking into the victim's house, and upon being discovered, killing the victim and later stealing the jewelry. This evidence contradicts scenario S_2 in which the victim's partner is the killer. More precisely, insofar as the evidence is taken at face value—that is, the video is taken to be truthful—scenario S_2 is inconsistent with the evidence, while scenario S_1 is consistent.

3.3 Probabilities

In the probability-based framework, conflicts are modeled as conflicts between pieces of evidence which support or attack a certain hypothesis, where "support" and "attack" are described in probabilistic terms.

Support can be characterized as "probability increase" or "positive likelihood ratio." A piece of evidence E supports an hypothesis H whenever E raises the probability of H, or in symbols, $\Pr(H|E) > \Pr(H)$. For example, a witness testifies that she saw the defendant around the crime scene at the time of the crime. The testimony supports the hypothesis that the defendant is factually guilty. This can be described probabilistically, as follows:

$$\Pr(\mathit{guilt}|\mathit{testimony}) > \Pr(\mathit{guilt}).$$

There is another characterization of evidential support. Instead of comparing the initial probability $\Pr(H)$ and the probability $\Pr(H|E)$ of the hypothesis given the evidence, a so-called likelihood ratio of the form $\Pr(E|H)/\Pr(E|\neg H)$ can also be used. On this account, E supports H whenever the likelihood ratio $\Pr(E|H)/\Pr(E|\neg H)$ is greater than one. This means that the presence of the evidence is regarded as more probable if the hypothesis is true than if the hypothesis is false. Given the example considered earlier, we have:

$$\frac{\Pr(\mathit{testimony}|\mathit{guilt})}{\Pr(\mathit{testimony}|\neg \mathit{guilt})} > 1.$$

These two characterizations of evidential support—in terms of probability increase and positive likelihood ratio—are in fact equivalent. For the following statement holds[8]:

$$\Pr(H|E) > P(H) \text{ iff } \frac{\Pr(E|H)}{\Pr(E|\neg H)} > 1.$$

The equivalence, however, only holds if the two hypotheses being compared in the likelihood ratio are one the negation of the other, such as *guilt* and $\neg$*guilt*.

Attack can be characterized as "probability decrease" or "negative likelihood ratio." By contrast, a piece of evidence E attacks a hypothesis H whenever E lowers the probability of H, or in symbols, $\Pr(H|E) < \Pr(H)$. For example, if a DNA test shows no match between the traces found at the crime scene and the defendant, this evidence attacks the hypothesis that the defendant is factually guilty. Probabilistically,

$$\Pr(\mathit{guilt}|\mathit{no\ DNA\ match}) < \Pr(\mathit{guilt}).$$

Similarly, a piece of evidence E attacks a hypothesis H whenever the likelihood ratio is lower than one. This means that the presence of the evidence is less probable if the hypothesis is true than if the hypothesis is false. For the example considered earlier, we have:

$$\frac{\Pr(\mathit{no\ DNA\ match}|\mathit{guilt})}{\Pr(\mathit{no\ DNA\ match}|\neg \mathit{guilt})} < 1.$$

Just as the two characterizations of evidential support are equivalent, so are the two characterizations of evidential attack, that is:

$$\Pr(H|E) < \Pr(H) \text{ iff} \frac{\Pr(E|H)}{\Pr(E|\neg H)} < 1.$$

The equivalence holds because the two hypotheses being compared in the likelihood ratio are the negation of each other.

[8] To see why, note that

$$\frac{\Pr(H|E)}{\Pr(\neg H|E)} = \frac{\Pr(E|H)}{\Pr(E|\neg H)} \cdot \frac{\Pr(H)}{\Pr(\neg H)},$$

which implies

$$\frac{\Pr(E|H)}{\Pr(E|\neg H)} > 1 \text{ iff } \frac{\Pr(H|E)}{\Pr(\neg H|E)} > \frac{\Pr(H)}{\Pr(\neg H)}.$$

To prove the left-right direction of the equivalence in the text, if $\Pr(H|E) > P(H)$, then $1 - \Pr(H|E) < 1 - \Pr(H)$. This means that $\frac{\Pr(H|E)}{1-\Pr(H|E)} > \frac{\Pr(H)}{1-\Pr(H)}$, and thus $\frac{\Pr(H|E)}{\Pr(\neg H|E)} > \frac{\Pr(H)}{\Pr(\neg H)}$. So, by the equivalence above, $\frac{\Pr(E|H)}{\Pr(E|\neg H)} > 1$. For the other direction, if $\frac{\Pr(E|H)}{\Pr(E|\neg H)} > 1$, then $\frac{\Pr(H|E)}{\Pr(\neg H|E)} > \frac{\Pr(H)}{\Pr(\neg H)}$, again by the equivalence above. The latter is the same as $\frac{\Pr(H|E)}{1-\Pr(H|E)} > \frac{\Pr(H)}{1-\Pr(H)}$. To establish $\Pr(H|E) > \Pr(H)$, suppose for contradiction that $\Pr(H|E) \leq \Pr(H)$, which implies $1 - \Pr(H|E) \geq 1 - \Pr(H)$. This means that $\frac{\Pr(H|E)}{1-\Pr(H|E)} \leq \frac{\Pr(H)}{1-\Pr(H)}$. This contradicts $\frac{\Pr(H|E)}{1-\Pr(H|E)} > \frac{\Pr(H)}{1-\Pr(H)}$, and thus $\Pr(H|E) > \Pr(H)$.

The conflict between two pieces of evidence can be described probabilistically. Two pieces of evidence come into conflict with one another insofar as one supports a hypothesis and the other attacks the same hypothesis. The conflict can be described probabilistically, in that one piece of evidence increases the probability of the hypothesis, while the other decreases it, or equivalently, the likelihood ratio is positive (for one piece of evidence) and negative (for the other).

For example, the testimony that the defendant was around the crime scene conflicts with the lack of a DNA match. Probabilistically, the testimony increases the probability of the defendant's guilt (or equivalently, the likelihood ratio is greater than one), while the lack of a DNA match decreases the probability of the same hypothesis (or equivalently, the likelihood ratio is lower than one).

4 Evidential Value

The evidence in a criminal case has different levels of evidential value: Some evidence is strong, other not so much. How is evidential value handled in each of the three normative frameworks? That is the topic of this section.

4.1 Probability

In the probabilistic framework, evidential value is quantified numerically using various concepts based on the probability calculus, that is, probabilistic difference, likelihood ratio, and conditional probability on the evidence.

The incremental evidential value is measured by probabilistic change. The incremental value of evidence for, or against, a hypothesis can be quantified probabilistically in various ways. One approach considers the difference between the probability of the hypothesis with and without the evidence, that is, $\Pr(H|E) - \Pr(H)$. The larger the positive difference, the higher the value of the evidence for the hypothesis. An alternative approach is given by the likelihood ratio $\Pr(E|H)/\Pr(E|\neg H)$. For any value greater than one, the higher the likelihood ratio, the higher the value of the evidence for the hypothesis. By contrast, a negative difference $\Pr(H|E) - \Pr(H)$ and a likelihood ratio lower than one quantify the value of the evidence *against* a hypothesis. The larger the negative difference and the lower the likelihood ratio (for any value below one), the higher the value of the evidence against the hypothesis.

Note that these two approaches parallel the two characterizations of evidential support and attack in the previous section, as probability increase/decrease and positive/negative likelihood ratio. While these notions were only qualitative, probability increases/decreases and likelihood ratios, as measures of evidential value, express quantities.

The overall evidential value is measured by the overall conditional probability. In contrast with the incremental evidential value of evidence that is measured by a probabilistic difference or likelihood ratio, the overall evidential value of the full body of evidence is measured by the conditional probability of the hypothesis given the evidence. The higher, or lower, the probability $\Pr(H|E)$, the higher the overall value of the evidence for, or against, the hypothesis. If there are different pieces of evidence $E_1, \ldots, E_k$, the overall evidential value of the evidence is measured as $\Pr(H|E_1, \ldots, E_k)$.

Overall and incremental evidential value should not be confused. To illustrate, suppose we have strong evidence E_1 for the hypothesis H that a suspect was at the crime scene, for instance, security camera footage in which the suspect is easily recognizable. In this case, the overall evidential value $\Pr(H|E_1)$ of the evidence is high. If this is the only evidence, then also the incremental evidential value is high: Before the evidence is considered, the hypothesis is not strongly supported, i.e., $\Pr(H)$ is low, whereas after the evidence is considered, the hypothesis is strongly supported, i.e., $\Pr(H|E_1)$ is high. In this case, the overall and incremental evidential value of E_1 are both high. But suppose a witness testifies that the defendant was not at the crime scene (evidence E_2), but as it turns out, the witness is unreliable as a known accomplice of the suspect. Consider now the overall evidential value $\Pr(H|E_1, E_2)$ of the two pieces of evidence together. This will not have changed much when compared to $\Pr(H|E_1)$. As a result, the incremental evidential value of E_2 is low, while still the overall evidential value $\Pr(H|E_1, E_2)$ is high, even though E_2 did not contribute much.

The difference between overall and incremental evidential value can be especially confusing when there is a single piece of evidence. Consider the hypothesis $\neg H$ that the suspect was not at the crime scene and the evidence E_2, the testimony of the unreliable witness. Now, if $\Pr(\neg H)$ is high, then $\Pr(\neg H|E_2)$ will be equally high because E_2 has no incremental value. Uncritically interpreted, the high value of $\Pr(\neg H|E_2)$ suggests that the testimony of the unreliable witness has a high evidential value. But incrementally E_2 did not change much. The hypothesis $\neg H$ is, in totality, still strongly supported after the incrementally weak evidence E_2, since the hypothesis was already strongly supported before that evidence.

The use of evidence with high incremental evidential value has complications. As an illustration, we discuss the likelihood ratio of a DNA match. When introduced in court, a DNA match comes with a so-called Random Match Probability or, as of late, with its Conditional Genotype Probability. This (roughly) is the probability that the DNA of a random person, who had nothing to do with the crime, would match. Let us denote this probability by γ.

With some simplifications (on these later), the evidential value of the DNA match M in favor of the hypothesis that the suspect is the source of the crime traces, abbreviated S, is as follows

$$\frac{\Pr(M|S)}{\Pr(M|\neg S)} = \frac{1}{\gamma}.$$

The numerator $\Pr(M|S)$ equals 1 because we assume that if the defendant is the source of the crime traces, the laboratory test will report a match. As for the denominator, assuming $\Pr(M|\neg S) = \gamma$ is plausible because the probability that a match would be reported if the defendant was *not* the source is roughly the same as the chance that someone who had no contact with the victim would randomly match. For example, if γ is 1 in 200 million, the likelihood ratio would be

$$\frac{\Pr(M|S)}{\Pr(M|\neg S)} = \frac{1}{\frac{1}{200\text{ million}}} = 200\text{ million.}$$

Since the likelihood ratio in question is a high number, the DNA match has strong evidential value in favor of the hypothesis that the suspect is the source. More generally, a low random match or genotype probability γ corresponds to a match with a rare profile and thus has a high evidential value.

Still, even with a low γ one should beware of the complications when using a DNA match in a criminal case. Consider the following non-equivalent hypotheses:

1. The *lab reports* that the defendant's genetic profile matches with the crime traces;
2. The defendant's genetic profile *truly matches* with the crime traces;
3. The defendant is the *source* of the traces;
4. The defendant *visited* the crime scene; and
5. The defendant is *factually guilty*.

The inferential path from "reported match" to "guilt," passing through the intermediate steps "true match," "source" and "visit," is a long one, and each step comes with sources of error that may undermine the inference along the way.

First, the inference from "reported match" to "true match" depends on whether or not the laboratory made a mistake. A key source of laboratory mistakes is human errors, much more frequent than DNA profiles. Second, the inference from "true match" to "source" can go wrong in several ways. For one, someone who is entirely unrelated with the crime could be, by sheer coincidence, a true match. This could happen, although the chance of this happening remains typically low as measured by the Random Match Probability or Conditional Genotype Probability. For another, a suspect who is not the source of the crime traces could still match because of close family relations with the actual perpetrator. Think of a perpetrator who has a genetically identical twin. Third, the inference from "source" to "visiting the crime scene" is not infallible. In particular, the traces can have been accidentally transferred to the crime scene or have been planted there. Fourth, the inference from "visiting the crime scene" to "factual guilt" can go wrong in many ways, because having visited a crime scene is not the same as having committed the crime.

4.2 Arguments

The evidential value of arguments can be analyzed in terms of the strength of the reasons they are built from, but also by asking critical questions about the reasons of the argument, its conclusion, and the connection between reasons and conclusion.

The reasons used can be conclusive or defeasible. A reason is conclusive when, given the reason, its conclusion is guaranteed. The main type of conclusive reason corresponds to deductive, logically valid reasoning. An example of a conclusive reason occurs in the logically valid argument from the reasons "John is shot" and "If someone is shot, he dies" to the conclusion "John dies." Its logical validity is connected to the underlying logical structure of the argument: from "A" and "A implies B" conclude "B."

Many reasons are not conclusive, but defeasible. There are circumstances in which the conclusion does not follow, although the reason obtains. The reason "The witness reports to have seen the suspect at the crime scene" supports the conclusion "The suspect was at the crime scene" but does not guarantee that conclusion, because the witness could have made a mistake. A defeasible reason can provide *prima facie* justification for a conclusion, which might later be withdrawn in light of countervailing reasons. Reasons that occur in so-called *abductive arguments* are also defeasible, where abductive arguments can be thought of as providing an explanation. For example, from "John's DNA matches the crime trace" conclude "John left the trace." The fact that John left the trace is put forward as an explanation for the fact that John's DNA matches the trace. Abductive arguments are typically defeasible because there often are alternative explanations. Someone with the same genetic profile as John might have left the trace.

Arguments can be evaluated by asking critical questions. Consider again the one-step argument from the reason "The witness reports that she saw the suspect at the crime scene" to the conclusion "The suspect was at the crime scene." Critical questions can be asked about the argument. They include, for example, whether there are reasons to doubt the suspect was at the crime scene, such as an alibi; whether there are reasons to doubt that the witness testimony supports the conclusion that the suspect was at the crime scene, for instance, the witness is lying; and whether there are reasons to doubt the existence of the witness testimony, such as a fraudulent report. The first of these questions is directed at the argument's conclusion, the second at the argument step from reason to conclusion, and the third at the argument's reason. These different kinds of critical questions are connected to the three kinds of argument attack discussed in Sect. 3.1 (see in particular Fig. 2, page 455).

But what do critical questions do? How do they help us assess the strength of arguments? Suppose that initially it is believed that the suspect was at the crime scene because of the witness testimony. A positive answer to any of the critical questions mentioned above will weaken the support for the conclusion that the suspect was at the crime scene, perhaps up to the point of making it no longer believable.

It can be subject to debate whether a reason supports or attacks a conclusion. Whether a reason supports a conclusion depends on an underlying general rule. For instance, the argument from a witness testimony (the reason) to the suspect's being at the crime scene (the conclusion) rests on the general rule that what witnesses say can generally be believed. Following Toulmin (1958)'s terminology, such general rules making explicit how to get from the reason to the conclusion are referred to as *warrants*. Support for a warrant is called the backing of the warrant.

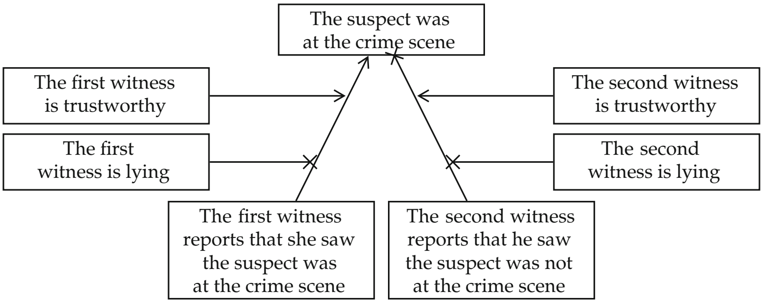

The suspect was at the crime scene
The first witness is trustworthy
The first witness is lying
The second witness is trustworthy
The second witness is lying
The first witness reports that she saw the suspect was at the crime scene
The second witness reports that he saw the suspect was not at the crime scene

Fig. 5 Arguments about whether a reason is supporting or attacking

More generally, a reason can either support or attack a conclusion, so the relation between reason and conclusion can be a supporting relation or an attacking relation. These supporting or attacking relations can, in turn, be themselves supported or attacked. This gives rise to four different combinations: support of a supporting relation; support of an attacking relation; attack of a supporting relation; and attack of an attacking relation. In Fig. 5, these situations are illustrated by two opposite witness testimonies.

4.3 Scenarios

The evidential value of a scenario depends on how well it matches up with the evidence. This matching up can be understood in three ways: the scenario's plausibility and logical consistency; its power to explain the evidence; its consistency with the evidence. We examine each in turn.

Scenarios can be plausible and logically consistent. Plausibility measures how well a scenario matches up with our background assumptions and knowledge of the world. At least, a scenario should not violate the laws of nature or commonsense. If a scenario asserts that the same individual was in two different locations at the same time, or moved from one location to another in too short amount of time, the scenario would lack plausibility. The scenario "an alien did it" lacks plausibility because it describes something that rarely happens. Lack of plausibility can become so pronounced that it amounts to a lack of *logical consistency*, for example, claiming that the defendant had and did *not* have a motive for killing the victim.

Recall now the two conflicting scenarios we considered earlier:

S_1: The defendant killed the victim when caught during a robbery.

S_2: The victim's partner killed the victim after a violent fight between the two.

Which one is the most plausible? Statistics suggest that people are less often killed by strangers than by people they know. If so, scenario S_2 would be initially more plausible. However, suppose we acquired more background information about the relationship between the victim and her partner, and it turned out their relationship was peaceful. In light of this new information, scenario S_2 will appear less plausible than S_1. It does not happen often that anger and violence manifest themselves unannounced, while it is natural that a robber, once he is discovered and has no alternative, will resort to violence.

In assessing plausibility, the evidence with which the scenario is expected to match up is not the evidence specific to the case, but rather, background information about the world. Plausibility has something to do with what we might call *normality*, that is, with what happens most of the time. It is true, however, that criminal cases are often about odd coincidences, unexpected and improbable events. Plausibility only measures the persuasiveness or credibility of a scenario *prior to* considering any more specific evidence about the crime. An initially plausible scenario may turn out to be weakly supported in light of more evidence presented about the crime.

The more evidence a scenario can explain, the better. When a case comprises several items of evidence, the more items of evidence a scenario can accommodate, preferably from both the prosecutor and the defense, the better the scenario. This depends on the scenario's explanatory power and consistency with the evidence.

Consider a case in which two items of evidence must be explained. The first is the presence of fingerprint traces at the crime scene, traces whose presence is consistent with just innocent contact. The second is that the fingerprints match with the defendant. Scenarios S_1 and S_2—the robber scenario and the victim's partner scenario, respectively—both explain the presence of fingerprint traces at the crime. They were left either by the robber, if S_1 is true, or by the victim's partner, if S_2 is true. Still, only S_1 can explain the fact that the traces match with the defendant (who is the alleged robber).

But suppose that in order to defend S_2—the victim's partner scenario—a new detail is added to the story: the victim's partner, right after killing the victim and with the intent to mislead the investigators, implanted fingerprint traces that match the defendant. This new scenario, however implausible, can explain both items of evidence: the presence of the fingerprint and the fact that they match the defendants. As far as explanatory power goes, scenarios S_1 and S_2, when properly supplemented, are now on a par with one another. Still, further evidence may distinguish the two. For example, if a witness testified she saw the defendant/robber walk toward the location of the crime immediately before the crime was committed, scenario S_2 cannot easily explain the testimony, even when supplemented with additional information. By contrast, S_1 can easily explain the testimony. Absent other evidence, scenario S_1 explains more evidence than the competing scenario S_2, in both its original and updated versions.

The more pieces of evidence a scenario is consistent with, the better. Besides plausibility and explanatory power, we can evaluate a scenario by checking whether it is consistent with the evidence presented in a case. The more pieces of evidence the scenario is consisted with, the better.

We can define consistency as lack of inconsistency between the evidence (taken at face value) and the scenario. For example, if a witness testifies that the defendant was at home with his girlfriend at 6 PM, while according to the scenario proposed by the prosecutor, the defendant was at the crime scene at 6 PM, the two are inconsistent. Here, we are dealing we what we earlier called quasi-inconsistency, in the sense that insofar as the evidence is taken at face value—that is, the witness is taken to be truthful—the scenario is inconsistent with the evidence. An inconsistency in this sense between the evidence and a proposed scenario need not be damning for the scenario. It might, in fact, turn out that the witness was untruthful or simply confused about the timing. If so, the evidence will be discarded, not the scenario.

But, if a scenario is inconsistent with several pieces of evidence, this becomes an increasingly powerful challenge against the scenario. For example, if the timing provided by the scenario is not only inconsistent with the first witness testimony but also with the testimony of a pizza delivery man, who claims to have delivered a pizza to the house of the defendant's girlfriend, around 6 PM, and remembers having received money from the defendant, then the prosecutor's scenario is further undermined. In short, the more pieces of evidence inconsistent with the scenario, the more powerful the challenges against the scenario. This conclusion can also be stated more positively. The more pieces of evidence consistent with the scenario, the higher the evidential value of the scenario.

5 Coherently Interpreting the Evidence

The dossiers of criminal cases can be large, and the coherent interpretation of the evidence in such a dossier can be daunting, whichever normative framework is used. For each framework, we discuss how the coherent interpretation of the evidence can be addressed.

5.1 Scenarios

Scenarios can provide coherent interpretations that make sense of the evidence. We examine three dimensions along which competing scenarios, considered holistically in their entirety, can be assessed: coherence; completeness; and explanatory power.

Scenarios are coherent clusters of events, ordered in time and with causal relations. Earlier we encountered an elementary murder case scenario; namely, a robber kills the victim when caught during a robbery (S_1). This scenario can be analyzed as having a specific temporal structure: First, the robber enters the victim's house (H_1); then, the victim accidentally encounters the robber (H_2); and finally, the robber kills the victim (H_3).

Some of the events in this temporally ordered scenario are also causally connected. The accidental encounter is the cause that triggers a reaction in the robber who then kills the victim. Causal relations among the different parts, or episodes, in a scenario are important to evaluate what we might call the *coherence* of a scenario.

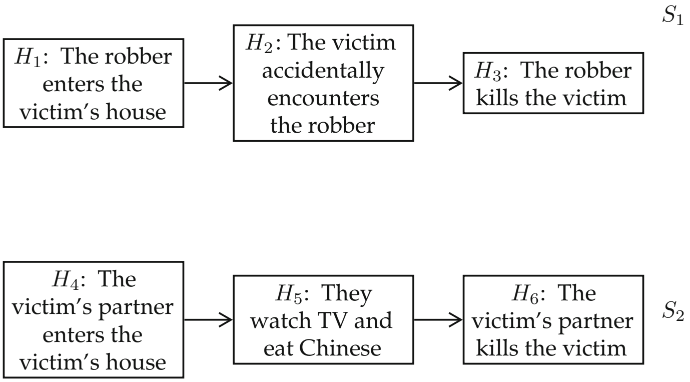

S_1
H_1: The robber enters the victim's house
H_2: The victim accidentally encounters the robber
H_3: The robber kills the victim
H_4: The victim's partner enters the victim's house
H_5: They watch TV and eat Chinese
H_6: The victim's partner kills the victim
S_2

Fig. 6 Scenarios and their structure. The second scenario lacks in causal structure

Compare the robbery scenario S_1 with the partner scenario S_2 we countered earlier. Suppose this scenario is articulated more in detail as follows: The victim and her partner were watching a show on TV and eating Chinese takeout, when the partner killed the victim. This scenario has a clear temporal structure: The victim's partner arrives at the victim's home (H_4); then, they watch TV while eating Chinese takeout (H_5); finally, the victim's partner kills the victim (H_6). Still, something is missing here, that is, the causal link between "watching TV" and "killing." Why would peacefully watching TV suddenly turn into fatal violence? In comparison, the first scenario scores better in terms of causal structure. The first scenario is more coherent than the second (Fig. 6).

Scenarios can be more or less complete. Another criterion to evaluate scenarios is their *completeness*. Since scenarios are discursive arrangements of events, ordered according to temporal and causal relations, they may contain gaps in time, space, and causality. A scenario may not describe the defendant's whereabouts between 4 and 6 PM, while it describes, rather precisely, what the defendant did at 7 PM, immediately before the killing took place. The temporal gap between 4 and 6 PM makes it less complete than a scenario which describes the defendant's whereabouts between 4 and 7 PM without gaps. Yet, this might not be the notion of completeness that is important here to evaluate scenarios.

The law is not very specific in this respect. Besides defining the crime and requiring that both *mens rea*—the intention to do harm—and *actus reus*—the occurrence of the physical harm—be established, the law does not say how detailed the prosecutor's reconstruction of the crime should be. So, how is completeness a criterion to evaluate a scenario?

Some suggest that scenarios must follow certain patterns, schematic structures, or scripts. For example, in most violent crimes, we can identify an initial moment of conflict. This triggers a psychological reaction that gives rise to the formation of an intention, which, in turn, later results in the violent act. On this account, a scenario is complete whenever it has *all of its parts*, at least given an appropriate scenario script or schematic structure. Scenario S_2, in this sense, is incomplete because it does say why and how the victim's partner formed the intention to kill the victim nor does it describe any initial moment of conflict. Scenario S_1 does not say, exactly, why the robber killed the victim. But the reason can be easily inferred. Presumably, the robber formed the intention to kill the victim when he was caught by surprise and saw no better alternative.

Weaker scenarios can be better supported by the evidence. The coherence and completeness of a scenario play a role in its evaluation. However, a weaker scenario in terms of coherence and completeness may be the best explanation of the evidence. Earlier we saw that the robbery scenario was more coherent than the scenario in which the victim's friend kills the victim while eating Chinese takeout in front of the TV. But now suppose that the pieces of evidence are as follows: The investigators find Chinese takeout in the victim's house (E_1); the saliva on one fork matches with the victim's friend DNA (E_2); there are no signs of forced entry into the victim's house (E_3). While the robbery scenario was more coherent, the Chinese takeout scenario explains the three items of evidence. In fact, the robbery scenario cannot explain any of them. So, a scenario might be superior to another on one dimension, for example, the robbery scenario is more coherent than the Chinese takeout scenario, but inferior on another dimension, for example, the robbery scenario has less explanatory power than the Chinese takeout scenario.

5.2 Arguments

An analysis of a case in terms of arguments can become complex. When Wigmore (1913) developed his charting method for analyzing the evidence in a criminal case, he was well aware of this complexity. Figure 7 provides a Wigmore diagram of the murder case *Commonwealth v. Umilian* (1901). The diagram for this relatively simple case already contains some two dozen nodes. Diagrams for more complex cases contain many more nodes.

Here, we describe three sources of complexity in the analysis of cases from the argumentation perspective: arguments and subarguments; attacks, counterattacks, and chains of attacks; and conflicts between reasons and their resolution.

The evaluation of an argument can depend on its subarguments. The structure of a complex argument influences its evaluation, and in particular, the subarguments of a larger argument determine the evaluation of the whole. For example, consider the argument in Fig. 4 (page 456). This can be analyzed as consisting of two subarguments. The first is that the witness testimony supports the intermediate conclusion that the suspect was at the crime scene. This intermediate conclusion, in turn, supports the conclusion that the suspect committed the crime, and this is the second subargument. If the first subargument is successfully attacked by a counterargument alleging that the witness is lying, the subargument supporting the intermediate conclusion that the suspect was at the crime scene breaks down and its conclusion no

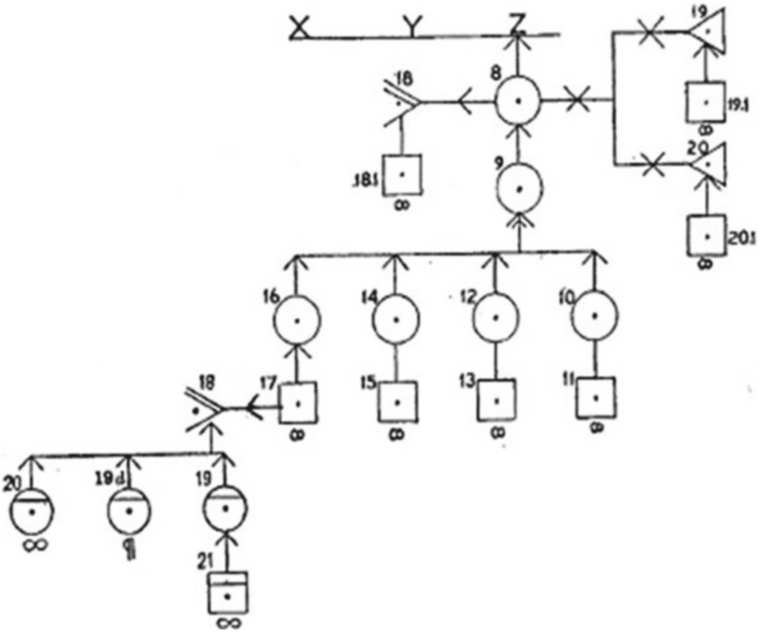


Fig. 7 A Wigmore chart

longer follows. Since the subargument does not successfully support the intermediate conclusion, also the larger argument for the final conclusion does not successfully support its conclusion.

The evaluation of an argument can depend on chains of attacks. Besides the modular relationship between arguments and subarguments, a further source of complexity in the analysis of arguments is that attacks can be chained. An attack against an argument can be countered by a further attack. When an attack is countered by a further attack, the original argument can be reinstated, in the sense that it again successfully supports its conclusion. Fig. 8 shows an example. A first witness, witness A, testifies that the suspect was at the crime scene. This testimony successfully supports the conclusion that the suspect was at the crime scene. However, suppose a second witness, witness B, claims that A is lying. The claim by witness B attacks the original argument, and thus, the conclusion that the suspect was at the crime scene is no longer supported. But if a third witness, witness C, claims that B is lying, the attack by B against A is countered. Witness B is no longer believable, so there is no reason to conclude that A is lying. As a result, A's testimony can again support the conclusion that the suspect was at the crime scene, and the original argument is thus reinstated.

Conflicts between reasons can be addressed by exceptions, preferences, and weighing. Another source of complexity in the analysis of arguments is conflicts between reasons. Reasons may support a certain conclusion or oppose it. When the same conclusion is supported by a reason (or set of reasons) and opposed by another

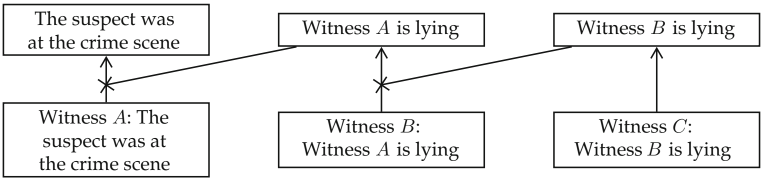

The suspect was at the crime scene
Witness A is lying
Witness B is lying
Witness A: The suspect was at the crime scene
Witness B: Witness A is lying
Witness C: Witness B is lying

Fig. 8 Reinstatement

reason (or set of reasons), different reasons come into conflict. Since a conclusion cannot be both supported and opposed, the question arises of how to address and resolve conflicts between reasons. We distinguish three ways in which conflicts between reasons can be addressed.

First, if a reason supports a conclusion and another reason opposes it, the conflict between them can be resolved if an exception exists which excludes one of the reasons. For instance, suppose two witnesses, A and B, make conflicting statements about whether the suspect was at the crime scene. If there is evidence that one witness is lying, the conflict is resolved in favor of the witness against whom there is no evidence of lying. If there is evidence that witness B is lying, the conflict is resolved in favor of A's testimony (see the top of Fig. 9). Evidence that B is lying undercuts the connection between B's testimony and its conclusion (cf. Sect. 3.1). This, by contrast, leaves intact the connection between A's testimony and its conclusion. The conflict is thus resolved in favor of A's testimony.

In the second way of addressing a conflict between reasons, there is again a reason for a conclusion and a reason against the same conclusion. This time the resolution of the conflict occurs not by excluding one of the two reasons, but rather, by giving preference to one reason over the other. If two witnesses give conflicting testimonies about whether the suspect was at the crime scene, the conflict will remain unresolved insofar as the two reasons oppose one another and are assigned equal weight. Taking into account further information that can justify preferring one reason over the other will resolve the conflict. A reason can be preferred over another, for instance, when it is stronger. A preference (indicated by the $>$-sign in Fig. 9) can be justified if one witnesses is shown to be more reliable than the other. In this case, the conclusion that follows from the testimony by the more reliable witness should be drawn, while the conclusion that follows from the other, weaker testimony should not be drawn. This resolves the conflict.

The third way of addressing conflicts between reasons involves more than two conflicting reasons. For instance, there can be more than two witnesses, offering conflicting testimonies (Fig. 9, bottom). Resolving such conflicts can be thought of as weighing the reasons involved, where the weighing of reasons is done by generalizing a preference ordering of reasons to an ordering of sets of conflicting reasons.

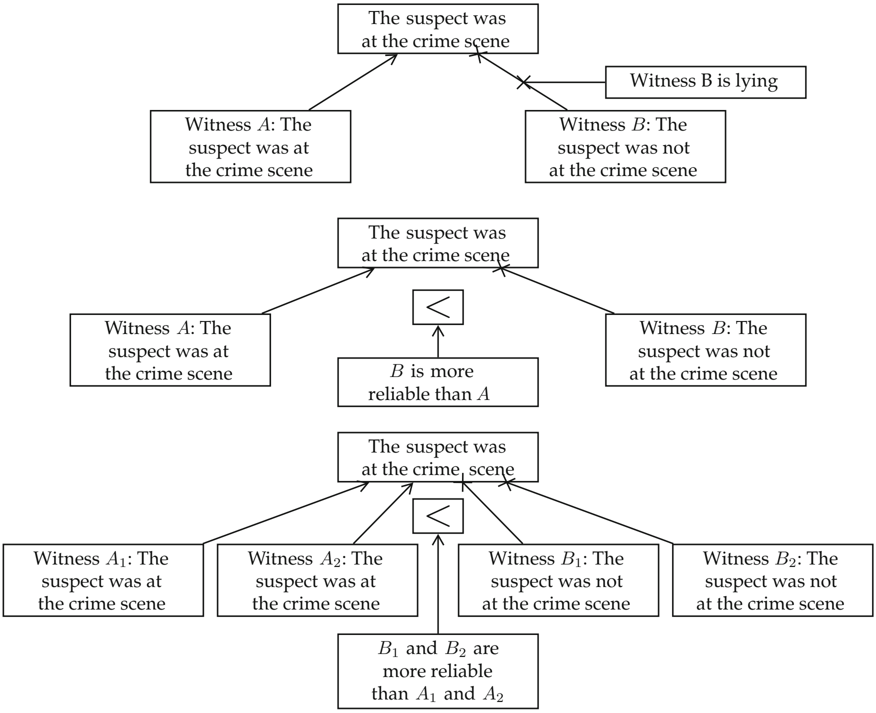

The suspect was at the crime scene
Witness B is lying
Witness A: The suspect was at the crime scene
Witness B: The suspect was not at the crime scene
The suspect was at the crime scene
<
Witness A: The suspect was at the crime scene
B is more reliable than A
Witness B: The suspect was not at the crime scene
The suspect was at the crime scene
<
Witness A_1: The suspect was at the crime scene
Witness A_2: The suspect was at the crime scene
Witness B_1: The suspect was not at the crime scene
Witness B_2: The suspect was not at the crime scene
B_1 and B_2 are more reliable than A_1 and A_2

Fig. 9 Three kinds of addressing conflicts of reasons. The sign $<$ indicates a preference ordering between one reason (or set of reasons) and another

5.3 *Probability*

The probability calculus provides formal rules for the coherent interpretation of the evidence. We begin by discussing an elementary application of Bayes' theorem when only one piece of evidence is involved, and then turn to an analysis that involves more than one piece of evidence. We briefly discuss Bayesian networks, formal and computational tools for handling complex bodies of evidence in the probabilistic setting.

The likelihood ratio formula shows how to find the posterior odds given the evidence. The odds of a hypothesis H are given by the ratio $\Pr(H)/\Pr(\neg H)$ of the probability of the hypothesis and the probability of its negation. The odds $\Pr(H)/\Pr(\neg H)$ of the hypothesis, unconditioned on the evidence, are called the *prior odds* of the hypothesis, and the odds $\Pr(H|E)/\Pr(\neg H|E)$ of the hypothesis, conditioned on evidence E, are called the *posterior odds*. The latter can be found by multiplying the prior odds with the likelihood ratio[9]:

[9] To derive the likelihood ratio formula, one first applies Bayes' theorem to both H and $\neg H$. We get $\Pr(H|E) = \Pr(E|H) \cdot \Pr(H)/\Pr(E)$ and $\Pr(\neg H|E) = \Pr(E|\neg H) \cdot \Pr(\neg H)/\Pr(E)$. Using these,

$$\frac{\Pr(H|E)}{\Pr(\neg H|E)} = \frac{\Pr(E|H)}{\Pr(E|\neg H)} \cdot \frac{\Pr(H)}{\Pr(\neg H)}.$$

This formula shows that the (incremental) evidential value of the evidence for a hypothesis, expressed by the likelihood ratio $\frac{\Pr(E|H)}{\Pr(E|\neg H)}$, does not by itself give the posterior odds. The prior odds are needed as well. If the posterior odds $\frac{\Pr(E|H)}{\Pr(E|\neg H)}$ are known, the *posterior probability* $\Pr(H|E)$ can be derived by applying the following formula[10]:

$$\Pr(H|E) = \frac{\frac{\Pr(H|E)}{\Pr(\neg H|E)}}{1 + \frac{\Pr(H|E)}{\Pr(\neg H|E)}}.$$

Consider an example. The incremental evidential value of a DNA match, call it M, relative to the hypothesis that the defendant is factually guilty, call it G, is given by the likelihood ratio $\frac{\Pr(M|G)}{\Pr(M|\neg G)}$. Suppose this ratio is assigned a numerical value, as follows:

$$\frac{\Pr(M|G)}{\Pr(M|\neg G)} = \frac{1}{\frac{1}{2{,}000{,}000}} = 2{,}000{,}000.$$

Suppose, also, that the prior odds are as follows:

$$\frac{\Pr(G)}{\Pr(\neg G)} = \frac{\frac{1}{200{,}000}}{\frac{199{,}999}{200{,}000}} \approx \frac{1}{200{,}000}.$$

The posterior odds of the hypothesis given the match $\frac{\Pr(G|M)}{\Pr(\neg G|M)}$ are therefore as follows:

$$\frac{\Pr(G|M)}{\Pr(\neg G|M)} = \frac{\Pr(M|G)}{\Pr(M|\neg G)} \cdot \frac{\Pr(G)}{\Pr(\neg G)} \approx 2{,}000{,}000 \cdot \frac{1}{200{,}000} = 20.$$

So the poster probability of the hypothesis is as follows:

$$\Pr(G|M) = \frac{\frac{\Pr(G|M)}{\Pr(\neg G|M)}}{1 + \frac{\Pr(G|M)}{\Pr(\neg G|M)}} \approx \frac{20}{1+20} \approx 95\%.$$

A generalization of the formula shows how to handle more pieces of evidence. So far we have considered only one piece of evidence. In a straightforward generalization

we find:

$$\frac{\Pr(H|E)}{\Pr(\neg H|E)} = \frac{\Pr(E|H) \cdot \Pr(H)/\Pr(E)}{\Pr(E|\neg H) \cdot \Pr(\neg H)/\Pr(E)} = \frac{\Pr(E|H) \cdot \Pr(H)}{\Pr(E|\neg H) \cdot \Pr(\neg H)},$$

proving the likelihood ratio formula.

[10] $\Pr(H|E) = \frac{\Pr(H|E)}{\Pr(H|E)+\Pr(\neg H|E)} = \frac{\frac{\Pr(H|E)}{\Pr(\neg H|E)}}{\frac{\Pr(H|E)+\Pr(\neg H|E)}{\Pr(\neg H|E)}} = \frac{\frac{\Pr(H|E)}{\Pr(\neg H|E)}}{\frac{\Pr(H|E)}{\Pr(\neg H|E)}+1}.$

of the formula for two pieces of evidence E_1 and E_2, the likelihood ratios of the individual pieces of evidence are multiplied, as follows:

$$\frac{\Pr(H|E_1 \wedge E_2)}{\Pr(\neg H|E_1 \wedge E_2)} = \frac{\Pr(E_2|H)}{\Pr(E_2|\neg H)} \cdot \frac{\Pr(E_1|H)}{\Pr(E_1|\neg H)} \cdot \frac{\Pr(H)}{\Pr(\neg H)}.$$

However, this generalization only holds provided that the two pieces of evidence are independent conditional on the hypothesis, that is, $\Pr(E_2|H) = P(E_2|H \wedge E_1)$.

To illustrate, consider now two pieces of evidence: a DNA match and a witness testimony. The DNA match, call it M, holds between the crime traces and the defendant, and the witness, call it W, in her testimony asserts that the defendant was seen at the crime scene. Both pieces of evidence, intuitively, support the hypothesis G that the defendant is factually guilty. To assign an explicit numerical value, assume the DNA match has a likelihood ratio $\frac{\Pr(M|G)}{\Pr(M|\neg G)}$ of 2 million, and the witness testimony a likelihood ratio $\frac{\Pr(W|G)}{\Pr(W|\neg G)}$ of 1,000. These numbers are purely illustrative, but are needed to perform the probabilistic calculations. (Of course, there remains the question of how the numbers can be obtained and whether the numbers needed to carry out the calculations are available in the first place. This is a topic of debate.) Finally, assume the two pieces are independent conditional on the hypothesis G, that is, $\Pr(W|G) = \Pr(W|G \wedge M)$.

The combined (incremental) evidential value of the two pieces of evidence is given by multiplying the two likelihood ratios, that is, $2{,}000{,}000 \times 1{,}000 = 2{,}000{,}000{,}000$, which is a higher value than the two pieces considered independently. If the prior odds $\frac{\Pr(G)}{\Pr(\neg G)}$ are roughly $\frac{1}{200{,}000}$ as before, the posterior odds are therefore as follows:

$$\frac{\Pr(G|M \wedge W)}{\Pr(\neg G|M \wedge W)} = \frac{\Pr(M|G)}{\Pr(M|\neg G)} \cdot \frac{\Pr(W|G)}{\Pr(W|\neg G)} \cdot \frac{\Pr(G)}{\Pr(\neg G)} \approx 2{,}000{,}000 \cdot 1{,}000 \cdot \frac{1}{200{,}000} = 20{,}000.$$

So the posterior probability of the hypothesis given the two pieces of evidence is as follows:

$$\Pr(G|M \wedge W) \approx \frac{20{,}000}{1 + 20{,}000} \approx 99\%.$$

Compare this probability with $\Pr(G|M)$, which was 95%, a lower value. The probability calculus can offer a numerical representation of the intuitive fact that two pieces of evidence, taken together, have a higher (overall) evidential value than one piece alone.

If the two pieces of evidence are not independent, the likelihood ratio formula for two pieces of evidence takes the following, more general form:

$$\frac{\Pr(H|E_1 \wedge E_2)}{\Pr(\neg H|E_1 \wedge E_2)} = \frac{\Pr(E_2|H \wedge E_1)}{\Pr(E_2|\neg H \wedge E_1)} \cdot \frac{\Pr(E_1|H)}{\Pr(E_1|\neg H)} \cdot \frac{\Pr(H)}{\Pr(\neg H)}.$$

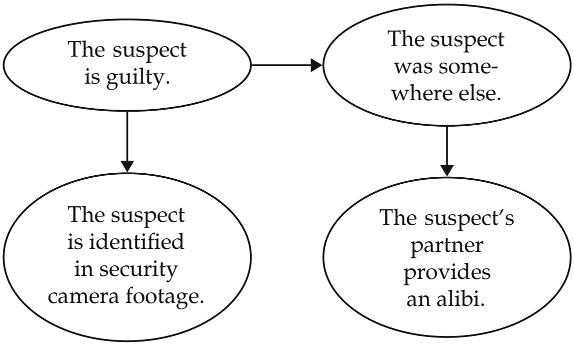

The suspect is guilty.
The suspect was somewhere else.
The suspect is identified in security camera footage.
The suspect's partner provides an alibi.

Fig. 10 An example of a Bayesian network: directed acyclic graph

The first generalization follows from the second, assuming independence between the two pieces of evidence conditional on the hypothesis of interest. The first generalization does not always hold because the evidential value of a piece of evidence, as measured by the likelihood ratio, can change in the face of other evidence.

More complex analytic tools can be used, in particular Bayesian networks. Probabilistic analyses become more complex when more elements are involved, and calculations quickly become unmanageable without appropriate modeling tools. We discuss Bayesian networks, a prominent example of a modeling tool that allows for complex probabilistic representations and calculations. A great strength of Bayesian networks is that, once the network structure is in place and the probabilities are assigned, many computer programs exist that can perform all calculations automatically.

A Bayesian network has a graphical and a numeric part. The graphical part is a directed, acyclic graph of the relevant probabilistic variables. Each node in the graph represents a variable—a hypothesis or a piece of evidence—that can take the value "true" or "false." Consider, for instance, the Bayesian network in Fig. 10. At the top, two hypotheses are shown. One is that the suspect is factually guilty (abbreviated, "Guilt"), and the other hypothesis is that the suspect was somewhere else when the crime was committed (abbreviated, "Elsewhere"). At the bottom, two pieces of evidence are shown: security camera footage that incriminates the suspect (abbreviated, "Camera"), and the testimony of the suspect's partner who provides an alibi (abbreviated, "Partner"). The arrows connecting the nodes indicate dependency relations. The "Elsewhere" variable has the "Guilt" variable as a parent, and each evidence node has a hypothesis as parent.

Some might wonder how a Bayesian network should be constructed. This is not a trivial task, and there can be different Bayesian network models of the same case. For example, some might prefer to the Bayesian network in Fig. 10 a network whose nodes are arranged in a slightly different way, or using different variables. The construction of a Bayesian network for the purpose of evidence assessment is ultimately an art.

Table 1 An example of a Bayesian network: conditional probability tables

Guilt	Pr(Guilt)
False	0.999
True	0.001

Guilt	Elsewhere	Pr(Elsewhere|Guilt)
False	False	0
False	True	1
True	False	1
True	True	0

Guilt	Camera	Pr(Camera|Guilt)
False	False	0.99
False	True	0.01
True	False	0.3
True	True	0.7

Elsewhere	Partner	Pr(Partner|Elsewhere)
False	False	0.9
False	True	0.1
True	False	0
True	True	1

Turning now to the numeric part of a Bayesian network, this consists of conditional probability tables. In these tables, the conditional probabilities of each variable are specified, conditioned on the different values of the parent variables. Table 1, for example, contains four tables of conditional probabilities, one per variable. Consider first the probability assignments in the top two tables. The prior probability of factual guilt is set to an arbitrary low value, 0.1%, or 1 in a 1000 (top left table). Since the two hypotheses—"guilt" and "elsewhere"—are assumed to exclude one another, it can never occur that they are both true or both false (top right table). Consider now the two bottom tables. The bottom left table shows how the camera identification depends on the suspect's factual guilt. If the suspect is factually guilty, there is a 70% probability that the suspect is identified in the security camera footage. If the suspect is not factually guilty, this probability is much lower, 1%. The likelihood ratio $\Pr(\text{Camera}|\text{Guilt})/\Pr(\text{Camera}|\neg\text{Guilt})$ expressing the strength of the camera footage evidence relative to hypotheses "guilt" and "not-guilt" is therefore 70. The bottom right table shows how the partner's alibi testimony depends on whether the suspect was in fact elsewhere when the crime was committed. If the "Elsewhere" hypothesis is true, the partner will surely provide an alibi in favor of the defendant. If the "Elsewhere" hypothesis is false, the table specifies that there is a 10% chance that the suspect's partner will still provide an alibi in favor of the defendant. The likelihood ratio $\Pr(\text{Partner}|\text{Elsewhere})/\Pr(\text{Partner}|\neg\text{Elsewhere})$ expressing the strength of the evidence (in this case, the partner's testimony providing an alibi) relative to hypotheses "Elsewhere" and "not-Elsewhere" is therefore 10.

The question now is how to combine the two pieces of evidence, the incriminating camera footage and the exculpatory partner's alibi testimony, and thus how to assess the probability of the defendant's guilt. Multiplying the likelihood ratios for each piece of evidence will not do. Bayesian networks can help here. For illustrative purposes, we write down the somewhat tedious calculations, although there is software that can do them automatically. Now, if the goal is to calculate $\Pr(\text{Guilt}|\text{Partner} \wedge \text{Camera})$—that is, the conditional probability of guilt given both the camera footage and the partner's alibi testimony—we need to calculate both Pr(Partner $\wedge$ Camera) and Pr(Guilt $\wedge$ Partner $\wedge$ Camera).

Pr(Partner $\wedge$ Camera) can be broken down as the sum of four probabilities. That is,

$$\begin{aligned}\mathrm{Pr}(\text{Partner} \wedge \text{Camera}) = {} & \mathrm{Pr}(\text{Partner} \wedge \text{Camera} \wedge \text{Elsewhere} \wedge \text{Guilt})\\ & + \mathrm{Pr}(\text{Partner} \wedge \text{Camera} \wedge \text{Elsewhere} \wedge \neg\text{Guilt})\\ & + \mathrm{Pr}(\text{Partner} \wedge \text{Camera} \wedge \neg\text{Elsewhere} \wedge \text{Guilt})\\ & + \mathrm{Pr}(\text{Partner} \wedge \text{Camera} \wedge \neg\text{Elsewhere} \wedge \neg\text{Guilt})\end{aligned}$$

Let us calculate each in turn. First, we have:

Pr(Partner $\wedge$ Camera $\wedge$ Elsewhere $\wedge$ Guilt)

= Pr(Partner|Camera $\wedge$ Elsewhere $\wedge$ Guilt) Pr(Camera|Elsewhere $\wedge$ Guilt) Pr(Elsewhere|Guilt) Pr(Guilt)
= Pr(Partner|Elsewhere) Pr(Camera|Guilt) Pr(Elsewhere|Guilt) Pr(Guilt)
= $1 \times 0.7 \times 0 \times 0.001 = 0$

Note that we relied on the equalities

Pr(Partner|Camera $\wedge$ Elsewhere $\wedge$ Guilt) = Pr(Partner $\wedge$ Elsewhere); and
Pr(Camera|Elsewhere $\wedge$ Guilt) = Pr(Camera $\wedge$ Guilt).

They hold for the Bayesian network because the node representing the partner's testimony has only the "Elsewhere" variable node as a parent, and the "Camera" node has only the "Guilt" node as a parent. Given the Bayesian network, the "Partner" variable does not depend on the "Camera" or "Guilt" variable, and the "Camera" variable does not depend on the "Elsewhere" variable.

Second, we have:

Pr(Partner $\wedge$ Camera $\wedge$ Elsewhere $\wedge\neg$ Guilt)

= Pr(Partner|Camera $\wedge$ Elsewhere $\wedge$ $\neg$Guilt) Pr(Camera|Elsewhere $\wedge$ $\neg$Guilt) Pr(Elsewhere|$\neg$Guilt) Pr($\neg$Guilt)
= Pr(Partner|Elsewhere) Pr(Camera|$\neg$Guilt) Pr(Elsewhere |$\neg$Guilt) Pr($\neg$Guilt)
= $1 \times 0.01 \times 1 \times 0.999 = 0.00999$

Note that we relied on the equalities

Pr(Partner|Camera $\wedge$ Elsewhere $\wedge$ $\neg$Guilt) = Pr(Partner|Elsewhere); and
Pr(Camera|Elsewhere $\wedge$ $\neg$Guilt) = Pr(Camera|$\neg$Guilt).

They hold, as before, because of the independence relations among variables in the Bayesian network.

Third, we have:

Pr(Partner $\wedge$ Camera $\wedge$ $\neg$Elsewhere $\wedge$ Guilt)

$$= \mathrm{Pr}(\mathrm{Partner}|\mathrm{Camera} \wedge \neg\mathrm{Elsewhere} \wedge \mathrm{Guilt})\ \mathrm{Pr}(\mathrm{Camera}|\neg\mathrm{Elsewhere} \wedge \mathrm{Guilt})\ \mathrm{Pr}(\neg\mathrm{Elsewhere}|\mathrm{Guilt})\ \mathrm{Pr}(\mathrm{Guilt})$$

$$= \mathrm{Pr}(\mathrm{Partner}|\neg\mathrm{Elsewhere})\ \mathrm{Pr}(\mathrm{Camera}|\mathrm{Guilt})\ \mathrm{Pr}(\neg\mathrm{Elsewhere}|\mathrm{Guilt})\ \mathrm{Pr}(\mathrm{Guilt})$$

$$= 0.1 \times 0.7 \times 1 \times 0.001 = 0.00007$$

Once again, we relied on independence relations. Fourth, since “Guilt” and “Elsewhere” cannot both be true, we have:

$$\mathrm{Pr}(\mathrm{Partner} \wedge \mathrm{Camera} \wedge \neg\mathrm{Elsewhere} \wedge \neg\mathrm{Guilt}) = 0$$

By adding these four probabilities, we find:

$$\mathrm{Pr}(\mathrm{Partner} \wedge \mathrm{Camera}) = 0 + 0.00999 + 0.00007 + 0 = 0.01006$$

Finally, we can compute the guilt probability based on the camera footage and the partner’s testimony, as follows:

$$\begin{aligned}\mathrm{Pr}(\mathrm{Guilt} \mid \mathrm{Partner} \wedge \mathrm{Camera}) &= \mathrm{Pr}(\neg\mathrm{Elsewhere} \wedge \mathrm{Guilt} \mid \mathrm{Partner} \wedge \mathrm{Camera})\\ &= \frac{\mathrm{Pr}(\mathrm{Partner} \wedge \mathrm{Camera} \wedge \neg\mathrm{Elsewhere} \wedge \mathrm{Guilt})}{\mathrm{Pr}(\mathrm{Partner} \wedge \mathrm{Camera})}\\ &= 0.00007/0.01006 \approx 0.7\%\end{aligned}$$

The 0.7% probability is low, but recall that the prior guilt probability was lower, 0.1%. Also note that Pr(Guilt | Camera) equals (roughly) 6.5%.[11] This is a low probability, but still greater than 0.7 or 0.1%. This shows that the camera evidence by itself has incriminating evidential value, as it raises the guilt probability from 0.1 to 6.5%. However, when the partner’s alibi testimony is added as evidence, the support for the guilt hypothesis is weakened to 0.7%.

Finally, a note about complexity. For probability functions of many variables that have many dependencies, a Bayesian network representation can be significantly more compact than a general full probability distribution over the variables. Our example has four variables, in general requiring 15 numbers to specify the full distribution. Given the independencies in the network, we only need 7 (note that half of the 14 numbers in Table 1 are superfluous as they follow from the other half).

6 Reasoning and Decision Making

So far we have focused on how the evidence can be evaluated and combined. But once the evidence has been introduced at trial, examined and cross-examined, it comes a time when the fact finders, either a trained judge or a group of lay jurors, must reason from the evidence, reach a conclusion, and decide whether to convict or acquit the

[11] Since $\frac{\mathrm{Pr(Guilt)}}{\mathrm{Pr}(\neg\mathrm{Guilt})} = \frac{0.001}{0.999}$ and $\frac{\mathrm{Pr(Camera|Guilt)}}{\mathrm{Pr}(\mathrm{Camera}|\neg\mathrm{Guilt})} = 70$, by Bayes’ theorem, $\frac{\mathrm{Pr(Guilt|Camera)}}{\mathrm{Pr}(\neg\mathrm{Guilt|Camera})} = \frac{0.001}{0.999} \times 70 \approx 0.07$ and thus Pr(Guilt|Camera)$\approx \frac{0.07}{1+0.07} \sim 6.5\%$.

defendant. The decision criterion is defined by law and consists of a standard of proof, sometimes also called burden of persuasion. If the decision makers are persuaded of the defendant's guilt beyond a reasonable doubt, they should convict, or else they should acquit.

Paraphrases of the formulation "proof beyond a reasonable doubt" abound in the case law. Yet, it is unclear whether they improve our understanding. The US Supreme Court might have been right when, in Holland v. United States, 348 U.S. 121 (1954), it wrote that that "attempts to explain the term 'reasonable doubt' do not result in making it any clearer" (140). The three frameworks we considered—probability, arguments, and scenarios—can be supplemented by a decision-theoretic layer and then used to characterize the standard of proof, although they are not immune from shortcomings, as we shall soon see.

6.1 Probability

In a probabilistic treatment, reasoning and decision making are analyzed using the probability calculus combined with elements of decision theory.

The guilt probability is assessed by weighing the evidence with the probability calculus. On the probabilistic framework, the goal is to assess the probability of the defendant's guilt based on all the available evidence. The assessment begins with a relatively low value for the guilt probability, prior to considering any evidence. After all, absent any incriminating evidence, the prior probability that an individual committed a crime should be rather low. As more evidence is presented, the guilt probability moves upward or downward depending on whether the evidence is incriminating or exculpatory. When all the evidence is considered, a final guilt probability value is reached. This forms the basis for the decision to convict or acquit.

The value of the guilt probability is arrived at by applying Bayes' theorem a repeated number of times and by plugging the values of the probabilities that are needed. Sometimes, these probabilities can be based on numerical data about population proportions, as in the case of DNA evidence, but often data are not available. For example, $\Pr(G)$ is required to calculate $\Pr(G|E)$ using Bayes' theorem, where $\Pr(G)$ is the probability of the defendant's guilt regardless of the evidence presented at trial. What probability value should be assigned to $\Pr(G)$? It is subject to debate how to assess this probability, and even whether it makes sense to assign a number to the prior probability of guilt in the first place.

The decision criterion is a guilt probability threshold. In probabilistic terms, proof of guilt beyond a reasonable doubt means that the defendant's *probability of guilt*, given the evidence presented at trial, meets a threshold, say >99 or $>99.9\%$. A numerical value for the threshold can be identified using expected utility theory. Let $c(CI)$ be the cost of convicting an innocent and $c(AG)$ the cost of acquitting a guilty defendant. For a conviction to be justified, the expected cost of convicting an innocent must be lower than the expected cost of acquitting an innocent, that is,

$$\Pr(G|E) \cdot c(AG) > [1 - \Pr(G|E)] \cdot c(CI).$$

The inequality holds just in case

$$\frac{\Pr(G|E)}{1 - \Pr(G|E)} > \frac{c(CI)}{c(AG)}.$$

Suppose $\frac{c(CI)}{c(AG)} = \frac{99}{1}$, as might be more appropriate in a criminal case in which the conviction of an innocent defendant is regarded as far worse than the acquittal of a guilty defendant. Then, the inequality holds only if $\Pr(G)$ meets the threshold 99%. More complicated models are also possible, but the basic idea is that the probability required for a conviction is a function of weighing the costs that would result from an erroneous decision.

It is not obvious how to assess all the required probabilities. The characterization of a decision criterion in terms of a probabilistic threshold is elegant, but its application in practice can be problematic. If a probabilistic threshold is understood as a criterion which the decision makers should mechanically apply whenever they confront the decision to convict or acquit, two difficulties arise. The first difficulty is that assigning a probability value to guilt itself might not be feasible. As seen earlier, the starting probability $\Pr(G)$ cannot be easily determined, and even if it could, other probabilities might be hard to assess. One solution here is that instead of aiming for a unique guilt probability, we can simply aim for an interval of admissible probabilities given the evidence. More generally, the assessment of the probability of guilt can be viewed as an idealized process, a regulative ideal which can improve the precision of legal reasoning.

Another problem with the probabilistic characterization is that it does not take into account the so-called weight of the evidence, that is, whether the evidence presented at trial contains all the evidence in the case or just a partial subset of the evidence. The guilt probability will vary dramatically depending on the evidence that is used to assess it. It is tempting to suggest that the guilt probability must be based on a body of evidence that is complete, or at least as complete as reasonably possible. And yet, it is unclear how to characterize this notion. No body of evidence is, strictly speaking, complete because new evidence could always be discovered and added.

6.2 Arguments

In an argumentative treatment, reasoning and decision making are analyzed in terms of the arguments that are collected.

Supporting and attacking reasons are collected and weighed. In a court of law, the prosecutor puts forward a conclusion and offers supporting reasons. The opposing side responds by offering attacking reasons. The dialectical process can be complex. As seen earlier, there are different attacking reasons: undermining, undercutting, and rebutting. The process is complex also because it can be iterated. A conclusion can be attacked by an attacking reason, and the latter in turn can be attacked and so on. When the dialectical process reaches an equilibrium point and the opposing parties have nothing more to contribute, the status of a claim and its supporting reasons can be assessed.

On the argument-based framework, the goal is to consider all the available reasons, by representing them in a comprehensive argumentation graph that keeps track of the relations of support and attack. The two competing theories of the cases, the prosecutor's and the defense's theory, will each be supported by a set of reasons. The argument framework, through the aid of argument graphs, allows us to compare the relative strength of the reasons in favor of one side of the case or the other. This comparison of the two sides forms the basis for the trial decision.

Defeating attacking arguments is the criterion for meeting the standard of proof. In order to establish the defendant's guilt beyond a reasonable doubt, all the attacks against the conclusion that the defendant is guilty must be defeated. Now, whether an attack is defeated is not always an all or nothing affair. It is often a matter of degrees. If the reasons for guilt are slightly stronger than all their attacks, this would not be enough yet. To meet the demands of the standard of proof beyond a reasonable doubt, the supporting reasons must be significantly stronger than all their attacks. On the other hand, defeating all the attacks with absolute certainty would be too much to expect. So, more realistically, all attacks must be defeated in an almost definitive way. Perhaps, we need to reintroduce some threshold, even though not in an explicitly probabilistic or numerical way.

It is not obvious when to stop collecting supporting and attacking reasons. The argumentation framework is rather realistic. The idea that meeting the standard of proof requires to answer all attacks against the conclusion that the defendant is guilty is natural enough. A problem is that if the opposing party puts forward no attacks, meeting the standard of proof would be effortless. A possible response here is that the attacks must be all the attacks which a reasonable objector could in principle put forward, not just the attacks that in fact are put forward. But who is this "reasonable objector"?

Another problem consists in identifying the threshold. While the probability-based account can identify a specific probability threshold, at least in theory, by applying the principle of expected utility theory, the argumentation-based framework cannot. How could expected utility theory be applied to the argument framework, as well?

6.3 Scenarios

In a scenario treatment, reasoning and decision making are analyzed by comparing the different scenarios.

Competing scenarios are collected and compared. On the scenario framework, the two parties will put forward competing scenarios, at least two or possibly more than two. This is partly problematic because in a criminal case, the defense does not have the burden of proof. So it might well be that the defense puts forward a scenario that weakens the prosecutor's scenario, but that is not a scenario that proves innocence. Be that as it may, the various competing scenarios will be evaluated along the different criteria we identified, such as consistency with the evidence, explanatory power, plausibility, coherence. The question arises, which scenario should be selected among the competitors?

The best explanatory scenario is the rule of decision. We can picture the process of evaluation of the competing scenarios as a process of elimination. At the beginning, several scenarios are viable, but as more evidence is considered and the scrutiny of each scenario continues, fewer scenarios will survive. The goal would be to select one scenario, or at least a limited set of scenarios, so that the answer to the question "guilty or not?" would be univocal. On this picture, a scenario meets the demands of the standard of proof whenever it is the *only* scenario left.

But, once again, we confront a recurrent problem. The selection of a scenario is not always an all-or-nothing affair. The term "abduction" or the expression "inference to the best explanation" is sometimes used in this context. The basic idea is that, when confronted with two or more competing scenarios, the best explanation must be chosen. The notion of "best explanation" here is wide-ranging. It includes criteria such as consistency with the evidence, explanatory power (predictive power and causal fit), plausibility, completeness, coherence (temporal and causal structure). Other criteria might also play a role, such as the simplicity of the scenario. The best explanation is the scenario that fares best on some combination of these criteria. In this background, the decision rule would stipulate that the best explanatory scenario should be selected.

It is not obvious how to identify the scenarios and how to compare them. The process of scenario analysis and selection resembles how jurors reason in trial proceedings, whereas—in contrast— it is hard to relate probability to judicial proceedings: Jurors do not naturally quantify guilt, and it can be difficult to quantify it even if we wanted to. Still, a problem with the scenario approach is that the method by which scenarios are identified and selected is not entirely transparent. When are all relevant scenarios identified? Should all scenarios mentioned in trial be taken into account, even when they seem far-fetched? Also, the different criteria, such as consistency, explanatory power, coherence, can pull the decision makers in opposite directions. For example, a scenario might be better in terms of explanatory power, while another might be more plausible. What to do, then? Perhaps a criterion for selecting the best scenario would ultimately be a qualitative version of selecting the most probable scenario, connecting a scenario-based approach with a probabilistic perspective.

7 Summary and Conclusion

We have discussed evidential reasoning in the law. We started out by discussing two common forms of evidence, eyewitness testimonies and DNA matches. We then distinguished three normative frameworks for theorizing about evidential reasoning: one focusing on the arguments for and against the positions taken; the second using probabilities to assess the evidential value of the evidence; and the third considering the scenarios that best explain the evidence. We then discussed four main themes: conflicting evidence; evidential value; the coherent interpretation of the evidence; and reasoning and decision making. For each theme, we discussed how they can be addressed in each of the three frameworks. We now summarize our discussion for each theme, using the highlighted phrases in the preceding sections.

7.1 Conflicting Evidence

Arguments Three kinds of attack can be distinguished: rebutting, undercutting, and undermining. Three kinds of support can be distinguished: multiple, subordinated, and coordinated. Arguments can involve complex structures of supporting and attacking reasons.

Scenarios There may be conflicting scenarios about what happened. Evidence can be explained by one scenario, but not by another. Scenarios can be contradicted by evidence.

Probabilities Support can be characterized as "probability increase" or "positive likelihood ratio." Attack can be characterized as "probability decrease" or "negative likelihood ratio." The conflict between two pieces of evidence can be described probabilistically.

7.2 Evidential Value

Probabilities The incremental evidential value is measured by probabilistic change. The overall evidential value is measured by the overall conditional probability. The use of evidence with high incremental evidential value has complications.

Arguments The reasons used can be conclusive or defeasible. Arguments can be evaluated by asking critical questions. It can be subject to debate whether a reason supports or attacks a conclusion.

Scenarios Scenarios can be plausible and logically consistent. The more evidence a scenario can explain, the better. The more pieces of evidence a scenario is consistent with, the better.

7.3 *Coherently Interpreting the Evidence*

Scenarios Scenarios are coherent clusters of events, ordered in time and with causal relations. Scenarios can be more or less complete. Weaker scenarios can be better supported by the evidence.
Arguments The evaluation of an argument can depend on its subarguments. The evaluation of an argument can depend on chains of attacks. Conflicts between reasons can be addressed by exceptions, preferences, and weighing.
Probabilities The likelihood ratio formula shows how to find the posterior odds given the evidence. A generalization of the formula shows how to handle more pieces of evidence. More complex analytic tools can be used, in particular Bayesian networks.

7.4 *Reasoning and Decision Making*

Probabilities The guilt probability is assessed by weighing the evidence with the probability calculus. The decision criterion is a guilt probability threshold. It is not obvious how to assess all the required probabilities.
Arguments Supporting and attacking reasons are collected and weighed. Defeating attacking arguments is the criterion for meeting the standard of proof. It is not obvious when to stop collecting supporting and attacking reasons.
Scenarios Competing scenarios are collected and compared. The best explanatory scenario is the rule of decision. It is not obvious how to identify the scenarios and how to compare them.

With the thematic discussion of the three normative frameworks, we have aimed to show how each framework contributes to the understanding of conflicting evidence, evidential value, the coherent interpretation of the evidence, and reasoning and decision making. In our perspective, there is no need to choose between the frameworks, since each adds to the normative analysis of evidential reasoning. At the same time, there is room for further studies of how the three normative frameworks relate to one another and how they can be integrated into a unified normative perspective on evidential reasoning.

8 Further Readings

For the interested readers, we now provide some reading suggestions, thematically organized following the headings of the previous sections. The full list of references is provided at the end.

8.1 Setting the Stage

Eyewitness Testimony: Eyewitness misidentification as one of the main causes of mistaken convictions (www.innocenceproject.org). Manipulations of the memory of witnesses (Loftus 1996). Selective attention in perception (Simons and Chabris 1999). Detecting lies and the psychology of lying (Vrij 2008). Holistic versus piecemeal face recognition (Tanaka and Farah 1993). Orientation of faces in identification tasks and the Tatcher illusion (Thomson 1980). Improving the probative value of eyewitness evidence (Wells et al. 2006). On a corroboration requirement for convictions solely based on eyewitness evidence (Thompson 2008; Crump 2009).
DNA Evidence: Basics of DNA evidence (Hicks et al. 2016; Wasserman 2008; Kaye and Sensabaugh 2000). DNA matches as a matter of degrees (Kaye 1993). Confusions between the source hypothesis and the guilt hypothesis and other exaggerations in the presentation of DNA evidence (Koehler 1993). Laboratory errors (Thompson et al. 2003). On whether DNA profiles are unique (Balding 1999; Weir 2007; Koehler and Saks 2010; Kaye 2013). History of the use of DNA evidence in court and the debate on the independence assumption Kaye (2010).

8.2 Three Normative Frameworks

The three frameworks for modeling evidential reasoning (Anderson et al. 2005; Kaptein et al. 2009; Dawid et al. 2011).

Arguments: Wigmore charts (Wigmore 1913). The New Evidence Scholarship (Anderson et al. 2005). Formal and computational study of arguments (Pollock 1987, 1995). Informal and formal argumentation theory (van Eemeren et al. 2014b).

Probabilities: Bayes' theorem (Swinburne 2002). Bayesian epistemology and updating (Bovens and Hartmann 2003a). Evidence and probabilities in the law (Dawid 2002; Schum 1994; Schum and Starace 2001; Mortera and Dawid 2007). Statistics in the law (Finkelstein and Levin 2001; Fenton 2011; Gastwirth 2012). Miscarriages of justice involving statistics (Dawid et al. 2011; Schneps and Colmez 2013). Debate on whether probabilistic calculations have a place in courts (Finkelstein and Fairley 1970; Tribe 1971; Fenton et al. 2016), and more recently, the 2012 special issue of *Law, Probability and Risk*; Vol. 11, No. 4.

Scenarios: Scenarios in evidential reasoning (Bennett and Feldman 1981; Pennington and Hastie 1993a, b). Scenarios and miscarriages of justice (Wagenaar et al. 1993). Inference to the best explanation (Pardo and Allen 2008). Hypothetical explanations of the evidence (Thagard 1989).

Combined Approaches: Combining arguments and scenarios (Bex et al. 2010; Bex 2011). Bayesian networks for evidential reasoning (Hepler et al. 2007; Fenton et al. 2013; Taroni et al. 2014). Combining arguments, scenarios and probabilities (Vlek et al. 2014, 2016; Timmer et al. 2017; Verheij et al. 2016; Verheij 2014, 2017).

8.3 Conflicting Evidence

Arguments: Argument structure and diagrams (Wigmore 1913; Toulmin 1958; Freeman 1991). Defeasible reasoning and nonmonotonic logic (Pollock 1987; Gabbay et al. 1994). Rebutting and undercutting attack (Pollock 1987, 1995). Undermining attack (Bondarenko et al. 1997). Formal evaluation of defeasible arguments (Pollock 1987, 1995; Dung 1995; Prakken 2010). Argumentative dialogue (Toulmin 1958; Walton and Krabbe 1995; Prakken 1997; Hage 2000). Argument diagramming and evaluation software (Pollock 1995; Reed and Rowe 2004; Kirschner et al. 2003; van Gelder 2003; Verheij 2005; Gordon et al. 2007).

Scenarios: Scenarios in evidential reasoning (Bennett and Feldman 1981; Pennington and Hastie 1993a, b). Scenarios and miscarriages of justice (Wagenaar et al. 1993). Inference to the best explanation (Pardo and Allen 2008). Hypothetical explanations of the evidence (Thagard 1989).

Probabilities: On confirmation theory and accounts of evidential support (Carnap 1950; Fitelson 1999; Skyrms 2000; Hacking 2001; Bovens and Hartmann 2003a; Crupi 2015). Probabilistic accounts of evidential support in the law (Lempert 1977). On whether the likelihood ratio should consider exhaustive hypotheses or not (Fenton et al. 2014; Biedermann et al. 2014).

8.4 Evidential Value

Probabilities: Introductions to using probability for weighing evidence (Finkelstein and Fairley 1970; Dawid 2002; Mortera and Dawid 2007). Critique of the probabilistic approach (Tribe 1971; Cohen 1977; Allen and Pardo 2007). Prosecutor's fallacy (Thompson and Schumann 1987). Introduction to DNA evidence (Wasserman 2008; Kaye and Sensabaugh 2000). Different hypotheses for evaluating DNA evidence (Koehler 1993; Cook et al. 1998; Evett et al. 2000). Probabilistic analyses of DNA evidence (Robertson and Vignaux 1995; Hicks et al. 2016; Balding 2005). Lab errors for DNA evidence (Thompson et al. 2003). Match is not all-or-nothing judgment (Kaye 1993). Uniqueness of DNA profiles (Balding 1999; Kaye 2013; Weir 2007). How DNA evidence can be synthesized and implanted (Frumkin et al. 2009). Cold hit controversy in DNA evidence cases (NRC 1996; Balding and Donnely 1996). Comparison between DNA evidence and fingerprints (Zabell 2005). Probabilistic analyses of eyewitness testimony (Friedman 1987; Schum 1994; Schum and Starace 2001).

Arguments: Nonmonotonic reasoning (Gabbay et al. 1994). Prima facie reasons, undercutting and rebutting defeaters (Pollock 1987, 1995). Warrants and backings (Toulmin 1958). Argument schemes and critical questions (Walton et al. 2008). Formal and computational argumentation (van Eemeren et al. 2014a).

Scenarios: Explanation in the deductive nomological model (Hempel and Oppenheim 1948). Explanation and causality (Salmon 1984). Abduction and inference to the best explanation (Lipton 1991). More the philosophical literature on scientific explanation (Woodward 2014). Two directions of fit (Wells 1992). Hypothetical explanations of the evidence (Thagard 1989). Scenario quality (Pennington and Hastie 1993b; Wagenaar et al. 1993; Bex 2011).

8.5 Coherently Interpreting the Evidence

Scenarios: Explanation and unification in philosophy of science (Friedman 1974). Coherence in epistemology (BonJour 1985). The crossword puzzle analogy for coherently evaluating a mass of evidence (Haack 2008). Explanatory coherence (Thagard 2001). Cognitive role of scripts (Schank and Abelson 1977). The story model (Pennington and Hastie 1993a). Scenarios as scripts (Wagenaar et al. 1993). Scenarios in legal cases (Griffin 2013). Evidence and scenario schemes (Bex 2011; Verheij et al. 2016; Vlek et al. 2014, 2016). Worries about scenarios in law (Velleman 2003). Scenarios shifting the legal perspective (Bex and Verheij 2013).

Arguments: Argument structure and their evaluation (Pollock 1995). Formalizing argumentation (Prakken and Vreeswijk 2002). Evaluating argument attack (Dung 1995). Formal argumentation models (Simari and Loui 1992; Vreeswijk 1997; Prakken 2010; Verheij 2003; Gordon et al. 2007). Informal and formal argumentation theory (van Eemeren et al. 2014b). Accrual of reasons and weighing (Pollock 1995; Hage 1997; Verheij 1996; Prakken 2005).

Probabilities: The conjunction paradox (Cohen 1977) and a response (Dawid 1987). Coherence and probability (Bovens and Hartmann 2003b). Probabilistic analysis of an entire legal case (Kadane and Schum 1996; Vlek et al. 2014, 2016). On the use of probability in law (Fenton 2011). Bayesian networks (Pearl 1988; Darwiche 2009; Jensen and Nielsen 2007; Fenton and Neil 2013). Bayesian networks for evidential reasoning (Taroni et al. 2014; Hepler et al. 2007; Fenton et al. 2013; Vlek et al. 2014, 2016; Timmer et al. 2017). Bayesian networks and causality (Pearl 2000/2009; Dawid 2010). Arguments, scenarios and probabilities (Keppens and Schafer 2006; Keppens 2012; Vlek et al. 2014, 2016; Timmer et al. 2017; Verheij et al. 2016; Verheij 2014, 2017).

8.6 Reasoning and Decision Making

Evidence law manuals (Fisher 2008; Méndez 2008). Criminal Procedure manuals (Allen et al. 2016; Roberts and Zuckerman 2010). On difficulties and confusions while defining proof beyond a reasonable doubt (Laudan 2006). Character evidence and its exclusion (Redmayne 2015).

Probabilities: Probabilistic accounts of the burden of proof (Kaplan 1968; Kaye 1986, 1999; Hamer 2004; Cheng 2013). Critique of probabilistic accounts (Cohen 1977; Nesson 1979; Thomson 1986; Stein 2005; Ho 2008; Pardo and Allen 2008; Haack 2014). On the question whether the threshold should be variable (Kaplow 2012; Picinali 2013). The problem of priors (Finkelstein and Fairley 1970; Friedman 2000). A critique of proof beyond a reasonable doubt as understood in the law (Laudan 2006). History of beyond a reasonable doubt standard (Shapiro 1991; Whitman 2008). Other measures, weight, resiliency and completeness of the evidence (Kaye 1999; Stein 2005; Nance 2016).

Arguments: Evaluating arguments and their attacks (Pollock 1995; Dung 1995). Burden of proof and argumentation (Gordon et al. 2007; Gordon and Walton 2009; Prakken and Sartor 2007, 2009). Weighing reasons (Hage 1997).

Scenarios: Inference to the best explanation (Lipton 1991). Application of inference to the best explanation to legal reasoning (Pardo and Allen 2008). Narrative based account of proof beyond a reasonable doubt (Allen 2010; Allen and Stein 2013).

Acknowledgements This chapter has been developed in the context of the project "Designing and Understanding Forensic Bayesian Networks with Arguments and Scenarios," funded in the NWO Forensic Science program (http://www.ai.rug.nl/~verheij/nwofs/). The first author would like to thank Infosys Limited which made possible his stay at the Institute for Advanced Study in Princeton for the academic year 2016–17 during which parts of this chapter were written. The second author would like to thank the Isaac Newton Institute for Mathematical Sciences at the University of Cambridge for its hospitality during the program "Probability and Statistics in Forensic Science" which was supported by EPSRC Grant Number EP/K032208/1. The authors would like to thank Ronald Allen, Alex Biedermann, Christian Dahlman, and Norman Fenton for helpful comments and suggestions on an earlier draft.

References

Allen, R.J. 2010. No plausible alternative to a plausible story of guilt as the rule of decision in criminal cases. In *Prueba y Esandares de Prueba en el Derecho*, ed. J. Cruz, and L. Laudan. Mexico: Instituto de Investigaciones Filosoficas-UNAM.

Allen, R.J., and M.S. Pardo. 2007. The problematic value of mathematical models of evidence. *Journal of Legal Studies* 36 (1): 107–140.

Allen, R.J., and A. Stein. 2013. Evidence, probability and the burden of proof. *Arizona Law Journal* 55: 557–602.

Allen, R.J., W.J. Stuntz, J.L. Hoffmann, D.A. Livingston, A.D. Leipold, and T.L. Meares. 2016. *Comprehensive criminal procedure*, 3rd ed. New York, N.Y.: Wolters Kluwer.

Anderson, T., D. Schum, and W. Twining. 2005. *Analysis of Evidence*, 2nd ed. Cambridge: Cambridge University Press.

Balding, D.J. 1999. When can a DNA profile be regarded as unique? *Science & Justice* 39.

Balding, D.J. 2005. *Weight-of-evidence for forensic DNA profiles*. West Sussex: Wiley.

Balding, D.J., and P. Donnely. 1996. Evaluating DNA profile evidence when the suspect is identified through a database search. *Journal of Forensic Science* 41: 603–607.

Bennett, W.L., and M.S. Feldman. 1981. *Reconstructing reality in the courtroom*. London: Tavistock Feldman.

Bernoulli, J. 1713. *Ars Conjectandi*.

Bex, F.J. 2011. *Arguments, stories and criminal evidence: A formal hybrid theory*. Berlin: Springer.

Bex, F.J., P.J. van Koppen, H. Prakken, and B. Verheij. 2010. A hybrid formal theory of arguments, stories and criminal evidence. *Artificial Intelligence and Law* 18: 1–30.

Bex, F.J., and B. Verheij. 2013. Legal stories and the process of proof. *Artificial Intelligence and Law* 21 (3): 253–278.

Biedermann, A., T. Hicks, F. Taroni, C. Champod, C. Aitken. On the use of the likelihood ratio for forensic evaluation: Response to Fenton, et al. 2014. *Science and Justice* 54 (4): 316–318.

Bondarenko, A., P.M. Dung, R.A. Kowalski, and F. Toni. 1997. An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence* 93: 63–101.

BonJour, L. 1985. *The structure of empirical knowledge*. Cambridge, MA: Harvard University Press.

Bovens, L., and S. Hartmann. 2003a. *Bayesian Epistemology*. Oxford: Oxford University Press.

Bovens, L., and S. Hartmann. 2003b. Solving the riddle of coherence. *Mind* 112: 601–633.

Carnap, R. 1950. *Logical foundations of probability*. Chicago, IL: University of Chicago Press.

Cheng, E. 2013. Reconceptualizing the burden of proof. *Yale Law Journal* 122 (5): 1104–1371.

Cohen, L.J. 1977. *The probable and the provable*. Oxford: Clarendon Press.

Cook, R., I.W. Evett, G. Jackson, P.J. Jones, and J.A. Lambert. 1998. A hierarchy of propositions: Deciding which level to address in casework. *Science and Justice* 38 (4): 231–239.

Crump, D. 2009. Eyewitness corroboration requirements as protections against wrongful conviction: The hidden questions. *Ohio State Journal of Criminal Law* 7 (1): 361–376.

Crupi, V. 2015. Confirmation. In *Stanford encyclopedia of philosophy*, ed. E.N. Zalta. Stanford University.

Darwiche, A. 2009. *Modeling and reasoning with bayesian networks*. Cambridge: Cambridge University Press.

Dawid, A.P. 1987. The difficulty about conjunction. *Journal of the Royal Statistical Society. Series D (The Statistician)* 36(2/3):91–92.

Dawid, A.P. 2002. Bayes's theorem and weighing evidence by juries. In *Bayes's Theorem*, vol. 113, 71–90, Oxford: Oxford University Press.

Dawid, A.P. 2010. Beware of the DAG! In *JMLR workshop and conference proceedings: Volume 6. Causality: Objectives and assessment (NIPS 2008 workshop)*, eds. I. Guyon, D. Janzing, and B. Schölkopf, 59–86. http://www.jmlr.org/

Dawid, A.P., W. Twining, and M. Vasiliki (eds.). 2011. *Evidence, inference and enquiry*. Oxford: Oxford University Press.

Dung, P.M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77: 321–357.

Evett, I., G. Jackson, J.A. Lambert, and S. McCrossan. 2000. The impact of the principles of evidence interpretation on the structure and content of statements. *Science and Justice* 40 (4): 233–239.

Fenton, N., D. Berger, D. Lagnado, M. Neil, and A. Hsu. 2014. When "neutral" evidence still has probative value (with implications from the barry george case). *Science and Justice* 54 (4): 274–287.

Fenton, N., M. Neil, and D. Berger. 2016. Bayes and the law. *Annual Review of Statistics and Its Application* 3.

Fenton, N.E. 2011. Science and law: Improve statistics in court. *Nature* 479: 36–37.

Fenton, N.E., and M.D. Neil. 2013. *Risk assessment and decision analysis with Bayesian networks*. Boca Raton, FL: CRC Press.

Fenton, N.E., M.D. Neil, and D.A. Lagnado. 2013. A general structure for legal arguments about evidence using Bayesian Networks. *Cognitive Science* 37: 61–102.

Finkelstein, M.O., and W.B. Fairley. 1970. A Bayesian approach to identification evidence. *Harvard Law Review* 83: 489–517.

Finkelstein, M.O., and B. Levin. 2001. *Statistics for lawyers*. Berlin: Springer.

Fisher, G. 2008. *Evidence*, 2nd ed. New York, N.Y.: Foundation Press.

Fitelson, B. 1999. The plurality of Bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science* 66: 362–378.

Freeman, J.B. 1991. *Dialectics and the macrostructure of arguments. A theory of argument structure*. Berlin: Foris.

Friedman, M. 1974. Explanation and scientific understanding. *Journal of Philosophy* 71: 5–19.

Friedman, R.D. 1987. Route analysis of credibility and hearsay. *The Yale Law Journal* 97 (4): 667–742.

Friedman, R.D. 2000. A presumption of innocence, not of even odds. *Stanford Law Review* 52: 873–887.

Frumkin, D., A. Wasserstrom, A. Davidson, and A. Grafit. 2009. Authentication of forensic DNA samples. *Forensic Science International: Genetics* 4 (2): 95–103.

Gabbay, D.M., C.J. Hogger, and J.A. Robinson (eds.). 1994. *Handbook of logic in artificial intelligence and logic programming. Volume 3. Nonmonotonic reasoning and uncertain reasoning*. Oxford: Clarendon Press.

Gastwirth, J.L. (ed.). 2012. *Statistical Science in the Courtroom*. Berlin: Springer.

Gordon, T.F., H. Prakken, and D.N. Walton. 2007. The Carneades model of argument and burden of proof. *Artificial Intelligence* 171 (10–15): 875–896.

Gordon, T.F., and D.N. Walton. 2009. Proof burdens and standards. In *Argumentation in artificial intelligence*, ed. I. Rahwan, and G.R. Simari, 239–258. Berlin: Springer.

Griffin, L.K. 2013. Narrative, truth, trial. *Georgetown Law Journal* 101: 281–335.

Haack, S. 2008. Warrant, causation, and the atomist of evidence law. *Journal of Social Epistemology* 5: 253–265.

Haack, S. 2014. Legal probabilism: An epistemological dissent. In *Science, proof, and truth in the law*, ed. Evidence Matters, 47–77. Cambridge: Cambridge University Press.

Hacking, I. 2001. *An introduction to probability and inductive logic*. Cambridge: Cambridge University Press.

Hage, J.C. 1997. *Reasoning with rules. An essay on legal reasoning and its underlying logic*. Dordrecht: Kluwer.

Hage, J.C. 2000. Dialectical models in artificial intelligence and law. *Artificial Intelligence and Law* 8: 137–172.

Hamer, D. 2004. Probabilistic standards of proof, their complements and the errors that are expected to flow from them. *University of New England Law Journal* 1 (1): 71–107.

Hempel, C., and P. Oppenheim. 1948. Studies in the logic of explanation. *Philosophy of Science* 15: 135–175.

Hepler, A.B., A.P. Dawid, and V. Leucari. 2007. Object-oriented graphical representations of complex patterns of evidence. *Law, Probability and Risk* 6 (1–4): 275–293.

Hicks, T., J. Buckleton, J.-A. Bright, and D. Taylor. 2016. A framework for interpreting evidence. In *Forensic DNA Evidence Interpretation (second edition)*, ed. J. Buckleton, J.-A. Bright, and D. Taylor. Boca Raton, FL: CRC Press.

Ho, H.L. 2008. *Philosophy of evidence law*. Oxford: Oxford University Press.

Jensen, F.V., and T.D. Nielsen. 2007. *Bayesian networks and decision graphs*. Berlin: Springer.

Kadane, J.B., and D.A. Schum. 1996. *A probabilistic analysis of the Sacco and Vanzetti evidence*. Chichester: Wiley.

Kaplan, J. 1968. Decision theory and the fact-finding process. *Stanford Law Review* 20: 1065–1092.

Kaplow, L. 2012. Burden of proof. *Yale Law Journal* 121 (4): 738–1013.

Kaptein, H., H. Prakken, and B. Verheij (eds.). 2009. *Legal evidence and proof: statistics, stories, logic (Applied legal philosophy series)*. Farnham: Ashgate.

Kaye, D.H. 1986. Do we need a calculus of weight to understand proof beyond a reasonable doubt? *Boston University Law Review* 66: 657–672.

Kaye, D.H. 1993. DNA evidence: Probability, population genetics and the courts. *Harvard Journal of Law and Technology* 7: 101–172.

Kaye, D.H. 1999. Clarifying the burden of persuasion: What Bayesian rules do and not do. *International Commentary on Evidence* 3: 1–28.

Kaye, D.H. 2010. *The double helix and the law of evidence*. Cambridge, Mass.: Harvard University Press.

Kaye, D.H. 2013. Beyond uniqueness: the birthday paradox, source attribution and individualization in forensic science. *Law, Probability and Risk* 12 (1): 3–11.

Kaye, D.H., and G.F. Sensabaugh. 2000. Reference guide on DNA evidence. In *Reference manual on scientific evidence*, 2nd ed., 576–585. Washington, D.C.: Federal Judicial Center.

Keppens, J. 2012. Argument diagram extraction from evidential Bayesian networks. *Artificial Intelligence and Law* 20: 109–143.

Keppens, J., and B. Schafer. 2006. Knowledge based crime scenario modelling. *Expert Systems with Applications* 30 (2): 203–222.

Kirschner, P.A., S.J.B. Shum, and C.S. Carr. 2003. *Visualizing argumentation: Software tools for collaborative and educational sense-making*. Berlin: Springer.

Koehler, J.J. 1993. Error and exaggeration in the presentation of DNA evidence in trial. *Jurimetrics Journal* 34: 21–39.

Koehler, J.J., and M.J. Saks. 2010. Individualization claims in forensic science: Still unwarranted. *Brooklyn Law Review* 75 (4): 1187–1208.

Laplace, P.-S. 1814. *Essai Philosophique sur les Probabilités*.

Laudan, L. 2006. *Truth, error, and criminal law: An essay in legal epistemology*. Cambridge: Cambridge University Press.

Lempert, R.O. 1977. Modeling relevance. *Michigan Law Review* 75 (5/6): 1021–1057.

Lipton, P. 1991. *Inference to the best explanation*. New York, N.Y.: Routledge.

Loftus, E.F. 1996. *Eyewitness testimony (revised edition)*. Cambridge, MA: Harvard University Press.

Méndez, M.A. 2008. *Evidence: The California code and the Federal rules*, 4th ed. Eagan, MN: Thomson West.

Mortera, J., and P. Dawid. 2007. Probability and evidence. In *Handbook of probability theory*, ed. T. Rudas. Los Angeles, CA: Sage.

Nance, D.A. 2016. *The burdens of proof: Discriminatory power, weight of evidence, and tenacity of belief*. Cambridge: Cambridge University Press.

Nesson, C.R. 1979. Reasonable doubt and permissive inferences: The value of complexity. *Harvard Law Review* 92 (6): 1187–1225.

NRC. 1996. *The evaluation of forensic DNA evidence*. Washington, D.C.: National Academy Press.

Pardo, M.S., and R.J. Allen. 2008. Juridical proof and the best explanation. *Law and Philosophy* 27: 223–268.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.

Pearl, J. 2000/2009. *Causality: Models, reasoning and inference*, 2nd ed. Cambridge: Cambridge University Press.

Pennington, N., and R. Hastie. 1993a. *Inside the juror*, chap. The story model for juror decision making, 192–221. Cambridge: Cambridge University Press.

Pennington, N., and R. Hastie. 1993b. Reasoning in explanation-based decision making. *Cognition* 49 (1–2): 123–163.

Picinali, F. 2013. Two meanings of "reasonableness": Dispelling the "floating" reasonable doubt. *Modern Law Review* 76 (5): 845–875.

Pollock, J.L. 1987. Defeasible reasoning. *Cognitive Science* 11 (4): 481–518.

Pollock, J.L. 1995. *Cognitive Carpentry: A blueprint for how to build a person*. Cambridge, MA: The MIT Press.

Prakken, H. 1997. *Logical tools for modelling legal argument. A study of defeasible reasoning in law*. Dordrecht: Kluwer.

Prakken, H. 2005. A study of accrual of arguments, with applications to evidential reasoning. In *Proceedings of the tenth international conference on artificial intelligence and law*, 85–94, New York (New York): ACM Press.

Prakken, H. 2010. An abstract framework for argumentation with structured arguments. *Argument and Computation* 1 (2): 93–124.

Prakken, H., and G. Sartor. 2007. Formalising arguments about the burden of persuasion. In *Proceedings of the 11th international conference on artificial intelligence and law*, 97–106, New York, N.Y.: ACM Press.

Prakken, H., and G. Sartor. 2009. A logical analysis of burdens of proof. In *Legal evidence and proof: Statistics, stories, logic*, chap. 9, ed. H. Kaptein, H. Prakken, and B. Verheij, 223–253, Farnham: Ashgate.

Prakken, H., and G.A.W. Vreeswijk. 2002. Logics for defeasible argumentation. In *Handbook of philosophical logic*, vol. 4, 2nd ed. D.M. Gabbay, and F. Guenthner, 218–319. Dordrecht: Kluwer Academic Publishers.

Redmayne, M. 2015. *Character evidence in the criminal trial*. Oxford: Oxford University Press.

Reed, C., and G. Rowe. 2004. Araucaria: Software for argument analysis, diagramming and representation. *International Journal of AI Tools* 14 (3–4): 961–980.

Robertson, B., and G.A. Vignaux. 1995. DNA evidence: Wrong answers or wrong questions? *Genetica* 96: 145–152.

Roberts, P., and A. Zuckerman. 2010. *Criminal evidence*, 2nd ed. Oxford: Oxford University Press.

Salmon, W. 1984. *Scientific explanation and the causal structure of the world*. Princeton, N.J.: Princeton University Press.

Schank, R., and R. Abelson. 1977. *Scripts, plans, goals and understanding, an inquiry into human knowledge structures*. Hillsdale: Lawrence Erlbaum.

Schneps, L., and C. Colmez. 2013. *Math on trial: How numbers get used and abused in the courtroom*. New York, N.Y.: Basic Books.

Schum, D.A. 1994. *The evidential foundations of probabilistic reasoning*. New York, N.Y.: Wiley.

Schum, D.A., and S. Starace. 2001. *The evidential foundations of probabilistic reasoning*. Evanston, Il.: Northwestern University Press.

Shapiro, B. 1991. *Beyond reasonable doubt and probable cause: Historical perspectives on the Anglo-American law of evidence*. Oakland, Calif.: University of California Press.

Simari, G.R., and R.P. Loui. 1992. A mathematical treatment of defeasible reasoning and its applications. *Artificial Intelligence* 53: 125–157.

Simons, D.J., and C.F. Chabris. 1999. Gorillas in our minds: Sustained inattention blindness for dynamic events. *Perception* 28: 1059–1074.

Skyrms, B. 2000. *Choice and chance: An introduction to inductive logic*, 4th ed. Belmont, CA: Wadsworth.

Stein, A. 2005. *Foundations of evidence law*. Oxford: Oxford University Press.

Swinburne, R. (ed.). 2002. *Bayes's theorem*. Oxford: Oxford University Press.

Tanaka, J.W., and M.J. Farah. 1993. Parts and whole in face recognition. *The Quarterly Journal of Experimental Psychology* 46A (3): 225–245.

Taroni, F., A. Biedermann, S. Bozza, P. Garbolino, and C. Aitken. 2014. *Statistics in practice. In Bayesian networks for probabilistic inference and decision analysis in forensic science*, 2nd ed. Chichester: Wiley.

Taroni, F., C. Champod, and P. Margot. 1998. Forerunners of Bayesianism in early forensic science. *Jurimetrics* 38: 183–200.

Thagard, P. 1989. Explanatory coherence. *Behavioral and Brain Sciences* 12: 435–502.

Thagard, P. 2001. *Coherence in thought and action*. Cambridge, MA: The MIT Press.

Thompson, S.G. 2008. Beyond a reasonable doubt? reconsidering uncorroborated eyewitness identification testimony. *UC Davis Law Review* 41: 1487–1545.

Thompson, W.C., and E.L. Schumann. 1987. Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior* 11: 167–187.

Thompson, W.C., F. Taroni, and C.G.G. Aitken. 2003. How the probability of a false positive affects the value of DNA evidence. *Journal of Forensic Science* 48: 47–54.

Thomson, J.J. 1986. Liability and individualized evidence. *Law and Contemporary Problems* 49 (3): 199–219.

Thomson, P. 1980. Margaret Thatcher: A new illusion. *Perception* 9 (4): 483–484.

Tillers, P. 2011. Trial by mathematics-reconsidered. *Law, Probability and Risk* 10: 167–173.

Timmer, S.T., J.J. Meyer, H. Prakken, S. Renooij, and B. Verheij. 2017. A two-phase method for extracting explanatory arguments from Bayesian networks. *International Journal of Approximate Reasoning* 80: 475–494.

Toulmin, S.E. 1958. *The uses of argument*. Cambridge: Cambridge University Press.

Tribe, L. 1971. Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review* 84: 1329–1393.

van Eemeren, F.H., B. Garssen, E.C.W. Krabbe, A.F. Snoeck Henkemans, B. Verheij, and J.H.M. Wagemans. 2014a. *Chapter 11: Argumentation in artificial intelligence*. In Handbook of argumentation theory. Berlin: Springer.

van Eemeren, F.H., B. Garssen, E.C.W. Krabbe, A.F. Snoeck Henkemans, B. Verheij, and J.H.M. Wagemans. 2014b. *Handbook of argumentation theory*. Berlin: Springer.

van Gelder, T. 2003. Enhancing deliberation through computer supported argument visualization. In *Visualizing argumentation: Software tools for collaborative and educational sense-making*, ed. P.A. Kirschner, S.J.B. Shum, and C.S. Carr, 97–115. New York, N.Y.: Springer.

Velleman, D. 2003. Narrative explanation. *The Philosophical Review* 112 (1): 1–25.

Verheij, B. 1996. *Rules, reasons, arguments. Formal studies of argumentation and defeat.*. Maastricht: Dissertation Universiteit Maastricht.

Verheij, B. 2003. DefLog: on the logical interpretation of prima facie justified assumptions. *Journal of Logic and Computation* 13 (3): 319–346.

Verheij, B. 2005. *Virtual arguments. On the design of argument assistants for lawyers and other arguers*. The Hague: T.M.C. Asser Press.

Verheij, B. 2014. To catch a thief with and without numbers: Arguments, scenarios and probabilities in evidential reasoning. *Law, Probability and Risk* 13: 307–325.

Verheij, B. 2017. Proof with and without probabilities. correct evidential reasoning with presumptive arguments, coherent hypotheses and degrees of uncertainty. *Artifical Intelligence and Law* 25(1):127–154.

Verheij, B., F.J. Bex, S.T. Timmer, C.S. Vlek, J.J. Meyer, S. Renooij, and H. Prakken. 2016. Arguments, scenarios and probabilities: Connections between three normative frameworks for evidential reasoning. *Law, Probability and Risk* 15 (1): 35–70.

Vlek, C.S., H. Prakken, S. Renooij, and B. Verheij. 2014. Building Bayesian Networks for legal evidence with narratives: a case study evaluation. *Artifical Intelligence and Law* 22 (4): 375–421.

Vlek, C.S., H. Prakken, S. Renooij, and B. Verheij. 2016. A method for explaining Bayesian Networks for legal evidence with scenarios. *Artifical Intelligence and Law* 24 (3): 285–324.

Vreeswijk, G.A.W. 1997. Abstract argumentation systems. *Artificial Intelligence* 90: 225–279.

Vrij, A. 2008. *Detecting lies and deceit: The psychology of lying and the implications for professional practice*. Chichester: Wiley.

Wagenaar, W.A., P.J. van Koppen, and H.F.M. Crombag. 1993. *Anchored narratives: The psychology of criminal evidence*. London: Harvester Wheatsheaf.

Walton, D.N., and E. Krabbe. 1995. *Commitment in dialogue. Basic concepts of interpersonal reasoning*. Albany (New York): State University of New York Press.

Walton, D.N., C. Reed, and F. Macagno. 2008. *Argumentation schemes*. Cambridge: Cambridge University Press.

Wasserman, D. 2008. Forensic DNA typing. In *A companion to genethics*, ed. J. Burley, and J. Harris. Malden, MA: Blackwell.

Weir, B.S. 2007. The rarity of DNA profiles. *The Annals of Applied Statistics* 1: 358–370.

Wells, G.L. 1992. Naked statistical evidence of liability: Is subjective probability enough? *Journal of Personality and Social Psychology* 62: 793–752.

Wells, G.L., A. Memon, and S.D. Penrod. 2006. Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest* 7 (2): 45–75.

Whitman, J.Q. 2008. *The origins of reasonable doubt: Theological roots of the criminal trial*. New Haven, CT: Yale University Press.

Wigmore, J.H. 1913. *The principles of judicial proof as given by logic, psychology, and general experience, and illustrated in judicial trials. Second edition 1931, third edition "The science of judicial proof" 1937*. Boston, MA: Little, Brown and Company.

Woodward, J. 2014. Scientific explanation. In *The Stanford encyclopedia of philosophy*, ed. E.N. Zalta. Stanford University.

Zabell, S.L. 2005. Fingerprint evidence. *Journal of Law and Policy* 13: 143–179.

Interpretive Arguments and the Application of the Law

J. J. Moreso and Samuele Chilovi

And this undulating imprecision, this uncertainty, is the strange matter of which he is made.

—Jorge Luis Borges (1999, 279).

1 The Charm of Interpretation[1]

To apply the law is to solve a legal dispute by subsuming an individual case under a general normative premise (a legal rule, legal standard, legal principle, legal precedent, etc.). In order to find this normative premise, it is usually thought that we need

[1] This Very Fortunate Expression Is Owed to Endicott (2012), Chap. 8.

J. J. Moreso is professor of legal philosophy at Pompeu Fabra University, Barcelona (Spain).
Samuele Chilovi is a doctoral student at the University of Barcelona, his research has been partially funded by FFI2012-35026 and FFI2015-66372-P MINECO/FEDER, UE.

J. J. Moreso (✉)
Departament de Dret, Universitat Pompeu Fabra, Barcelona, Spain
e-mail: josejuan.moreso@upf.edu

S. Chilovi
Departament de Filosofia, Universitat de Barcelona, Barcelona, Spain
e-mail: samuele.chilovi@gmail.com

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_17

to interpret some legal texts (constitutional text, statutes, judicial decisions).[2] Therefore, legal interpretation is the usual way to ascertain the legal obligations, powers, and rights of citizens. Legal interpretation is the door to legal content.

On the other hand, legal theory has elaborated an assemblage of arguments (analogy, *a contrario*, *a fortiori*, and so on) which can help us understand what the law requires. They are like paths to the content of the law. Perhaps for this reason, it has sometimes been suggested that legal reasoning is a special kind of reasoning, with a proper and different logic. However, for present purposes we are going to assume that applying norms to the facts requires nothing more than classical logic, perhaps extended to embrace deontic logic. Accordingly, we will regard legal reasoning as a kind of (classical) deductive reasoning (Alexander 1998, 522).

Legal reasoning is not limited to judicial reasoning, as any rational being can produce a legal argument. For instance, if the municipal law concedes the right to vote in the general elections to every adult citizen, and NN is an adult citizen, then she has the right to vote in the general election. However, the relevance of judicial reasoning is undisputable. Often, in legal contexts, controversies are reasoned out from a judicial point of view. Moreover, judicial reasoning has legal effects and put an end to controversies. This is the legal doctrine of *res iudicata* (see Raz 1995b, 2009b). Thus, here we shall understand legal reasoning mainly as judicial reasoning.

A distinctive approach to legal interpretation—called "communicative-content theory of law" (CT), see Greenberg 2011b, 217—has recently been put forward and defended by some philosophers of language, proceeding from the insight that legal interpretation can be modelled on utterance interpretation. In this chapter, we shall defend a modified version of it that does not depend on some controversial aspects of the theory of communication to which proponents of this theory have typically appealed (Neale 2008, 2012; Soames 2008b, 2013). To do so, we first evaluate a sceptical objection directed at theories of legal interpretation generally and point to the way in which it can be easily overcome; secondly, we elaborate an argument that casts light on the flaws of CT by illuminating its tension with the conventional nature of law; accordingly, the result we reach suggests how the theory can be modified to accommodate the criticism triggered by the objection. Throughout, we clarify the claims and assumptions of CT and extract some of their consequences partly by introducing them through examples. Since nowhere, to our knowledge, has the enterprise so far been taken up in detail, we think the descriptive side of the inquiry carries independent theoretical interest. Relatedly, we use the *expressio unius est exclusio alterius* canon of construction both for explanatory purposes in dealing with the above-mentioned task and, eventually, for the purpose of making an argument to the conclusion that the role played by interpretive arguments in law gives us reasons

[2] See, for instance, Pound (1922), Chap. III: "Three steps are involved in the adjudication of a controversy according to law: (1) Finding the law, ascertaining which of the many rules in the legal system is to be applied, or, if none is applicable, reaching a rule for the cause (which may or may not stand as a rule for subsequent cases) on the basis of given materials in some way which the legal system points out; (2) interpreting the rule so chosen or ascertained, that is, determining its meaning as it was framed and with respect to its intended scope; (3) applying to the cause in hand the rule so found and interpreted."

to reject the account of legal content defended by CT in favour of a minimal version of the theory.

As noticed, legal reasoning proceeds by drawing an inference from the premises to the conclusion of an ordinary deductive argument. The argument's major premise encapsulates the content of the applicable legal norm,[3] a content that itself stands in need of explanation. As an effort in meeting this demand, legal interpretation is meant to be the activity through which the content of the applicable law is uncovered and displayed, thereby providing the general normative premise from which a legal decision is eventually derived. Competing approaches to the interpretation of legal sources differ in the way they think their content is determined. If talk about sources and the contents thereof is not vacuous, the arguable options can't be many. Some think that the lawmakers' intentions alone are determinative of the sources' propositional content, while others will dispute this and argue that the literal meaning of the text's sentences uniquely does, plus a number of hybrids in between. Hybrids will differ by the extent to which intentions are taken to be determinative of the law's content, and on the nature of those intentions. Relatedly, interpretive theories sometimes frame their disagreement in terms of the different analyses they give of the interpreted object. This leads to divergent conceptions of the identity of the source, focusing on utterer meaning (intentionalism), sentence meaning (textualism), and utterance meaning (hybrids), respectively.[4] It thus seems clear that in order to identify the law's content, the interpreter must specify the object whose meaning he aims to display and further needs an account of the things in virtue of which its content is as it is. This will then provide an account of those facts that make it the case that an interpretation is correct.

The communication theory of law holds the characteristic view that facts relating to the nature of language, perhaps together with facts relating to the nature of law, are able to tell us what theory of legal interpretation is correct in general, irrespective of the local features of the particular legal system which is being taken into consideration. In this paper, we argue that there is a tension between CT and the doctrine of

[3] We will be using the terms *norm*, *requirement*, and *obligation* interchangeably as shorthand for the facts about the content of the law in a given legal system, at a given time.

[4] All three variants share the common assumption that the object (source) whose meaning they purport to account for is some sort of *act* of meaning. If the assumption is correct, this may seem to rule out what jurists sometimes call "evolutive interpretation" from the arguable interpretative strategies, since the relevant feature of the speech act (be it sentence or utterer meaning) should in any event be fixed at the time the speech act is performed. We don't rule out this possibility, though we notice that there are at least two ways in which some form of evolutive interpretation may still be argued for. One is by arguing that (i) word meaning is the interpretative target and, (ii) though it is fixed at the time of the text's enactment, (iii) some terms' extension may vary over time even without there being any change in meaning. (This position is defended by Perry (2011), who views it as a form of textualism, since it takes word and sentence meaning [at the time of the enactment] as authoritative.) Alternatively, evolutive interpretation may be advocated for the interpretation of some terms by (i) assuming utterance meaning as the interpretative target, and (ii) defending an account of certain terms as *interpretation-sensitive*, thereby conferring a content-creating role on the interpreter in relation to them. (This view is defended in Cappelen (2009) with respect to the interpretation of legal texts.)

the rule of recognition—a central element of Hartian positivism—and, hence, that insofar as the doctrine is plausible, this suggests that CT is in fact problematic. This tension derives from the role the two views assign to interpretive arguments within legal interpretation. We think that an account of the role and status of interpretive arguments within legal interpretation should meet three desiderata.

First, insofar as lawmaking takes place through the enactment of a text, the content of the resulting legal norms cannot be wholly separated from the authoritative pronouncement's linguistic import; this follows as a result of acknowledging that lawmaking actions and utterances have the nature of performative speech acts of *some* kind.[5] Secondly, the nature of meaning and communication sets constraints on the way legal content might be determined by the enactment of provisions framed in a natural language, in particular by explicating what role (if any) linguistic meaning, context, authorial intentions, and possibly other factors play in determining different levels of utterance content, which thereby helps identify suitable candidate relations between levels of content and legal facts.[6] Thirdly, the principles that map certain features of authoritative utterances to facts about the content of the law largely depend on the social practices consistently and systematically followed by the legal agents of the relevant community (Rosen 2011), if not on normative elements too (Greenberg 2004; Dworkin 1986) .

The first two tenets are common ground among many scholars and certainly are shared by communication theorists themselves. By contrast, the reading of the third tenet we favour—according to which the mapping principles (relative to a given system at a given time) depend on the interpretive rules followed by the courts (in a system, at a time)—will be shown to be at odds with the communication theory of law. A thorough analysis both of the nature of interpretive arguments in law and of the role they play within the communication theory of legal interpretation will be key to understanding why this is so.

The paper's structure is as follows: Sect. 2 takes up the challenge, posed by a sceptical argument, to the effect that it is impossible to identify the content of legally valid rules through standard interpretive methods; Sect. 3 starts by offering a brief statement of CT and then articulates the theory of interpretation it lends support to; Sect. 4 analyses CT's account of the validity and content of legal norms and argues that it cannot bear scrutiny without appealing to further grounds; Sect. 5 examines an attempt to save the theory by expanding its resources in a certain way and points to some critical aspects of it; Sect. 6 concludes that the analysis warrants endorsement of only a minimal version of CT.

[5] This aspect is emphasized by Marmor (2011a, b), who goes further and analyses lawmaking in analogy to exhortatives.

[6] Soames (2008b) and Neale (2008, 2012) correctly criticize both the majority and the dissent opinion in the judgment issued by the US Supreme Court on *Smith v. United States* (508 U.S. 113 S. Ct. 2050, 1993) for their failure to appreciate the ways in which an expression's or a sentence's linguistic meaning may underdetermine what a speaker uses that sentence to assert on a given occasion.

2 General Scepticism About Legal Rules

The sceptical objection we consider seeks to undermine the ability of legal interpretation to provide determinately correct answers regarding the content of any law. The way it does so is by contending that legal interpretation can virtually never univocally identify the applicable legal content. When we use the interpretive methods and the canons of legal interpretation, we always have at our disposal several methods and canons that are in tension with one another and produce different and contradictory results.

This is clearly the view of Karl N. Llewellyn in the context of the American legal realist criticism of the traditional conceptions of interpretation. In a famous article, Llewellyn (1950) pointed out twenty-six cases in which we have at least two canons of construction providing incompatible solutions. And for this reason[7]:

> The major defect in that system is a mistaken idea which many lawyers have about it—to wit, the idea that the cases themselves and in themselves, plus the correct rules on how to handle cases, provide one single correct answer to a disputed issue of law. In fact the available correct answers are two, three, or ten. The question is: *Which* of the available correct answers will the court *select*—and *why*? For since there is always more than one available correct answer, the court always has to select. (Llewellyn 1950, 396)

Sometimes this approach is also adopted in European legal theory. For example, in the context of so-called Italian legal realism, Guastini (2011, 148–149) puts the same idea in the following terms:

> In most cases one and the same normative sentence may express different meanings depending on the interpretive technique to which it is submitted. Take, for example, an Italian constitutional provision referring to "statutes." Arguing *a contrariis*, one can conclude that such a provision applies to any kind of statute and only to statutes. Arguing by analogy, one can conclude that the provision at stake applies to statutes as well as to executive regulations (since both are "sources of law"). Arguing by the distinguishing technique, one can conclude that, since the class of "statutes" includes different subclasses (constitutional and ordinary, on the one hand; state and regional, on the other), the provision—in the light of its ratio—only applies to one of such subclasses. As a matter of fact, the set of interpretive methods (commonly accepted in the legal community at stake) is sufficient to warrant a great deal of competing results.

If the sceptical objection were compelling in general, then the picture provided by *communication theory* would be very unsatisfactory. In almost all legal cases, the interpretive methods would be unable to lead us to a right answer.

However, the objection can be answered (Moreso 1998, 147–156). Sinclair (2005–2006, 2006–2007, 2008–2009) is elaborating a very detailed and convincing counter-objection to Llewellyn's attack, showing that not all canons are always applicable, since the context of application determines the accuracy of a canon in individual cases, and consequently privileges some canons over others (Sinclair 2005–2006, 992):

[7] Sometimes scholars working in Critical Legal Studies (CLS) make use of this argument as a part of the defence of the so-called thesis of radical indeterminacy of law. See Kennedy (1976) and Singer (1984).

> A canon gets its status from the regularity of occurrence of its conditions of application and from the robustness of the reasoning associated with it. One might think of a canon as a conduit: it collects a set of applicable conditions and focuses them, producing a conclusion. Of course, if the applicable conditions do not obtain, then the canon will not apply and the conclusion will not follow. It is a mistake to apply a canon without paying attention to its required conditions of application; it is this mistake that often produces the appearance of antipathy between otherwise compatible canons.

Moreover, the interpretive methods should not be conceived as separate elements of analysis: they should be considered as parts of an integrated method we use to determine legal content. In the most important historical contribution to legal interpretation, Friedrich Karl Savigny distinguished four elements in every interpretation: grammatical, logical, historical, and systematic, and he added[8]:

> These elements are not four kinds of interpretation among which we could arbitrarily choose. On the contrary, they are four different operations which only jointly are able to interpret the legal statutes, even though in certain circumstances one of them could be more relevant than other. (Savigny 1840, 215; our translation)

So it seems fair to say that, when integrated and adapted to the context, methods and canons of interpretation cannot produce such pervasive indeterminacy as the sceptical objection claims.

3 The Communication Theory of Law

According to the communication theory of law, legal content is determined in the same way that the content of ordinary linguistic texts or utterances is (Soames 2013). On this view, in framing and voting for a written text, the lawmakers perform an illocutionary act such that the combination of their being endowed with de facto legal authority and of their expressing a substantive directive by means of a linguistic pronouncement by itself explains why their saying that so-and-so makes so-and-so law in their legal system. On the purest version of the theory, the only elements which we would need in order to identify the norms of a legal system and to provide a full explanation of their content could be given by an account of authority which told us who is in power, of the enactment procedure legally mandated in the system under consideration, of the enacted text, and of the intentions held by those who framed its provisions.

Moving from these insights, CT takes the principle of *epistemic asymmetry* that characterizes ordinary speech as straightforwardly applicable to the legal domain. Neale (2008, 22) presents *epistemic asymmetry* in the following way:

[8] This is Savigny's original German: "Mit diesen vier Elementen [grammatischen, logischen, historischen, systematische] ist die Einsicht in den Inhalt des Gesetzes vollendet. Es sind also nicht vier Arten der Auslegung, unter denen man nach Geschmack und Belieben wählen könnte, sondern es sind verschiedene Thätigkeiten, die vereinigt wirten müssen, wenn die Auslegung gelingen soll."

> The epistemic situations of producers (speakers and writers) and consumers (hearers and readers) are fundamentally asymmetric. The producer has a message he wants to get across to some other party (specific or non-specific). The producer airs a form of words he hopes will do the trick. By these words, aired in this fashion, on this occasion, the producer *meant* something. If the consumer succeeds in identifying what the producer meant, then the consumer *interpreted* the producer correctly.

Then, on the assumption that nothing special about the way laws are created prevents *epistemic asymmetry* from governing the relation between the parties to the "legal conversation," so to speak, the theory extends the parallel to the conclusion that the content of the law is what the lawmakers communicate. Further work would of course need to be done to accommodate those peculiar traits inhering in the structure of legal speech within the general model. This would presumably involve accounting for a kind of communication that is usually collective and carried out through strategic—non-cooperative—behaviour (Marmor 2011a), finding out who the speaker exactly is, and selecting, among the various intentions the speakers might have had in enacting the text, those that should count as relevant in shaping new law.

Up to this point, we have remained silent on what level of utterance content is regarded as relevant for determining legal content, according to CT. Three main views can be singled out in this respect: (a) what a speaker means by uttering a sentence X in a given context, (b) what a speaker says or asserts by uttering X on a given occasion, and (c) what a hearer would rationally take the speaker to have meant by uttering X on a given occasion.[9] There are at least two reasons why these distinctions matter. First, there is a way of meaning something without saying it—namely, by means of implicating it (Grice 1989) . Second, a reasonable reader may rationally take the author to have meant something she has not, or may take the author to have not meant something she actually has. Third, although more controversially, what is said by a speaker in uttering a sentence X may have not been meant by her.[10] And fourth, still controversially, what a rational hearer takes the speaker to have meant may differ from what she said.

The divergence between the (a), (b), and (c) variants of the theory reflects the differences manifested in accounting for the target of interpretation by the statements which we find in the works of e.g. Stephen Neale and Scott Soames. Theory (a) has been explicitly endorsed by Neale (2008, 2012), where he takes the view that what should be looked for in judicial adjudication is what the authors of the applicable provisions meant by their very enactments, as much as we try to retrieve what is meant by our interlocutors when we are engaged in ordinary conversation. Indeed, as his discussion of conversational implicatures in statutory enactments makes clear (Neale 2008), he thinks that insofar as lawmakers implicate something, the implicitly communicated proposition should be given authoritative status. A similar stance is taken by Soames (2008b), where he claims that "the *content of the law* includes everything asserted and conveyed in adopting the relevant legal text," though the

[9] We call theories (a), (b), and (c) the three versions of CT corresponding to each of these levels of meaning.

[10] Unless one agrees with Grice (1989, 87) that "'S (utterer) said that *p*' entails 'S did something *x* by which S meant that *p*.'"

position he endorses fails to make entirely clear whether, in cases where implicit and explicit content diverge, the correct interpretation should be faithful to the former or to the latter. Moreover, in a later paper (Soames 2013, 102), he seems to embrace yet another view—indeed a form of theory (c)—according to which the law is said to be "*what any reasonable person who understood the linguistic meanings of [the] words, the publically available facts, the recent history in the lawmaking context, and the background of existing law into which the new provision is expected to fit, would take* [the lawmakers] *to have meant*."

Given the complexities that emerge from the resulting landscape, it is worth pausing to consider the theory's exact import and implications.[11] Bearing CT's claim and unsettled facets in mind, we turn to the task of examining what it takes to be the correct model of legal interpretation, what role it assigns to interpretive arguments within it, and where it deems interpretation to run out and discretion to take over.

3.1 The Communication Theory of Legal Interpretation

We are now in a position to state what legal interpretation is according to the communication theory of law. It is the activity through which the interpreter forms a hypothesis about either (a),[12] (b),[13] or (c)[14]—the object depending on the preferred variant of CT—whenever needed to find out what (a), (b), or (c) exactly is.[15]

Similarly to what was noticed while dealing with the constitutive side of the matter, also on the epistemology of CT there is nothing special that distinguishes legal interpretation from utterance interpretation.[16] The evidential tools that may help the interpreter in identifying the content of the law may include canons of construction, background knowledge shared by the author and the reader, the linguistic and extra-linguistic context of the utterance, the conversational topic, previous remarks, lexical

[11] In doing so, we restrict the scope of the analysis to comparing theories (a) and (b), partly because it is still unclear what theory (c) exactly amounts to.

[12] Theory (a) is advocated by Neale (2008), defended by intentionalists in legal interpretation, and presupposed by Endicott (2012).

[13] Theory (b) is defended by Soames (2008b).

[14] Soames (2013).

[15] Marmor (1992), Endicott (2012), and Neale (2008) appear to agree (and we agree with them) that it is pointless to speak of "interpretation" when the content of the law is obvious; jurists usually express this thought with the maxim *in claris non fit interpretatio*. So it would probably be useful to employ a different term in reference to such cases (Marmor's and Endicott's usages converge on *understanding*).

[16] For simplicity, we are ignoring some aspects of legal practices (e.g. precedent) that have no counterpart in ordinary conversational contexts (while this is not to say that CT cannot account for them). However, it is still correct to say that, according to CT, since the principle of epistemic asymmetry holds in the legal sphere, too, the relation between interpretation and meaning here is the same as it is in conversational contexts. In this regard, it is also worth noting that there may be interesting parallels to draw between the phenomenon of overruling and that of retraction. Many thanks to Eliot Michaelson and Chiara Valentini for urging us to tackle this issue.

choices, the presumption that the speaker is operating in conformity with certain (Gricean or neo-Gricean) maxims (see Levinson 2000 and Horn 1995), and others. It is worth noting that the fact that anything useful may in principle be used as an aid to interpretation doesn't mean that the communication theorist cannot make sense of authoritatively prescribed restrictions on usable hermeneutic evidence. For instance, CT has no problem acknowledging that in some legal systems legislative history may be forbidden as a tool for figuring out what the law says, so long as this is taken to be an *epistemic* limitation which does nothing to alter the relevant constitutive relation between legislative action and legal content (however opaque or inaccessible such action and content may be).

So far we have been concerned with offering a schematic account of the metaphysics and epistemology of the communication theory of law. Now let us explore and assess how one legal argument, the canon of construction usually referred to by jurists as *expressio unius est exclusio alterius*, which we shall take as a case study, figures in that account.

Suppose a directive read

(p) The requirements set forth in statute ϕ must be fulfilled by the subjects in class A.

And suppose a court was asked to decide whether there are any requirements set out in legal texts other than ϕ which the subjects included in class A should fulfil. Imagine, further, that a previously enacted statute μ prescribed for the very same class the fulfilment of some additional requirements r_1, r_2. There are three interrelated aspects to this question: should we apply the *expressio unius* canon and exclude r_1, r_2 from the conditions applicable to A? What would the rationale for doing so be? Does answering to the first question involve interpretation or rather discretion?

Our aim in what follows is to answer them by the standards that follow from variants (a) and (b) of the communication theory of law. Each variant will be applied to each of three hypothetical scenarios, for this should enable us to see how they behave in handling these cases, if they behave differently from one another, and how they treat the *expressio unius* canon of construction in dealing with them.[17]

[17] Our description of theories (a) and (b) tracks the assumption, implicit in the work done by the proponents of these views, that the propositional content of *assertoric* uses of sentence tokens (assertive utterances, for short) is not determined in a way which is different from that of *prescriptive* ones. Therefore, we think it is fair to take (b) theorists to hold the view that the content of a conditional sentence used to make a legislative utterance is the same as the corresponding indicative conditional used to make an assertive speech act. Further, we assume the truth conditions of indicative conditionals to be those given by the truth table for the material conditional. The account of theory (b) we present owes its plausibility to the fact that, in general, Soames himself takes it for granted that the content of the law is whatever the lawmakers *assert* or *stipulate*—which means that at least he thinks lawmakers perform speech acts whose content is no different from that which they would have were the lawmakers instead making assertions; this, indeed, leads him to speak interchangeably of lawmakers asserting and stipulating things. Similarly, Neale (2008) does not distinguish what is said or implicated by speakers in ordinary conversation from lawmakers saying or implicating something in writing and voting for directives—at least not in the relevant sense that would make the content of (say) conditionals expressed in legislative utterances differ from those figuring in assertoric uses of language.

Scenario #1: In enacting *p*, the legislature lacks any sort of intention as to how ϕ relates to μ; the lawmakers are simply unaware that μ even exists.

Let us start by considering how theory (b) would handle this case. According to it, the content of the law is what is said by the lawmakers in enacting the provision from which the applicable legal norm is derived. In the case of *p*, the statement made (what is said) by the legislature is captured by the conditional

$p_{(b)}$ a requirement must be fulfilled by A if it is set forth in statute ϕ,

which is insensitive to whether the legislature has any further intentions (e.g. an intention to be implicating something beyond $p_{(b)}$). A related point is that the conditional would be true even if there were another statute or regulation prescribing for A the fulfilment of conditions other than those set out in ϕ. Therefore, $p_{(b)}$ would render ϕ compatible with μ, thus making room for the additional application of r_1 and r_2 to A. On this view, the interpretive outcome we would get from applying *expressio unius* to the interpretation of *p* in the case at hand is simply false and would accordingly have to be discarded. In answering to the third question, theory (b) claims the reasoning leading to this solution to be purely interpretive, thus making no use of discretion.

Turning to (a), absent the relevant intention behind the performance of the speech act in question, there would be no fact of the matter as to what the lawmakers meant in the crucial respect under consideration. Thus, the problem could not be solved by interpreting pre-existing legal directives, for we lack those facts about previous institutional behaviour that, had they obtained, alone would have settled the issue at hand. So the act of meaning in question would fall short of determining a unique correct result, and we would be left with a necessary recourse to discretion.[18] Accordingly, each of the available solutions—applying or blocking the application of the standards enshrined in μ—would have to be defended by recourse to a moral argument, and to say that in such cases, by arguing via *expressio unius*, we would still be in the business of interpreting pre-existing law would simply amount to deploying a juristic trick hiding the performance of disguised judicial lawmaking.

Scenario #2: We don't know what intentions the lawmaking body had in relation to μ in enacting *p*.

The result that theory (b) would yield in the present case coincides with that arrived at in Scenario #1, and for the same reasons: since the lawmakers' intentions are irrelevant, our ignorance of them is equally so. Turning to (a), the difference from Scenario #1 is that here, absent our knowledge of the legislators' intentions, it is

[18] It is a central feature of CT that if a text's *linguistically based content* (see Soames 2008b), i.e. a sentence's semantic content plus the speaker's intentions, fails to yield a single correct result for a given case, then the law's content is underdetermined with respect to that case. The fact that in his discussion of the open texture of language, and hence of law, in chap. VII of *The Concept of Law*, Hart appeared to embrace a similar view has sometimes led scholars to think that Hart's views on adjudication commit his theory of law to CT. That seems a stretch to us; for it seems implausible that Hart's central thesis—the doctrine of the rule of recognition—is compatible with the view that the relation between words and obligations is grounded in an account of the sort defended by CT (see Sect. 4.1–4.5).

the available evidence that, similarly to what ordinarily happens in many scientific inquiries, underdetermines the interpretive outcome. However, for all practical purposes no difference will result between Scenarios #1 and #2, since the conclusion to be reached in the latter is bound to be the same as we saw in the former.

> Scenario #3: In inscribing *p*, the legislature intended to create a legal norm prescribing that the sole conditions applicable to the subjects in A be those set forth in ϕ, thus intending to prevent r_1, r_2 from counting as fulfilment requirements relevant to them.

Given the irrelevance of the lawmakers' intentions in conveying a generalized conversational implicature through the enactment of *p*, (b) theorists would then reason from that assumption to the interpretive conclusion we found in dealing with Scenarios #1 and #2. For since the law is only what the lawmakers say, implicatures aren't part of it.[19] Further, the reasoning used in the current situation would be no less interpretive in character than that which was used in the previous ones, and so to apply the *expressio unius* canon to the case would be tantamount to committing an interpretive error. Theory (a), by contrast, views the legal norm obtaining in the present case as correctly captured by the biconditional,

> $p_{(a)}$ A requirement must be fulfilled by A if, and only if, it is set forth in statute ϕ,

that was implicitly communicated by the utterance of *(p)*,[20] for that is the proposition the author meant to convey. According to (a)-theorists, conversational implicatures can indeed be law, since for them the law is what the lawmakers mean, and since one way for a speaker to mean something is for her to implicate it. In the particular case at hand we are dealing with a generalized conversational implicature, a kind of implicature that is produced in especially systematic ways by speakers uttering sentences containing logical terms such as *and*, *or*, *if*, or *some*. Now, as we are dealing with *ifs*, that is, with cases involving a conditional's so-called "perfection" to the corresponding biconditional, what drives the hearer's non-deductive inference from the former to the latter is the presumption that the speaker is operating in conformity with certain maxims of conversation—the maxim of relevance and, perhaps, the second maxim of quantity (do not make your contribution more informative than is required)[21]—absent contextual elements to the contrary. In the legal case, the analogous interpretive outcome, consisting of $p_{(a)}$, would be reached by appealing to the *expressio unius* canon, whereas this would figure in support of an argument

[19] Theory (b) is not for this reason committed to denying that the propositions entailed by true propositions of law are also part of the law. Indeed, one property of implicatures, which distinguishes them from logical implications, is that they are cancellable without contradiction. So (b) theorists could consistently maintain that entailed law is part of the law while excluding implicatures from it.

[20] This way of putting the point is somewhat loose, for in order to be correct, one would have to say "that was implicitly communicated by the speaker in uttering *p*," since utterances do not implicate anything, nor do sentences, of course: speakers do. However, the context should make clear what we mean in framing the point in this way.

[21] The maxims of conversation were first formulated by Paul Grice (1989) and represent one of his greatest contributions to pragmatic theory and to the philosophy of language more generally.

to the conclusion that the list of the requirements included in statute ϕ is exclusive. On this view, *expressio unius* is an interpretive heuristic functioning as an inferential engine which takes the reader from the meaning of the text to the content the author most likely intended to convey through the text's enactment. It plays the role in legal interpretation that is played by the Gricean maxims in utterance interpretation. In conclusion, the reasoning in this scenario would be clearly interpretive, and the *expressio unius* canon is the only form of argument that could be correctly employed in resolving the interpretative problem at issue.

	Scenario #1	Scenario #2	Scenario #3
Theory (a)	Linguistic underdetermination	Evidential underdetermination	$p_{(a)}$
Theory (b)	$p_{(b)}$	$p_{(b)}$	$p_{(b)}$

	Scenario #1	Scenario #2	Scenario #3
Theory (a)	Discretion	Discretion	Interpretation: *expressio unius*
Theory (b)	Interpretation: not *expressio unius*	Interpretation: not *expressio unius*	Interpretation: not *expressio unius*

Time to draw the moral. First, we notice that the different views about the interpretative target held by (a) and (b) impinge both on the interpretive outcome they reach in the hypothetical scenarios we imagined and on the extent to which discretion is called for in solving the problems that arise.

Secondly, we see that according to both theories (a) and (b), *expressio unius* is not simply conceived of as one of the forms of argument that lawyers are entitled to invoke, and judges to appeal to, in interpreting the law. Rather, on both views, recourse to *expressio unius* is justified only insofar as it is conducive to the (level of) content that is regarded as legally valid by the theory. Generalizing on their treatment of this particular canon, one might say that according to them, a canon should only be used for interpreting the law if it makes the right prediction on the level of content that they posit as relevant. This exhibits a revisionary attitude towards legal arguments. While according to a common self-understanding of legal practitioners, arguments and canons function as strategic moves—either authoritatively licensed by positive law or conventionally accepted through the community's interpretive practice—in the language game of judicial disputes, CT sees them as interpretive heuristics functioning as presumptions that the legal content is displayed by the reading of the text according to their standards. A canon, then, would be no better as an interpretive heuristic than is the assumption that the lawmakers abided by its standards in framing and voting for a given provision. And this raises the question, what guarantees that they do in fact constitute good presumptions?

This question leads us to a third observation concerning the relations that theory (a) takes to hold between interpretation and meaning generally, and between the canons of construction and lawmaking in particular. According to theory (a), even though it is only an appropriately qualified subset of communicative intentions that constitutively determines speaker meaning, the genuine formation of those intentions is heavily conditioned by the speaker's expectations about the interpreter's capacity to recognize them, and hence about the likelihood of achieving illocutionary uptake or success in performing a given speech act (Neale 1992, 552). The point is clearly stated by Stephen Neale (2004, 77) :

> Despite the epistemic asymmetry, the perspectives of *A* [speaker] and *B* [hearer] are not independent. The asymmetry is *reciprocal* or *complementary* as in adjoining pieces of a jig-saw puzzle. In producing his utterance, *A* relies on what he takes to be *B*'s capacity to identify what he intends to convey; *B* assumes that *A* is so relying. And, possibly, so on. The ways in which *A* and *B* operate form a *dovetail joint* and are *mutually sustaining*. And to this extent, there is simply no possibility of making sense of *B*'s capacity to interpret *A* without making sense of *A*'s capacity to exploit that capacity, and vice versa.

The central idea is that in order for a speaker to mean a certain content—(say) the proposition *that q*—by uttering X, he must think that it will be possible for the hearer to recognize that content on the basis of X. This is why the formation of the speaker's meaning intentions is causally constrained by facts about his audience and the communicative context more generally. More precisely, in planning any given utterance, the speaker will take into account the hearers' capacity to figure out the message he tries to get across, and in so doing, he will rely on his own take on a number of features of the audience and of the context of utterance. Figuring among these is the speaker's estimation of the extent to which the hearer can be presumed to presume that the speaker will comply with certain norms guiding the conversation (e.g. Gricean maxims). In this way, through the speaker's presumption of the hearer's reliance on certain conversational norms—which in turn is the consequence of the hearer's presuming that the speaker is so reliant—the fact that the hearer follows certain interpretive norms indirectly affects the speaker's planning of his utterance and, consequently, the formation of his meaning intentions. The speaker's beliefs about the way in which the hearer will interpret his utterance will bear on his choice to utter X to communicate that *q*.

CT views the positions of the agents involved in the "legal conversation"—lawmakers and judicial interpreters—as analogously interdependent and accordingly posits a similar connection between lawmaking on the one hand and legal interpretation carried out through canons of construction on the other. In this vein, we may suppose that the more frequently a court relies on a given canon in interpreting the law, the more the legislators will (or at least should, assuming that they are rational agents) frame their provisions in a way that, if interpreted by the standards deriving from that canon, would deliver an outcome corresponding to the content they had meant to convey. This, in turn, will enhance the likelihood that judges appealing to that canon manage to capture what was meant by the lawmakers in enacting the provisions they enact, and this, in the end, will ensure that the canon is a good or reliable one.[22]

Having outlined the communication theory of legal interpretation at some length and having accounted for the role that in it is assigned to the hermeneutic maxim *expressio unius est exclusio alterius*, we shall now see what kinds of objections may be thought to undermine it.

4 Challenges to the Communication Theory

The motivating thought that underlies the communication theory is both simple and intuitively appealing. The idea is that legal texts are linguistic texts, and so the question of how the law is related to authoritative legal sources is "an instance of a more general question of what determines the contents of ordinary linguistic texts" (Soames 2008b, 403). Evidence that communication theorists have taken it to be obvious that the enactments of provisions in legislative processes are a special case of the ordinary utterances of sentences in conversational exchanges comes from considering how little argument has been put forward in order to defend the claim. Yet neither *simple* nor *intuitively appealing* means the same as "obvious." So let us scrutinize whether the relation between law and language is as the communication theorists take it to be and, if not, why. In what follows, we draw on Greenberg's insightful characterization of CT in order to show why its account of lawmaking isn't obvious, and we further develop an argument, partly inspired by recent work by Joseph Raz,[23] to the effect that the kind of metaphysical determination which CT supposes to be at work in the legal domain is more problematic than it seems.

[22] For an assessment of the project of analysing the relations between the legal agents taking part in the approval and interpretation of statutes along the lines of Grice's cooperative principle and maxims of conversation, see Chiassoni (1999) and Marmor (2011a, b). The analysis they offer is partly critical of the project, for they emphasize the strategic character of the agents' behaviour on three different levels of the legal context: (i) in legislative speech, between the legislators involved in the enactment process; (ii) in judicial interpretation, the "conversation" between lawmakers and judges often involves ongoing efforts on each side to redefine the division of labour between the judiciary and the legislative branch of government; and (iii) in adjudication, between the parties to the judicial dispute. However, it should be noticed that even if this were correct (which it probably is), it would not by itself pose an objection to the thesis that statutory implicatures are part of the law. This is so because even if the maxims were strategically violated by the agents involved—indeed, even if there was no such thing as a Gricean-like set of maxims in the legislative and judicial context—implicatures could still be produced, since the function of the maxims is epistemic, not constitutive, with respect to them. Endicott (2014) makes the related point that, while even in ordinary conversation agents happen to behave strategically, this is no ground for denying that they produce implicatures. Many thanks to Eliot Michaelson for discussion on these issues.

[23] Raz (2009a) defends the view that there can be interpretive pluralism in law—a thesis which, as we saw, contrasts with the communication theory of legal interpretation. Indeed, it is not surprising that the position Raz advocates is premised on the explicit denial of there being a close affinity between simple speech acts and laws (ibid. 300).

According to Greenberg, the core claim of the communication theory is that an authoritative pronouncement's semantic and pragmatic import bears directly on the content of the legal requirements stemming from the enacted text.

In a series of recent papers (Greenberg 2010, 2011a, b), Mark Greenberg has challenged CT's central thesis by trying to show that there is always a gap standing between the linguistic (semantic and pragmatic) import of any utterance performed by lawmaking actors and the content of the norms which obtain partly as a result of the utterance performed. The argument he advances (Greenberg 2011b) starts with the observation that "the communication theory moves from an understanding of what the legislature communicated to a thesis about the statute's contribution to the content of the law," and thereby claims that insofar as "a move from a text's *meaning* to the existence of certain legal *obligations* requires argument," no argument of the relevant sort can be given on the basis of linguistic considerations alone. The critical passage from what the lawmakers communicate to the statute's contribution to the content of the law[24] tends to be obscured by the lack of full articulation of the passage itself by the proponents of the theory. However, Greenberg points out, for the passage to be a viable one, a few crucial premises must hold, which he summarizes under the label of *explanatory directness thesis* (EDT):

> in the complete constitutive account of the obtaining of a legal norm: 1) the authoritativeness of the pronouncement is prior in the order of explanation to the obtaining of the legal norm; 2) the authoritativeness of the pronouncement is independent of the pronouncement's (specific) content and consequences; 3) there are no explanatory intermediaries between the authoritative pronouncement's being made and the norm's obtaining. (Greenberg 2011a, 44)

Together, these three elements articulate CT's account of the obtaining of legal norms. On this account, what is communicated is legally valid simply in virtue of being authoritatively pronounced; authoritative utterances uniquely determine legal content, and do so in such a way that the content of the law *is* the communicated content. The second thesis advocated by CT is what Greenberg calls the *linguistic content thesis* (LCT), according to which the content of the norms that obtain in virtue of the utterances of an authority must correspond to the meaning of the utterance made.

We can now recast CT's central claim in a better form:

> (M) For any authority A and any provision x, if A enacts x—by EDT—some legal norms $n_1,\ldots, n_n$ obtain such that—by LCT—the content of $n_1,\ldots, n_n$ is identical with some of the propositions expressed by A's utterance (inscription) of x.

A line of criticism that has been levelled against (M) is based on the multiplicity of candidate legal norms which any utterance may be thought to support. Greenberg (2011b) himself raises this objection: since many utterances express multiple contents, (M) generates indeterminacy by failing to identify a specific level of content as uniquely legally relevant. In appealing to the multiple levels of content which an

[24] This way of framing the issue is meant to remain neutral between (a), (b), and (c) with regard to the notion of "communicative content," and beyond statutes, it should be applicable to any kind of lawmaking performed through the enactment of provisions framed in a natural language (e.g. constitutions, regulations, directives).

utterance may be used to convey, what Greenberg has in mind is mainly (a), (b), and (c).[25]

Now, the natural reply on the part of the CT-theorist would be to select one particular level of utterance content as relevant, and thereby eliminate the indeterminacy in this way.

In what follows, we raise an objection to this reply strategy. To do so, we first carve out two consequences of (M) and then argue from premises concerning some essential features of law to their implausibility.

4.1 An Alternative Route for the Contribution Objection

Let us begin by highlighting some central aspects of law. A central feature of law is that facts about its content are partly determined by certain social practices. One way in which this is so is for acts of enactment by lawmakers to figure among the lower-level entities in virtue of which higher-level facts about the content of the law obtain.[26] In modern legal systems, the relevant social facts typically include facts about the lawmakers' actions, sayings, doings, and mental states. A consequence of this is that, on any plausible account, the subject with authority and the pronouncements it issues will play some role in explaining the content of the legal norms which are part of the system under consideration. But even then, we would still be in need of an answer to this question: What determines how lawmaking acts contribute to the content of the law? And at this point, CT's answer will prove unsatisfying. For consider the following two commitments of CT:

> *entailment*: if *a* is the authority in legal system L (at time t) and *u* the class of its utterances, all the truths about the content of the law in L (at t) are entailed by the truths about *a* and about the content of the utterances included in *u*.
>
> *invariance*: if *a* and *u* exist both in L and L_1, then the content of the law in L and in L_1 will be the same.

The point of these two principles is that they posit the *irrelevance* of the interpretive practices of the relevant legal community in explaining what contribution authoritative pronouncements make to the content of the law. For it doesn't matter what these practices are, since the explanation of the legal facts will be entirely dependent on factors other than them (thus *entailment*). And from this it would follow that a change in the criteria for successful interpretation used by a system's officials won't produce any change in the impact the system's legal sources have on the content of the law (thus *invariance*). Moreover, this also means that we could know what obligations there are in a given system without knowing what interpretive norms the system's officials are following. This, indeed, is precisely what enabled us (see Sect.

[25] On some occasions, he hints at semantic content and at the legislators' achievement intentions as other possible candidates.

[26] By *legal facts,* we mean facts about the content of the law in a given jurisdiction (at a given time).

3.1) to establish the status accorded by (a)- and (b)-theorists to the canon *expressio unius* without taking into account the role accorded to it by the legal agents acting in our hypothetical scenarios.

Now, even setting aside the fact that this upshot will look highly counter-intuitive to many legal practitioners, its trouble lies in its tension with the social dependence of law. For the law's social dependence is by no means exhausted by acknowledging that legal facts obtain in virtue of facts about the lawmakers' actions and mental states. Interpretative conventions matter no less than these.[27] And crucially, they are what gives rise to the web of principles that determine the contribution of legal sources (e.g. authoritative utterances) to the content of the law. This is the reason why, and the way in which, the social practice of interpretation and the interpretive arguments it embeds have both an epistemic and a partially constitutive nature. We can think of their twofold character by imagining a layered structure; on the surface, they are the key to our knowledge of laws, in that we use them to identify, or at least to form hypotheses about, the legal requirements that govern particular cases; at a deeper level, patterns of law-recognizing activities carried out by reference to certain interpretive rules and criteria determine the ways in which legal content facts come into existence as a result of institutional law-making actions and attitudes. Of course, legal interpreters do not (and could not) constitutively determine the meaning of the words and sentences they interpret; nor could they determine what the lawmaker uses those sentences to assert. But the point is different. It is that, in virtue of the customary rule that results from the practice of interpreting the law by reference to certain standards, interpreters constitute the true set of principles that map authoritative utterances (together with their linguistic and mental contents) to corresponding legal facts.

This line of thought is further supported and vindicated by a Hartian positivist account of law. According to it, in every possible legal system there is a conventional rule of recognition that sets out the criteria of validity for the system's rules. The criteria of validity give necessary and sufficient conditions for membership in the class of legally valid norms relative to a system at a given time, so that for every norm *n*, *n* is a valid norm in a system L at a time t if *n* satisfies the criteria set out by the rule of recognition followed by the officials in L at t (see Himma 2003). Since the existence and character of the rule of recognition are a function of the officials' joint practices, they will typically vary depending on the time and jurisdiction that is taken into consideration. To say that they are a function of the officials' practices means that they depend on the officials' convergent pattern of conduct, coupled with the adoption by the officials themselves of a critical reflective attitude towards the rule, an attitude that Hart called "acceptance." According to Hart's (1994, 103) own formulation of the relevant phenomenon, "to say that a given rule is valid is to recognize it as passing all the tests provided by the rule of recognition and *so as a rule of the system*." Further, he made clear that the rule of recognition was, according to him, "a form of judicial customary rule existing only if it is accepted and practiced in the law-identifying and law-applying operations of the courts" (1994, 256). Taking

[27] One point related to ours has been made in Bayón (2002).

these elements together, it may very well be that rules of recognition are what fills the gap between the contents of authoritative speech acts and the content of the law. And, if this is true, both *entailment* and *invariance* would then seem to be problematic.

5 Complex CT

What is shown by the argument we developed in the previous section is that there is no direct route from the lawmakers' acts of meaning to the content of the norms which hold in virtue of such acts. It may well be the case, for instance, that in a system L the courts pick out valid norms from the legal sources by reference to a rule according to which, say, the law is whatever the lawmakers mean in enacting the provisions they adopt. And, at the same time, it may also be that in L_1, as a matter of settled practice, the officials ascribe legal validity to only those propositions which the lawmakers explicitly assert—or to the literal meanings of the words and sentences they utter, or to the morally best reading of the text, for that matter. In such a case, it seems to us, assuming that the lawmakers in L and in L_1 have enacted the same texts and with the same intentions (and assuming that the judges in both systems have complete epistemic access to them), it would be wrong to regard the judges belonging to either L or L_1 as being mistaken about what the law requires. No doubt, it could be a contingent truth about L that the norms obtaining within it are made valid by the same rule of recognition followed by the agents in L_1. But it won't follow merely from the authorities in the two regimes having performed identical communicative acts, that the content of the law in the two systems is the same. In order for this to be the case, the further identity of secondary rules would be required, for these second-order rules, followed by the courts, are what determines what is to count as law within their system.

The considerations that we appealed to in objecting to CT have been utterly general, having to do with the essential features of law—its conventional character, structured as the union of primary and secondary rules (Hart 1994, Chap. V)—and with the way legal content essentially relates to facts about language—namely, through the conduit provided by secondary rules of recognition. These two elements set a constraint on any account of the relation between utterances and norms. It demands from any theory that it be about a *certain* system, and that it looks at the secondary practice of that community, if the relation it claims to be true of that system is to be correct. In other words, the constraint requires in general that the theory be partly parochial.

Perhaps driven by the need to accommodate this demand, Soames (2013) has recently recast the foundations for his theory of interpretation in a new shape. Two crucial innovations are of interest here: first, its scope is now taken to be limited to the US legal system; secondly, in order for the theory to be descriptively correct, the justification he advances makes appeal to a further ground. In Soames's own words (2013, 117):

> When legal interpretation is understood as the application of law to the facts of particular cases by authorized legal actors, one expects the task to be governed by legal rules that determine the responsibilities of those charged with it. [...] It is just such deeply and commonly accepted norms that, at bottom, constitute the authority of any system of laws.

Part of the difficulty involved in understanding this passage is due to the way we should read the term *responsibility*. If what is meant are the rules setting out judges' duties and the consequences, in terms of civil liability, deriving from their violation, then the expectation of finding legal rules on that matter will be quite reasonable and often satisfied. But it is unlikely that such norms could be those whose acceptance "constitutes the authority of any system of laws." Indeed, what he means by the "legal responsibilities" of judges is something quite different, which he states shortly afterwards:

> (OJR) Courts are not to legislate, but are to apply the laws adopted by legislative authorities to the facts of particular cases. To do so they must determine what the lawmakers in a given case asserted or stipulated in adopting the relevant legal texts, and apply that content to the facts of the case to arrive at a legal result. (Ibid.)

The first sentence is a concise statement of the judicial counterpart of the principle of legislative supremacy, the tenet according to which judges ought to act as faithful agents of the legislature when interpreting legal materials. The first part of the second sentence encapsulates the theory of interpretation Soames is arguing for—originalism—which is presented as having explicit content as its interpretive target.

There are several problems with OJR. On the one hand, while it is surely possible for there to be a legal rule of that sort, positively enshrined in some document and treated as binding by those who are subject to it, this is quite unlikely in general, and certainly not the case for the US system. In particular, while the principle of legislative supremacy may have a written formulation, originalism surely doesn't—it is the very object of contention, in and outside the courtrooms. In this respect, Soames's way out might be to concede that originalism is not a legal rule, strictly speaking, but rather the interpretative methodology most commonly used by judges. That would be a social rule, and, when invoked to justify the truth of theory (b), it would serve as the further ground needed to meet our ontological constraint. But is OJR so commonly accepted? This being an empirical claim, it should be tested against the body of available case law, of which we currently lack sufficient knowledge. However, we think there is room for doubt. Firstly, because matters of interpretive methodology tend to generate widespread disagreement, as is testified by a number of pivotal cases, some of which have been decided without respecting the originalist canon. Secondly, even if something along the originalist lines were shown to be the prevailing method, the specific version defended by Soames could hardly be the one commonly endorsed. This is due to the fact that the notion of *what is asserted*, as used by Soames and in the philosophy of language at large, is highly technical. So it is hard to expect from jurists that they appeal to it when arguing before the bench.

Perhaps one could try a different strategy. One may argue that originalism being the correct interpretative method in a given system is a direct consequence of the principle of legislative supremacy being in force in that system. That sounds like a more promising path, for it would seem to provide a powerful tool for dealing with disagreement. Few, in effect, would deny that the rule of separation of powers, which governs the division of labour between the lawmakers and the judiciary in our systems, requires some kind of legislative supremacy.

This line of reasoning is interesting, though perhaps too quick. Firstly, (a), (b), and (c) all involve some deference to original sources and, to this extent, no preference seems to be accorded by the rule itself to any of them in particular, which threatens to raise indeterminacy worries similar to those previously discussed. Secondly, the supremacy principle is not the only general tenet at the heart of our systems; others are fairness, equality, and the rule of law, plus those which belong to specific areas of law. Thus, even if OJR were supported by the principle of legislative supremacy, it may still be trumped by competing interpretative methods grounded in other principles.

Finally, the supremacy principle is itself open to interpretation. The problem here somehow resembles the situation Hart portrays in the last section of Chap. VII of *The Concept of Law*: uncertainty in the rule of recognition. One can imagine a regime where the courts regarded "whatever the parliament enacts is law" as the ultimate criterion for the identification of valid rules. Without further agreement on the details, we would still have to address the question: "Which propositions are made true by the parliament's enactment?" As Hart noticed, the answer to be given and the extent to which it is determinate will typically vary depending on the time and place where the question is asked. And similarly, the interpretation of the principles that would guide the interpretive process, including legislative supremacy, could itself be subject to such variations. Surely, in our systems there are limits to how the courts may truly interpret the enacted text, for, at a minimum, correct interpretations will have to be grounded in some feature of the text itself—the communicative intentions behind it or the content of its words and phrases. And still, since the rule of recognition itself can be open-textured, at its margins the question may lack a settled answer.

6 Conclusion

The aim of this paper has been to explore the communication theory of law in some detail, to flesh out its claims and implications. In doing so, we used the canon of construction *expressio unius* and the role accorded to it by CT as a way to shed light on the theory itself and on the role of interpretive arguments in legal interpretation.

The conclusion we reached resonates, we think, with the view expressed by Rosen (2011, 133–134) in the following passage:

> If textualism or intentionalism or any other determinate view of these matters is correct as an account of any part of the law—Bolivian criminal law, for example, or US constitutional law—this will be thanks to special features of the relevant field of law and to contingent features of the legal regime in question. To a significant degree, the true principles connecting the linguistic features of legal utterances to their legal effects are a matter of positive legal fact, grounded in the legal history and practice of the community.

In this paper, we tried to take this view seriously, and to show that a strong version of the communication theory of law is actually incompatible with it.

If our claim is correct, the philosophy of language will be extremely useful to legal theorists and interpreters in providing them with a sound and linguistically adequate framework for detecting what candidate relations between different levels of utterance content and the content of the law there are. However, legal theory will still have to specify what relation is true of a given system (at a given time), and what are the facts responsible for the holding of such relations.[28]

References

Alexander, L. 1998. The banality of legal reasoning. *Notre Dame Law Review* 73: 517–533.

Bayón, J.C. 2002. Derecho, convencionalismo y controversia. In *La relevancia del Derecho: ensayos de filosofía jurídica, moral y política*, ed. J.P. Navarro, and C. Redondo, 57–92. Barcelona: Gedisa.

Borges, J. L. 1999. *'The False Problem of Ugolino', nine dantesque essays 1945–1951.* In *Selected non-fictions*, ed. E. Weinberger, trans. E. Allen, S. J. Levine, and E. Weinberger. London: Penguin Books.

Cappelen, H. 2009. The creative interpreter: Content relativism and assertion. In *Philosophical perspectives. Philosophy of language*, vol. 22, ed. Hawthorne. 23–46. Hoboken: Wiley-Blackwell.

Chiassoni, P. 1999. Interpretive games: Statutory construction through gricean eyes. In *Analisi e Diritto. Ricerche di giurisprudenza analitica,* ed. P. Comanducci and R. Guastini. 79–99. Torino: Giappichelli.

Dworkin, R. 1986. *Law's empire*. Cambridge, Mass.: Belknap Press of Harvard University Press.

Endicott, T. 2012. Legal Interpretation. In *The Routledge companion to philosophy of law*, ed. A. Marmor, 109–122. London: Routledge.

Endicott, T. 2014. Interpretation and indeterminacy. *Jerusalem Review of Legal Studies Jerusalem Review of Legal Studies* 10: 46–56.

Green, L., and B. Leiter (eds.). 2011. *Oxford studies in philosophy of law*, vol. 1. Oxford: Oxford University Press.

Greenberg, M. 2004. *How Facts Make Law. Legal Theory* 10: 157–198.

Greenberg, M. 2010. The communication theory of legal interpretation and objective notions of communicative content. In *UCLA school of law working paper series, public law & legal theory working paper*, 10–35. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1726524.

Greenberg, M. 2011a. The standard picture and its discontents. In *Oxford studies in philosophy of law*, vol. 1, ed. L. Green, and B. Leiter, 39–106. Oxford: Oxford University Press.

Greenberg, M. 2011b. Legislation as communication? Legal interpretation and the study of linguistic communication. In *Philosophical foundations of language in the law*, ed. A. Marmor, and S. Soames, 217–256. New York, NY: Oxford University Press.

Grice, H.P. 1989. *Studies in the way of words*. Cambridge, Mass.: Harvard University Press.

Guastini, R. 2011. Rule-Scepticism Restated. In *Oxford studies in philosophy of law*, vol. 1, ed. L. Green, and B. Leiter, 138–161. Oxford: Oxford University Press.

Hart, H.L.A. 1994. *The concept of law*. Oxford: Oxford University Press, 1st ed. 1961.

Himma, K.E. 2003. Making sense of constitutional disagreement: Legal positivism, the bill of rights, and the conventional rule of recognition in the united states. *Journal of Law in Society* 4: 149–218.

[28] Thanks to Guglielmo Feis, Dan López de Sa, Eliot Michaelson and Chiara Valentini for helpful comments and discussions on previous drafts of the paper.

Horn, L.R. 1995. Vehicles of meaning: Unconventional semantics and unbearable interpretations. *Washington University Law Quarterly* 73: 1145–1152.

Kennedy, D. 1976. Form and substance in private law adjudication. *Harvard Law Review* 89: 1685–1778.

Levinson, S.C. 2000. *Presumptive meanings: The theory of generalized conversational implicatures*. Cambridge, Mass.: The MIT Press.

Llewellyn, K.N. 1950. Remarks on the theory of appellate decision and the rules or canons about how statutes are to be construed. *Vanderbilt Law Review* 3: 395–406.

Marmor, A. 2005. *Interpretation and legal theory*. Portland: Hart Publishing.

Marmor, A. 2011a. Can the Law Imply more than it Says? On Some Pragmatic Aspects of Strategic Speech. In *Philosophical foundations of language in the law*, ed. A. Marmor, and S. Soames, 83–104. Oxford: Oxford University Press.

Marmor, A. 2011b. *Philosophy of law*. Princeton, NJ: Princeton University Press.

Marmor, A. (ed.). 2012. *The Routledge companion to philosophy of law*. London: Routledge.

Marmor, A. 2013. Truth in law. In *Law and language*, *current legal issues*, vol. 15, ed.

Marmor, A., and S. Soames (eds.). 2011. *Philosophical foundations of language in the law*. Oxford: Oxford University Press.

Moreso, J.J. 1998. *Legal indeterminacy and constitutional interpretation*. Dordrecht: Kluwer.

Neale, S. 1992. Paul Grice and the philosophy of language. *Linguistics and Philosophy* 15: 509–559.

Neale, S. 2004. This, That, and the Other. In *Descriptions and beyond*, ed. A. Bezuidenhout, and M. Reimer, 68–82. Oxford: Oxford University Press.

Neale, S. 2008. Textualism with intent. Excerpt from manuscript for discussion at Oxford University, 2008.

Neale, S. 2012. Convergentism & the nature of law. In *Draft for the Oslo Workshop on the Pragmatics of Legal Language*.

Perry, J. 2011. Textualism and the Discovery of Rights. In *Philosophical foundations of language in the law*, ed. A. Marmor, and S. Soames, 105–129. Oxford: Oxford University Press.

Pound, R. 1922. An introduction to the philosophy of law. New Haven: Yale University Press. http://oll.libertyfund.org/?option=com_staticxt&staticfile=show.php%3Ftitle=2222&chapter=208866&layout=html&Itemid=27.

Raz, J. 1995a. *Ethics in the public domain*. Oxford: Oxford University Press.

Raz, J. 1995b. On the autonomy of legal reasoning. In *Ethics in the public domain*, ed. J. Raz, 310–324. Oxford: Oxford University Press.

Raz, J. 2009a. *Between authority and interpretation*. Oxford: Oxford University Press.

Raz, J. 2009b. Interpretation: Pluralism and innovation. In *Between authority and interpretation*, ed. J. Raz, 299–322. Oxford: Oxford University Press.

Rosen, Gideon. 2011. Textualism, intentionalism, and the law of the contract. In *Philosophical foundations of language in the law*, ed. A. Marmor, and S. Soames, 130–164. Oxford: Oxford University Press.

Savigny, F.K. von. 1840. *System des heutigen römischen Rechts*. Berlin: Band I. http://dlib-pr.mpier.mpg.de/m/kleioc/0010/exec/books/%22199236%22.

Sinclair, M. 2005–2006. 'Only a Sith thinks like that': Llewellyn's 'dueling canons', one to seven. *New York Law School Law Review* 50: 919–992.

Sinclair, M. 2006–2007. 'Only a Sith thinks like that': Llewellyn's 'dueling canons', eight to twelve. *New York Law School Law Review* 51: 1003–1056.

Sinclair, M. 2008–2009. 'Only a Sith thinks like that': Llewellyn's 'dueling canons', pairs thirteen to sixteen. *New York Law School Law Review* 53: 953–999.

Singer, J.W. 1984. The player and the cards: Nihilism and legal theory. *The Yale Law Journal* 94: 1–70.

Soames, S. 2008a. *Philosophical essays*, vol. 1. Princeton, NJ: Princeton University Press.

Soames, S. 2008b. Interpreting legal texts: What is, and what is not, special about the law. In *Philosophical essays*, vol. 1, ed. S. Soames, 403–423. Princeton, NJ: Princeton University Press.

Soames, S. 2011. Toward a theory of legal interpretation. *NYU Law School Journal of Law and Liberty* 6: 231–259.

Soames, S. 2013. Deferentialism: A post-originalist theory of legal interpretation. *Fordham Law Review* 82: 101–122.

Statutory Interpretation as Argumentation

Douglas Walton, Giovanni Sartor and Fabrizio Macagno

1 Introduction

Interpretation is regarded as the passage from a legal text to a legal rule (Hage 1996, 214; Tarello 1980), namely a normative premise under which an individual case is "subsumed" or classified (see Moreso and Chilovi, chapter 2, part III, this volume, on "Interpretive Arguments and the Application of the Law"). This passage can be compared to the common understanding and processing of utterances in ordinary conversation (Smolka and Pirker 2016), in which semantic content is only a vehicle for getting to the "speaker's meaning" or what is communicated—a richer content "to which meaning and obvious background assumptions have both contributed" (Soames 2008, 411; see also Butler 2016; Carston 2013; Horn 1995; Miller 1990). Legal interpretation does not differ essentially from ordinary interpretation, even though legislative speech is one-sided (there is nobody who can immediately answer back) and the basic presumption governing such texts is that the author used the language to convey ideas (Sinclair 1985, 390). However, pragmatic principles constitute a dimension of rationality which is necessary for the understanding of

D. Walton
University of Windsor, Centre for Research in Reasoning,
Argumentation and Rhetoric (CRRAR), Windsor, ON, Canada
e-mail: waltoncrrar@gmail.com

G. Sartor
Dipartimento di Scienze Giuridiche, Università di Bologna, Bologna, Italy
e-mail: giovanni.sartor@gmail.com

G. Sartor
European University Institute, Florence, Italy

F. Macagno (✉)
IFILNOVA, Instituto de Filosofia da Nova, Universidade Nova de Lisboa,
Lisbon, Portugal
e-mail: fabrizio.macagno@fcsh.unl.pt

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_18

legal texts (Sinclair 1985, 401). As Soames puts it, the statutory language provides incomplete semantic content, which needs to be completed by pragmatic (contextual) factors and processes:

> Just as what I say, and commit myself to, by uttering a sentence, is often a function of more than its semantic content, so "what the law says," and is committed to, is often a function of more than the semantic contents of relevant legal texts. Just as you have no standing to reinterpret my remark to conform to your moral and political views, simply because the meaning of my sentence doesn't fully determine the content of my remark, so judges applying the law have no standing to reinterpret it, simply because the linguistic meanings of the relevant legal texts don't fully determine the content of the law. There are other principles at work filling the gap between sentence meanings and the contents of texts, legal or otherwise (Soames 2008, 404).

In pragmatics, the reconstruction of meaning in ordinary conversation is regarded as characterized by both default reasoning and systematic and critical inferences (Jaszczolt 2005, 46; Wilson 2005). Default inferences are triggered when information about the current context is absent or not necessary for comprehension (i.e., when the inferential conclusion is not in conflict with the present context). When default inferences cannot be drawn (Kecskes 2008, 2013, 129, 131; Kecskes and Zhang 2009), more complex inferences need to be made. In legal theory, this twofold process is mirrored by the concepts of understanding and interpretation. Interpretation is defined as "an ascription of meaning to a linguistic sign in the case its meaning is doubtful in a communicative situation, i.e., in the case its "direct understanding" is not sufficient for the communicative purpose at hand" (Dascal and Wróblewski 1988, 204). In case there is an "eventual 'mismatch' between the 'computed' utterance-meaning and some contextual factor" resulting from the background or the specific case to which the law is applied (Dascal and Wróblewski 1988, 213, 216), interpretation needs to be justified through reasons (Atlas 2008; Atlas and Levinson 1981; Dascal 2003, 635).

This chapter addresses the problem of representing and assessing the reasons provided in favor of a specific interpretation and more precisely justifying why and how an interpretation is more acceptable than others (Macagno 2017). At this functional level, such interpretive reasons are regarded as arguments (Macagno and Capone 2016) aimed at showing why a particular rule, rather than another, is valid on the basis of the statutory text (Hage 1996, 215). In statutory interpretation, such arguments are usually analyzed using specific maxims of interpretation, which can be translated into a formal language (Hage 1997). In this chapter, we will show how the canons of interpretation can be represented as schemes, namely patterns of defeasible argument advanced in support of the interpretation of a text (or part thereof). This formalization can be then used to bridge the gap between legal interpretation and argumentation theory, and more specifically the argumentation schemes used for representing and evaluating natural arguments (Macagno and Walton 2015; Walton et al. 2008).

The functional analysis of legal interpretation in terms of arguments and the formalization of the interpretive arguments as schemes (advanced in Sects. 2 and 3) allows modeling legal interpretation combining the formal argumentation system ASPIC+ with a logical language (Sartor et al. 2014). After introducing the Carneades Argumentation System (Sect. 4) and applying it to two cases (Sects. 5 and 6), Sects. 7–10 will be devoted to developing a logical model for reasoning with interpretive canons, conceived as defeasible rules (see Sartor, chapter 3, part II, this volume, on "Defeasibility in Law"). The logical structure that will be developed will not be framed in deontic terms, but rather will concern terminological assertions concerning what should count as the best interpretations of the contested or potentially contested expressions.

2 Interpretive Arguments

The justification of an interpretation can be regarded as an argumentation-based procedure in which the best interpretation is the one supported by the strongest or more justified arguments (Atlas and Levinson 1981; Macagno et al. 2018). On this perspective, the "canons" or maxims of interpretation can be reframed as arguments (Macagno and Walton 2017), which can be classified according to their communicative purpose and the types of warrants. This classification allows detecting the relationship between interpretive canons and the schemes commonly used in argumentation theory.

2.1 The Existing Types of Interpretive Arguments

Macagno et al. (2012) compiled a list of eleven interpretive arguments identified by MacCormick and Summers (1991). Below, each type of argument recognized in that prior list is explained in a condensed manner to give the readers some idea of how each of them can be reconfigured as a distinct defeasible form of argument.

- *Argument from ordinary meaning* requires that a term should be interpreted according to the meaning that a native speaker would ascribe to it.
- *Argument from technical meaning* requires that a term having a technical meaning and occurring in a technical context should be interpreted in its technical meaning.
- *Argument from contextual harmonization* requires that a term included in a statute or set of statutes should be interpreted in line with whole statute or set.
- *Argument from precedent* requires that a term should be interpreted in a way that fits previous judicial interpretations.
- *Argument from statutory analogy* requires that a term should be interpreted in a way that preserves the similarity of meaning with similar provisions of other statutes.
- *Argument from a legal concept* requires that a term should be interpreted in line with the way it has been previously recognized and doctrinally elaborated in law.

- *Argument from general principles* requires that a term should be interpreted in a way that is most in conformity with general legal principles already established.
- *Argument from history* requires that a term should be interpreted in line with the historically evolved understanding of it.
- *Argument from purpose* requires that a term should be interpreted in a way that fits a purpose that can be ascribed to the statutory provision, or whole statute, in which the term occurs.
- *Argument from substantive reasons* requires that a term should be interpreted in line with a goal that is fundamentally important to the legal order.
- *Argument from intention* requires that a term should be interpreted in line with the intention of the legislative authority.

These eleven types of interpretive argument are comparable to and overlap with the fourteen types previously identified by Tarello (1980, Chap. 8), listed as follows in Sartor et al. (2014):

- *Arguments a contrario* rejects interpretations of a term departing from the term's literal meaning.
- *Analogical arguments* support interpretations according to which the meaning of a term or expression of a legal provision is extended to apply a rule to a case not regulated by the given provision (it is included in neither the core nor the periphery of its application area), but presenting a relevant similarity with the cases covered by it (Damele 2014; Gray 2013, 35).
- *Arguments a fortiori* support interpretations according to which the meaning of a term or expression in a legal provision is extended to apply that provision a case that is not regulated by such a provision (it is included in neither the core nor the periphery of the application area of the provision in question), but deserves, to a higher degree, the same discipline as the cases covered by it.
- *Arguments from completeness of the legal regulation* exclude interpretations that create legal gaps.
- *Arguments from the coherence of the legal regulation* exclude interpretations of different legal statements that make them conflicting.
- *Psychological arguments* support interpretations driven by the actual intent of the authors of legal text.
- *Historical argument*s support interpretations giving a legal statement the same meaning that was traditionally attributed to other statements governing the same matter.
- *Apagogical arguments* exclude interpretations that generate absurdities.
- *Teleological arguments* support interpretations contributing to a purpose pertaining to the goals or interests that the law is supposed to promote.
- *Non-redundancy arguments* exclude interpretations that would make the interpreted expression redundant, under the assumption that the legislator does not make useless normative statements.
- *Authoritative arguments* support interpretations already given by authoritative courts or scholars.

- *Naturalistic arguments* support interpretations aligning a legal statement to human nature or the nature of the matter regulated by that statement.
- *Arguments from equity* support (exclude) (un)fair or (un)just interpretations.
- *Arguments from general principles* support (exclude) interpretations that are supported by (incompatible with) general principles of the legal system.

The two lists complement each other, even though Tarello's list emphasizes the kinds of input on which interpretive argument is based, such as ordinary language, technical language, and so forth, while MacCormick and Summers' list emphasizes the reasoning steps involved in the interpretive process.

In comparing the two lists of types of interpretive arguments, some common elements stand out, but there are also significant differences. Some of the argument types in Tarello's list—such as analogical arguments, teleological arguments, and arguments from general principles—appear to be already included in the list of MacCormick and Summers. Tarello's psychological arguments seem to fit under McCormick and Summers' category of argument from intention. It looks like Tarello's authoritative arguments might fit under MacCormick and Summers' category of argument from precedent. Others types of argument are distinctively different, while in still other cases it is unclear how the type of interpretive argument described in the one list is related to the type described in the other list.

One of the crucial problems concerning types of interpretive arguments is their use (in training legal practitioners or scholars) and their relations with the works in argumentation theory and logic on argument analysis and reconstruction. Recently, the canons or maxims that express the general principle characterizing each type of argument have been represented as defeasible rules, to be integrated within a prioritized defeasible logic system (Rotolo et al. 2015). The purpose of this chapter is to analyze types of interpretive arguments as argumentation schemes, or rather dialogical patterns of arguments, in which an interpretation is regarded as a defeasible viewpoint that needs to be supported by a pattern of reasoning and can be subject to default in case specific critical questions are successfully advanced. On this perspective, interpretive reasoning is framed within a broader dialectical framework, involving a specific burden of bearing out and defeating a specific interpretation (Gizbert-Studnicki 1990).

Some of the interpretive argumentation schemes in both lists clearly relate to argumentation schemes already widely known and studied in argumentation that are not specifically designed to deal with interpretive issues (Macagno and Walton 2015; Walton et al. 2008). Hence, there are many questions about how some of the new interpretive schemes relate to these more general schemes that have been already widely recognized. For example, the category of authoritative arguments in Tarello's list might relate to scheme for argument from expert opinion. Since laws formulated in statutes are binding on the courts, it can be said that the statement made in this context can be held to hold by reason of authority. But a legal scheme for argument from administrative authority that is a variant on argument from authority already has some recognition in the field of argumentation studies. Hence, there are questions raised about how this new interpretive scheme proposed by Tarello distinguishes between the two kinds of argument from authority. As mentioned above, there is also the question of how Tarello's version of interpretive argument from authority fits in with schemes from MacCormick and Summers' list such as argument from precedent, argument from a legal concept, argument from general principles, and argument from history. None of these questions can be discussed in this chapter, for reasons of length, but they need to be recognized here as problems for future research.

Another similar problem is how the interpretive argument from precedent, as it is called in MacCormick and Summers' list, is related to the general scheme for argument from precedent, already recognized in the argumentation literature. The problem is that there are great divisions of opinion on precisely how the scheme should be modeled. Many think that argument from precedent is always based on argument from analogy, that is, on a comparison between and source case and a target case. But others might think that legal argument from precedent needs to be based on *ratio decidendi*. Another question raised by this difference of opinion is whether *ratio decidendi* represents some kind of analogy between the two cases where the rationale used to arrive at the conclusion in the source case is supposed to be similar to a comparable rationale that can fit the target case.

In this chapter, we recognize the existence of these problems without delving into a detailed analysis thereof, so that we can forge ahead with building a framework for interpretive argumentation schemes that can later be applied to studying specific schemes and issues. The starting point is to provide a general classification of the most important arguments of the two lists, identifying the more generic identities between them. Then, we move through a sequence of examples of legal arguments where interpretation of a statute or law is an issue, applying the model to the examples. As always, the work of applying formal structures to real cases of argumentation in natural language discourse raises problems and difficulties in its own right.

2.2 Classifying Interpretive Arguments

MacCormick (2005, 124–125) proposed that there are three main categories of interpretive argument, over the above eleven categories of interpretive arguments acknowledged as persuasive in grounding a selected interpretation of a text in a disputed case in a broad variety of legal systems. First, there are so-called *linguistic arguments* that appeal to the linguistic context itself to support an interpretation (which we can call definitional arguments, Macagno and Walton 2014). Second, there are the *systemic arguments* that take the special context of the authoritative text, within the legal system, into account. Such schemes merge the authority of the source with the reconstruction of the definition from the text. Third, there are the *teleological–evaluative arguments* that make sense of the text in light of its aim or goal (which we can refer to as pragmatic arguments, see Macagno and Walton 2015). A fourth category is what McCormick (2005) calls "*appeal to the lawmaker's intention*." McCormick does not consider this type of interpretive argument alongside the other main categories of interpretive argument, because of the ambiguity and indeterminacy of the notion of intention. He rather views it a trans-categorical type of argument that ranges across all the other categories and their types, as linguistic, systemic or teleological–evaluative considerations can support the attribution of intentions to legislators.

If we try to analyze the lists of arguments in terms of patterns of argument, explaining the arguments of legal interpretation using the categories of argumentation schemes, we need to draw a first crucial distinction between arguments that support an interpretation and arguments that reject an interpretation. Some interpretive canons, however, are bivalent, in the sense that they provide for two interpretive schemes: one (positive or negative) when the canon's condition is satisfied, and the opposite (negative or positive) when the canon's condition is not satisfied. For instance, while the contextual coherence of an interpretation supports the adoption of an interpretation, lack of contextual coherence supports rejection. In such cases, we use the symbol **+** and **–** to denote the use of a scheme to support and reject an interpretation, for instance **+** contextual coherence and **–** contextual coherence.

The arguments supporting an interpretation are different in nature (Macagno 2015). Pragmatic arguments, definitional arguments (of different types, including the systemic ones), and analogical arguments represent distinct reasoning patterns, which are often merged with authority arguments. Such arguments are intended to back up a specific definition based on previous interpretations (epistemic authority) or on the reconstruction of a possible "intention" of the lawmaker (deontic authority), or on the alleged "nature" of a concept (the commonly shared definition). Such categories often merge with each other, but they can be classified in Fig. 1 based on a distinctive feature, namely their distinctive reasoning pattern.

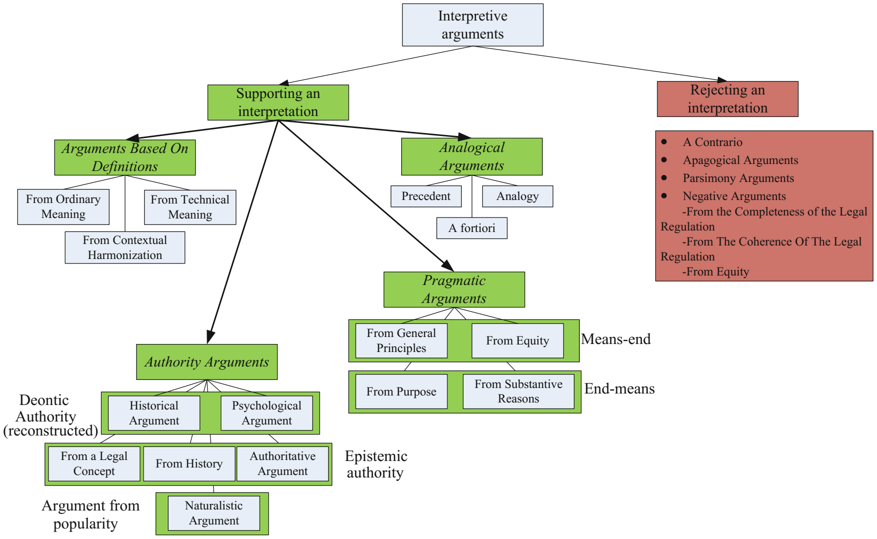


Fig. 1 Classifying the arguments of interpretation

It was recognized by MacCormick (2005) that there can be conflicts between interpretive arguments, pitting one form of interpretive argument against another (Rotolo et al. 2015). Some legal traditions provide general criteria for dealing with conflicts of this sort based on certain kinds of priorities. Alexy and Dreier (1991, 95–98) have cited criteria such as the following: (a) In criminal law, arguments from ordinary meaning have priority over arguments from technical meaning; (b) in criminal law, generic arguments based on the intention of the legislator have priority over arguments not based on authority, but not over linguistic arguments. In this chapter, we will use argumentation tools to represent such conflicts and priorities.

3 Translating Interpretive Arguments into Schemes

The classification of interpretive arguments can be the starting point for translating the arguments (and canons or maxims) into formal (or rather, quasi-formal) schemes representing how a conclusion is supported by premises. In particular, we will provide the schemes for the two general categories (positive versus negative) and the definition-based arguments (in particular, from ordinary and technical meaning). These schemes will be the ground for the further formal representations in Sects. 4, 5 and 6 and the logical formalization in the remaining sections.

3.1 Assumptions and Common Template

Statutes are written in natural language. Our concern is with the interpretation of sentences expressed in natural language that are susceptible to differing interpretations (Atlas 2005; Horn 1995). The major philosophical concern is how the notion of meaning is to be defined in relation to the task of finding the evidential basis for preferring one interpretation or another (Atlas 2005; Atlas and Levinson 1981; Dascal 2003, 635). In this chapter, we find it most highly suitable to adopt a pragmatic approach to meaning, namely to understand statutory meaning as the intention expressed through the legal text (Carston 2013), an approach that corresponds to the trans-category understanding of interpretation in McCormick (2005). The syntax representing the structure of a sentence, as well as the individual semantic meanings of each term contained in the sentence, are important. But over and above such factors, it needs to be acknowledged that the meaning of the sentence composed of these elements, especially in the examples considered in this chapter, needs to be placed in the context of a broader text or corpus in which it is embedded. For example, the issue of whether a contested word should be taking it as expressing and ordinary meaning or a technical meaning is a dispute about whether the word can be interpreted the one way or the other in a special context of use. For these reasons, although we acknowledge the importance of semantics and syntax in matters of statutory interpretation, we need to study the notion of meaning in a broad manner to include not only these aspects, but also the aspect of the placement of the sentence in a broader context of use in different kinds of discourse.

From our perspective, making an interpretation consists in associating a linguistic occurrence and a meaning within a specific context and use, i.e., in claiming that a certain expression E in certain document D has a certain meaning M. Interpretations are not necessarily correct. They may be right or wrong, preferable or not to other interpretations.

We shall model the application of interpretation canons by using a uniform template, so that for each canon we obtain an argument scheme including a major premise, a minor premise, and an interpretive conclusion.

- The major premise is a general canon: If interpreting an expression (word, phrase, sentence) in legal document (source, text, statute) in a certain way satisfies the condition of the canon issue, then *the expression* should/should not be interpreted (depending on whether the canon is a negative or positive one) in that way.
- The minor premise is a specific assertion: Interpreting an expression in a particular document in a certain way satisfies the condition of the canon.
- The conclusion is a specific claim: The expression in that document indeed should/should not be interpreted in that way.

In this chapter, we shall apply this template to provide schemes for the following canons: (1) argument from ordinary language (*OL*); (2) argument from technical language, whose requirement is correspondence to technical language (*TL*); (3) *a contrario* argument (*AC*); (4) argument from purpose (*Pu*); (5) argument from precedent (*Pr*); (6) argument from contextual harmonization (*CH*). This list of schemes will be added to as new schemes are formulated. Here is our system of notation for labeling the nodes in an argument diagram to indicate a scheme. We use + for schemes uses to argue for an interpretation, – for schemes used to argue against an interpretation, +e for exclusion, and +i for inclusion. Hence, we put +e as the use is in favor of exclusion (for the exclusionary conclusion). In Carneades, + indicates an argument in favor of its conclusion, so if the conclusion is exclusionary, it should be +e. So, for example, the notation *+iPr* labels a pro argument from inclusive argument from precedent.

3.2 Positive Interpretive Schemes

As mentioned above, two fundamental macro-categories of interpretive argument schemes need to be distinguished, the positive ones supporting an interpretation and the negative ones rejecting an interpretation. Here is the template for positive interpretive argument schemes. In presenting this template, we shall use uppercase letters for variables and lowercase letters for constants:

Major premise	*C*: If the interpretation of *E* in a *D* as *M satisfies C's condition*, then *E* in *D* should be interpreted as *M*
Minor premise	The interpretation of *e* in *d* as *m satisfies C's condition*
Conclusion	*e* in *d* should be interpreted as *m*

In applying this template, we need to substitute in the major premise the condition that characterizes a canon, for instance, fitting *ordinary language* (OL).

In order to show how positive interpretive canons can be applied with this pattern, we use the case of *Dunnachie v Kingston-upon-Hull City Council*, also used by MacCormick (2005), as a running example. This case concerns an employee who claimed to have been unfairly dismissed, and as a result to have suffered humiliation, injury to feelings and distress. The employer argued that the relevant section of the current UK legislation, called the Employment Rights Act of 1996, only permits recovery of *financial loss*. The employee argued that a proper construction of all the relevant section of the statute allows for recovery of *losses* other than financial losses narrowly construed. The question posed was whether the term "*loss*," as used in the statute, referred only to financial loss or could be given a more extended meaning so that it included losses such as emotional loss that are not strictly financial.

If we use the canon *Ordinary Language*, we obtain the following structure:

Major premise	*OL*: If The interpretation of *E* in *D* as *M* fits *ordinary language*, then *E* in *D* should be interpreted as *M*
Minor premise	The interpretation of "*loss*" in *Employment Relations Act* as *PecuniaryLoss* fits *ordinary language*
Conclusion	"*loss*" in the *Employment Relations Act* should be interpreted as *PecuniaryLoss*

Note that we use inverted commas for linguistic occurrences ("*loss*") and a single word, with capitalized initials for meanings (*PecuniaryLoss*).

By substituting the conditions of the *OL* canon, with the requirement of other canons listed above it is possible to generate other interpretation schemes. For instance, we can obtain the following scheme for *Technical Language* (*TL*):

Major premise	*TL*: If the interpretation of *E* in *D* as *M fits technical language,* then *E* in *D* should be interpreted as *M*
Minor premise	The interpretation of "*loss*" in the *Employment Relations Act* as *PecuniaryOrEmotionalLoss* fits *technical language*
Conclusion	"*loss*" in the *Employment Relations Act* should be interpreted as *PecuniaryOrEmotionalLoss*

Obviously, our interpretive schemes only provide the top-level step in the reasoning that is needed to apply an interpretive canon. For supporting the application of a canon, we need to establish the minor premise of the corresponding scheme, namely to show that the interpretation we are proposing indeed satisfies the canon we are considering. This requires specific arguments, according to scheme being considered. For instance, for establishing that interpretation "*pecuniary loss*" of expression "*loss*" in document *Employment Relations Act* fits canon *ordinary language*, we will have to establish, by providing adequate evidence, that this interpretation matches the current linguistic usage. Thus, for instance, to support the application of the *ordinary language* canon, we would need an inference like the following:

Major premise	If E is commonly understood as M, then the interpretation of E in D as M fits *ordinary language*
Minor premise	The "*loss*" is commonly understood as *PecuniaryLoss*
Conclusion	The interpretation of "*loss*" in *Employment Relations Act* as *PecuniaryLoss* fits *ordinary language*

Here, the minor premise is a substitution instance of the antecedent of the major premise.

3.3 Negative Interpretive Schemes

According to negative canons, if an interpretation meets the canon's condition, then it is to be rejected.

Major premise	C: If the interpretation of E in D as M *satisfies condition of C's canon*, then E in D should not be interpreted as M
Minor premise	The interpretation of e in d as m *satisfies condition of negative canon*
Conclusion	e in d should not be interpreted as m

The most common negative canon is the *a contrario* (*AC*), which rejects an interpretation which is over- or under-inclusive with regard to the usual semantic meaning of that expression, according to the idea that *Ubi lex voluit, dixit; ubi noluit, tacuit* (what the law wishes, it states, what the law does not want, it keeps silent upon). The *a contrario* canon can also be viewed as a counterfactual appeal to the intention of the legislator: If the legislator had meant to express a meaning that is different from the usual meaning (the semantic meaning) of the expression at issue, he would have used a different expression. Here is for instance an example of application of the *a contrario* canon.

Major premise	*AC*: If the interpretation of *E* in *D* as *M* conflicts with the usual meaning of *E* (is over or under-inclusive), then *E* in *D* should not be interpreted as *M*
Minor premise	The interpretation of the expression "*loss*" in the *Employment Relations* as *PecuniaryOrEmotionalLoss* conflicts with the usual meaning of "*loss*"
Conclusion	"*loss*" in *Employment Relations Act* should not be interpreted as *PecuniaryOrEmotionalLoss*

There is also a more specific kind of *a contrario* argument, which we may call subclass *a contrario*: Rather than rejecting an interpretation as a whole, it addresses the exclusion or inclusion of a certain subclass in the interpretation at issue, based on the fact that the subclass is included in or excluded from the usual meaning. Here are the two variants: the exclusionary *a contrario* (*eAC*) and the inclusionary a contrario (*iAC*). Note that the *iAC* has a positive interpretive conclusion, as the non-exclusion, i.e., the non–non-inclusion is an inclusion.

Here is the first variant, namely the exclusionary *a contrario* argument.

Major premise	*eAC*: If the interpretation of *E* in *D* as including *S* conflicts with the usual meaning of *E*, then *E* in *D* should be interpreted as excluding *S*
Minor premise	The interpretation of "*loss*" in the *Employment Relations* as including *EmotionalLoss* conflicts with the usual meaning of "*loss*"
Conclusion	"*loss*" in *Employment Relations Act* should be interpreted as excluding *EmotionalLoss*

Here is the second variant, the inclusionary *a contrario* argument.

Major premise	*iAC*: If the interpretation of *E* in *D* as excluding *S* conflicts with the usual meaning of *E*, then *E* in *D* should be interpreted as including *S*
Minor premise	The interpretation of "*loss*" in the *Employment Relations* as excluding *EmotionalLoss* conflicts with the usual meaning of "*loss*"
Conclusion	"*loss*" in *Employment Relations Act* should be interpreted as including *EmotionalLoss*

The *a contrario* scheme can also be used in a meta-dialogical sense that concerns the choice of the scheme. A clear example is the following argument taken from *R. v. Barnet London Borough Council* (1 All ER 97, 2004):

> The words 'ordinarily residing with' are common English words and here there is no context requiring that they should be given other than their natural meaning in accordance with the accepted usage of English. Even in such circumstances, however, there can be difficulty and doubt as to their applicability to particular facts, because the conception to which the words have reference does not have a clearly definable content or fixed boundaries.

The reasoning can be represented as follows, where *mAC* stands for meta-*a contrario*.

Major premise	*mAC*: If *E* in *D* is an ordinary English expression, and *E* in *D* has no context requiring a technical meaning, then the *technical language* is inapplicable to expression *E* in a document *D*
Minor premise 1	"*Ordinarily residing with*" in the *Local Education Authority Awards Regulations* is an ordinary English expression
Minor premise 2	"*Ordinarily residing with*" in the *Local Education Authority Awards Regulations* has no context requiring a technical meaning
Conclusion	The *technical language* canon is inapplicable to expression "*Ordinarily residing with*" in the *Local Education Authority Awards*

In this case, the absence of a context requiring a technical language (such as a definition, or the technical nature of the object of the regulation at issue) leads to the inapplicability of the *technical language* canon. This scheme is not a mere rebuttal (exclusion of a determinate meaning), but an undercutter (an attack to the grounds of an argument, in this case the possibility of using a major premise) (Pollock 1995; Walton 2015). Thus, the fact that the *technical language* argument cannot be used to support that interpretation does not exclude that the same interpretation can be successfully proposed through a different argument, such as the teleological one (argument from purpose).

The meta-dialogical analysis of the *a contrario* argument raises two issues concerning its nature. The first one is the relationship between the exclusion of alternative canons of interpretations and the idea of default. According to Alexy and Dreier (1991, 95–98), the *ordinary language* scheme should be taken as the default setting. The general principle at work here is the following conditional: Any expression in a legislative document should be interpreted using *ordinary language*, unless there are superior reasons to interpret the expression as fitting one of the other ten schemes. However, all interpretive canons are defaults. The difference here is that for any expression we can raise the defeasible claim that it should be interpreted according to its ordinary language meaning, while claims based on other canons can only be raised under specific conditions (e.g., a technical context is required to substantiate the claim that a term should be interpreted in a technical meaning).

The second controversial issue about the *a contrario* argument is whether it ought to be treated only as an argumentation scheme or also as a meta-level principle that can be applied in conjunction with interpretive argumentation schemes. Argument from ignorance has traditionally been treated as an argumentation scheme in logic (Macagno and Walton 2011; Walton 1995), whereas the closed world assumption has been treated in AI as a meta-level principle rather than as a specific form of argument in its own right (Reiter 1980). The *a contrario* argument is similar to the argument from lack of evidence as it supports an inference from a negative finding to a positive conclusion.

4 Attacking, Questioning, and Defending Interpretive Arguments

Since the basic defeasible schemes share a general pattern for interpretive arguments, there is no need to formulate critical questions for each of these schemes individually. The critical questions for each of them follow the general pattern indicated by the three critical questions presented below.

(CQ_1) What alternative interpretations of E in D should be considered?

(CQ_2) What reasons are there for rejecting alternative interpretations?

(CQ_3) What reasons are there for accepting alternative interpretations as better than (or equally good as) the one selected?

The function of the critical questions is to help someone dealing with interpretive issues to probe into an interpretive argument in order to get an initial idea of what some of the weak points of it might be. Critical questions have a heuristic function of suggesting to an arguer who is at a loss, on how to respond by suggesting possible avenues of attack. In this instance, the CQs are not independent of each other, and they have an ordering. CQ_1 should be asked first.

The way we will analyze interpretive arguments, as well as critical questions matching them and counterarguments attacking them, is to build an argumentation tree which includes a contested interpretive argument and provides an analysis of how the chains of argumentation on both sides of the dispute connect with each other and to the ultimate claim at issue. This can be done using tools from formal argumentation systems such as the Carneades Argumentation System (Carneades) or the ASPIC+ system. Both ASPIC+ and Carneades are based on a logical language comprising both strict and defeasible inference rules that can be used to build arguments, and both systems use argumentation schemes. Sartor et al. (2014) have applied ASPIC+ to build a logical analysis of interpretative schemes, and we will use here a simplified version of Carneades which will prove to have some tools that can be applied to examples illustrating the distinctive argumentation approach to interpretative arguments.

Both ASPIC+ and Carneades use a scheme called defeasible *modus ponens*, also used in the DefLog argumentation system of Verheij (2008). This scheme is a variant of *modus ponens* in which the antecedent of the conditional premise takes the form of a conjunction. Verheij (2008, 24) observed that if you look at a typical argumentation scheme with eyes slightly narrowed, it appears to have a *modus ponens* format in outline. In the formalism that will be used in the second part of the present contribution, a scheme fits the following type of argument structure, where the major premise is a defeasible conditional with a conjunctive antecedent.

Major Premise: $A, B, C, \ldots \Rightarrow Z$

Minor Premise: $A, B, C, \ldots$

Conclusion: Z

It was shown in Walton (2004, 134–139) how a majority of the schemes recognized in the argumentation literature can be tailored to fit this defeasible *modus ponens* form.

In all three systems, arguments are modeled as graphs containing nodes representing propositions from the logical language and edges from nodes to nodes. In these systems, an argument can be supported or attacked by other arguments, which can themselves be supported or attacked by additional arguments. The outcome in a typical case of argumentation is a graph structure representing a series of supporting arguments, attacks, and counterattacks in a sequence that can be represented using an argument map, also often called an argument diagram.

Carneades models arguments as directed graphs consisting of argument nodes connected to statement nodes. The premises and conclusions of an argument graph are represented as statement nodes, shown as rectangles in Fig. 3 (Gordon 2010). Argument nodes represent different structures of different kinds of arguments, such as linked or convergent arguments. A linked argument is one where two or more premises function together to support a conclusion. In the argument maps below, the name of the argumentation scheme is inserted in the node (the circle) joining the premises to the conclusion. As will be shown in the figures, there can be two kinds of arguments shown in the node, a pro (supporting argument) or a con (attacking) arguments. A supporting argument is represented by a plus sign in its argument node, whereas a con argument is represented by a minus sign in the nodes containing argumentation schemes such as *modus ponens*, argument from expert opinion, and so forth (http://carneades.github.com). Conflicts between pro and con arguments can be resolved using proof standards such as preponderance of the evidence (Gordon and Walton 2009b). Argument graphs are evaluated relative to audiences, modeled as a set of assumptions and an assignment of weights to argument nodes. An audience is defined as a structure <*assumptions*, *weight*>, where *assumptions* is a consistent set of literals assumed to be acceptable by the audience and *weight* is a partial function mapping arguments to real numbers in the range 0.0–1.0. These numbers represent the relative weights assigned by the audience to the arguments (Gordon and Walton 2011).

In Carneades, there can be compound arguments consisting of several argument nodes joined together by edges in the graph so that an argument represents a chain of reasoning from the supporting premises down to the ultimate proposition to be proved, the so-called statement at issue. Arguments are evaluated on the basis of whether the audience accepts the premises or not, and on how strong the various arguments making up the graph are. A very simple example of how an argument evaluation works in the Carneades system is shown in Fig. 2. The rounded nodes represent argumentation schemes accepted by the audience. A pro argument is indicated by the plus sign in its node. A con argument is represented by a minus sign in its argument node. A green (light gray) node means the proposition in it is accepted by the audience. A red (dark gray) node means the proposition in it is rejected by the audience. If the node is white (no color), the proposition in it is neither accepted nor rejected. In the printed version, green appears as light gray and red appears as dark gray.

In both argument diagrams shown in Fig. 2, the ultimate conclusion, statement 1, is shown on the far left of the diagram. First, let us consider which premises the audience accepts or rejects, as shown in the argument diagram on the left. Argument 2 is a pro argument supporting statement 1, while argument 3 is a con argument

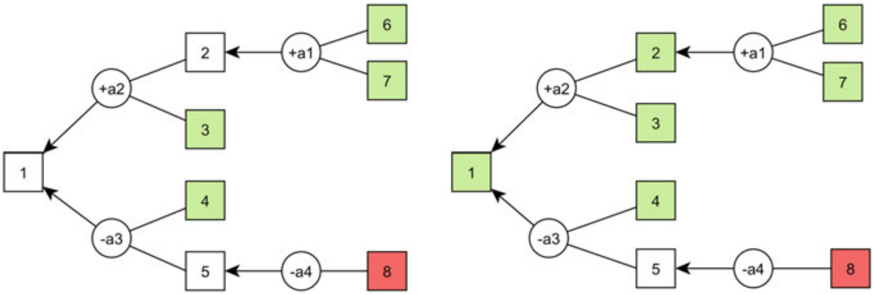


Fig. 2 Carneades graphs displaying an argument evaluation

attacking statement 1. The audience accepts proposition 3 as a premise in argument 2, but the other premise, statement 2, is neither accepted nor rejected by the audience. Both premises of this additional argument, argument 1, are accepted by the audience. Argument a3 is a con argument but one of its premises, statement 5, is not accepted. Moreover, this premise is attacked by a con argument, but the only premise in this con argument statement 6 is rejected.

To see how this conflict is resolved, look at the diagram on the right. Since both statements 6 and 7 are accepted by the audience, Carneades automatically calculates that the conclusion 2 is accepted. However, what about the con argument against statement 1 shown at the bottom, namely argument 3? This con argument could defeat statement 5, but its premise 8 is rejected by the audience. Therefore, pro argument a2 wins out over con argument a3, and so conclusion 1 is shown in green as acceptable.

Carneades also formalizes argumentation schemes. Schemes can be used to construct or reconstruct arguments, as well as to determine whether a given argument properly instantiates the types of argument deemed normatively appropriate according to the scheme requirements.

The critical questions matching an argumentation scheme cannot be modeled in a standard argument graph straightforwardly by representing each critical question as an additional implicit premise of the scheme. The reason is that there are two different variations on what happens when a respondent asks a critical question (Walton and Gordon 2005). These variations concern the pattern of how the burden of proof shifts from the proponent to the respondent and back as each critical question is asked by the respondent in a dialogue. With some critical questions merely asking the question is enough to defeat the proponent's argument, because the burden of proof is shifted onto the proponent's side, and if the proponent fails to meet this burden of proof, the initial argument is immediately defeated. With other critical questions, merely asking the critical question is not enough by itself to defeat the proponent's argument. For example, if the respondent asks the bias critical question when the proponent has put forward an argument from expert opinion, the proponent can simply reply, "What proof do you have that might expert is biased?" On this approach, merely asking the question does not defeat the proponent's argument until the respondent offers some evidence to back it up. Carneades deals with this problem of burden of proof for critical questioning by distinguishing three types of premises in an argumentation scheme, called ordinary premises, assumptions, and exceptions. Assumptions are assumed to be acceptable unless called into question. Exceptions are modeled as premises that are not assumed to be acceptable and which can block or undercut an argument as it proceeds. Hence, an exception, which is modeled in Carneades as an undercutter, only defeats the argument it was attacking if it is supported by other arguments which offer reasons to back up the undercutting argument. Ordinary premises of an argumentation scheme are treated as assumptions. They are assumed to be acceptable in case they are put forward, but must be supported by further arguments to remain acceptable after being challenged by critical questions or counterarguments.

For any one of these critical questions to be effective in defeating the original interpretive argument, the respondent must give some indication of what he takes this alternative interpretation to be. Thus, it would appear that each of these critical questions only defeats the original interpretive argument if some evidence is presented by the respondent pinpointing an alternative interpretation which might challenge the one originally appealed to by the proponent's argument.

Like ASPIC+, Carneades has three ways in which one argument can attack and defeat another. An opponent can attack one or more of the premises of an argument. This is called an undermining attack. Or an opponent can attack the conclusion by presenting an argument to show that the conclusion is false or unacceptable. This type of attack is called a rebutter. But thirdly, the opponent can attack the inferential link joining the premises to the conclusion. This type of attack is called an undercutter. For example, if the inference is based on a rule, the attack could claim that there is an exception to the rule that applies in the present case at issue. This way of modeling argumentation is based on Pollock's distinction (Pollock 1995, 40) between two kinds of argument attacks called rebutters and undercutters. On Pollock's view, a rebutter is a counterargument that attacks the conclusion of a prior argument, whereas an undercutter is a counterargument that attacks the argument link between the premises and the conclusion. For example, an argument that fits the argumentation scheme for argument from expert opinion can be critically questioned by asking whether the expert is biased. In Carneades, such a critical question is modeled as an undercutter, and an undercutter is modeled as an argument that defeats the original argument it was aimed at only if it is backed up by some additional evidence that supports it.

Next, we use Carneades to show how the interpretative statutory schemes can be applied to an extended sequence of argumentation in a typical case using a large argument graph to connect the individual interpretive arguments to each other.

5 The Education Grants Example

According to the account of the following case described in Cross (2005, 90), Section 1 of the Education Act of 1962 required local education authorities to make grants to students who were “ordinarily resident” in their area, so that the student could attend higher education courses. A requirement in the Education Act stipulated that to be eligible, the student had to have been ordinarily resident in the UK for three years prior to his or her application. The following issue arose: Could someone who had come to the UK for education count the period spent in education as ordinary residence to qualify for a mandatory grant under the Education Act?

There were two sides to the issue. The Court of Appeal held that such a person could not count this period as ordinary residence, offering the following argument (Cross 2005, 90). Lord Denning MR and Everleigh LJ were impressed by the need to relate this Act to the policy of the Commonwealth Immigrants Act 1962 and its successor, the Immigration Act 1971. Under the latter Act, students coming only for study had a conditional leave to stay in the country limited to the purpose of study which did not involve ordinary residence for the general purposes of everyday life. Denning and Everleigh considered that consistency with this Act required the term “ordinarily resident” in the Education Act to be interpreted as living as an ordinary member of the community would, which would not include residence for the limited purpose of study.

Arriving at a different interpretation, the House of Lords unanimously reversed this decision. They felt that the Court of Appeal had given too much weight to arguments drawn from the Immigration Act. They offered the following argument, quoted from Cross (2005, 91).

> Parliament’s purpose expressed in the Education Act gave no hint of any restriction on the eligibility for a mandatory award other than ordinary residence in the United Kingdom for three years and a satisfactory educational record. There was nothing expressed in the Immigration Act which gave guidance as to the interpretation of the Education Act and, indeed, despite a series of immigration measures since 1962, nationality had not formed part of the regulations under the Education Act until 1980. Accordingly, the ordinary natural meaning of the Education Act prevailed to make the students eligible for a mandatory grant if they had resided in the United Kingdom for the purposes of study.

In this case, it was concluded that the role of the judge should not be to reconcile legislative provisions. Instead, it was proposed that the basis for interpretation should be that of the ordinary language meaning of the expression “ordinarily resident.”

The argumentation in this case can be analyzed as an interpretive argument put forward by its proponents Denning and Everleigh and countered by an interpretive argument put forward in the House of Lords. Below, we use a sequence of three argument maps to model the structure of the argumentation sequence in the case.

The first argument, shown in Fig. 3, cites the Immigration Act of 1971, which stated that students coming to a country for study only had a conditional leave to stay in the country, adding that this conditional leave does not involve ordinary residence for the general purposes of everyday life. Because a related document is cited as the basis for drawing a conclusion in support of statutory interpretation, the argumentation scheme which is the basis of this argument is the one for argument from contextual harmonization (CH), recognized by MacCormick and Summers. For present purposes, this scheme is taken to represent the following kind of argument: A certain expression that occurs in a document is best interpreted as fitting with its usage in a set of related documents; therefore, in this document it will interpreted in

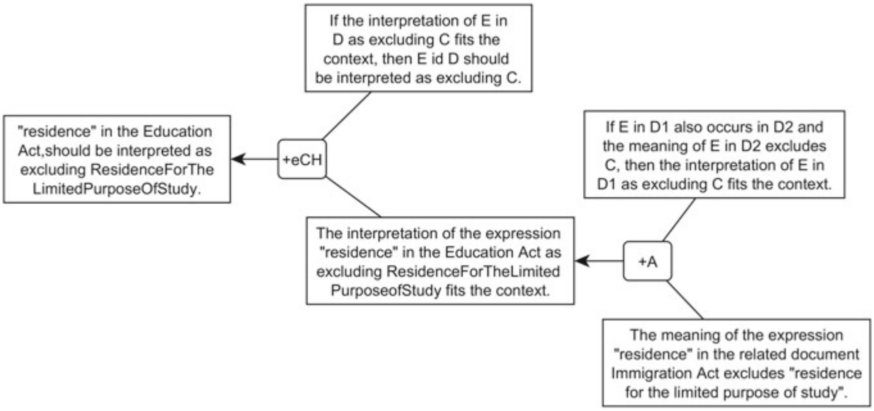

If the interpretation of E in D as excluding C fits the context, then E id D should be interpreted as excluding C.
"residence" in the Education Act,should be interpreted as excluding ResidenceForThe LimitedPurposeOfStudy.
+eCH
If E in D1 also occurs in D2 and the meaning of E in D2 excludes C, then the interpretation of E in D1 as excluding C fits the context.
The interpretation of the expression "residence" in the Education Act as excluding ResidenceForTheLimited PurposeofStudy fits the context.
+A
The meaning of the expression "residence" in the related document Immigration Act excludes "residence for the limited purpose of study".

Fig. 3 Proponent's argument in the educational grants example

the same way. In other words, if there is an issue about how to interpret an expression in a document, such as a statute, then it can be argued that the best way to interpret it is within a context of related documents so that it fits with the way the term has been interpreted in these other documents.

Let us apply the scheme for the argument from contextual harmonization to the first part of this example. The notation +*CH*, referring to a supporting use of argument from contextual harmonization, has been inserted in the node linking the two premises in the middle of Fig. 2 to the ultimate conclusion shown at the left. Here is a textual representation of the arguments, which corresponds to the graph of Fig. 3. Let us first examine the top argument by Lord Denning.

Major premise	*eCH*: If the interpretation of *E* in *D* as excluding *C* fits the context, then *E* in *D* should be interpreted as excluding *C*
Minor premise	The interpretation of "*residence*" in the *Education Act* as excluding *ResidenceForTheLimitedPurposeOfStudy* fits the context
Conclusion	"*residence*" in *Education Act* should be interpreted as excluding *ResidenceForTheLimitedPurposeOfStudy*

The supporting argument may appeal to the fact that in other pieces of legislation "*ordinary residence*" excludes indeed "*residence for the limited purpose of study*."

The ultimate conclusion is the statement that non-UK students cannot count the period as ordinary residence.

Next, we turn to an analysis of the argumentation in the second quoted text above, where the opponent, in this instance the House of Lords, put forward a counterargument.

Major premise	*eCH*: If an expression *E* in document D_1 also occurs in a related document D_2, and the meaning of *E* in D_1 excludes a concept *C,* then the interpretation of the expression *E* in D_2 as excluding *C* fits the context
Minor premise	The meaning of "*residence*" in the related document *Immigration Act* excludes concept "*residence for the limited purpose of study*"
Conclusion	The interpretation of an expression "*residence*" in the *Education Act* as excluding *ResidenceForTheLimitedPurposeOfStudy* fits the context

> Parliament's purpose expressed in the Education Act gave no hint of any restriction on the eligibility for a mandatory award other than ordinary residence in the United Kingdom for three years and a satisfactory educational record.

This argument fits the scheme for inclusionary argument from intention (+*iAI*):

Major premise	+*iAI*: If the interpretation of *E* in *D* as excluding *S* conflicts with legislative purpose, then *E* in *D* should be interpreted as including *S*
Minor premise	The interpretation of an expression "*residence*" in the *Education Act* as excluding *ResidenceForTheLimitedPurposeOfStudy* conflicts with legislative purpose
Conclusion	"*residence*" in *Education Act* should be interpreted as including *ResidenceForTheLimitedPurposeOfStudy*

The reason why the minor premise holds is provided by the following supporting counterfactual argument.

Major premise	If the linguistic meaning of *E* in *D* includes *S*, and there are no hints that the legislator intended to exclude *S* from the meaning of *E* in *D*, then the interpretation of *E* in *D* as excluding *S* conflicts with legislative intention
Minor premise 1	The linguistic meaning of "*residence*" in the *Education Act* includes *ResidenceForTheLimitedPurposeOfStudy*
Minor premise 2	There are no hints the legislator intended to exclude *ResidenceForTheLimitedPurposeOfStudy* from the meaning of "*residence*" in *Education Act*
Conclusion	The interpretation of an expression "*residence*" in the *Education Act* as excluding *ResidenceForTheLimitedPurposeOfStudy* conflicts with legislative intention

This argument is shown in Fig. 4 as a counterargument to the one in Fig. 3.

We leave it as an open problem how the argument on the right could be more fully represented, for example, by including the "there are no hints" statement as a premise in an *a contrario* argument. This would make the argument on the right more complex. Hint: it is possible to solve this problem by invoking the notion of an enthymeme.

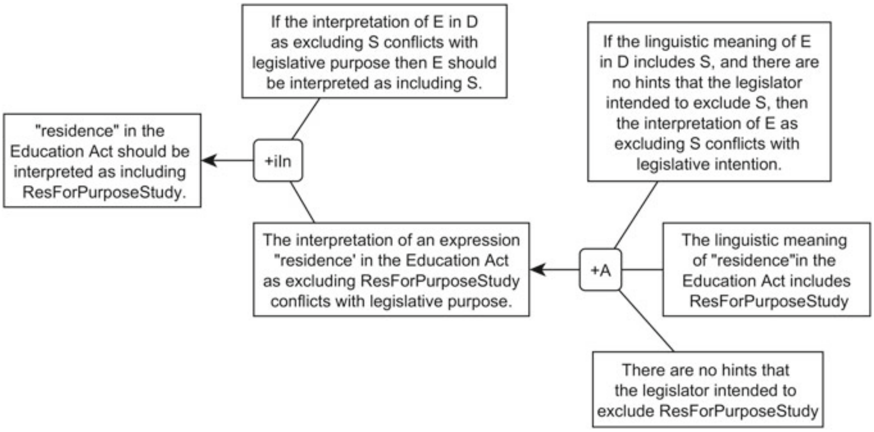

If the interpretation of E in D as excluding S conflicts with legislative purpose then E should be interpreted as including S.
"residence" in the Education Act should be interpreted as including ResForPurposeStudy.
+iIn
If the linguistic meaning of E in D includes S, and there are no hints that the legislator intended to exclude S, then the interpretation of E as excluding S conflicts with legislative intention.
The interpretation of an expression "residence" in the Education Act as excluding ResForPurposeStudy conflicts with legislative purpose.
+A
The linguistic meaning of "residence"in the Education Act includes ResForPurposeStudy
There are no hints that the legislator intended to exclude ResForPurposeStudy

Fig. 4 Respondent's rebuttal to the educational grants example

Next let us look at the other argument just below this one. Cross (2005, 91–92) offers this account of this part of the case.

> Lord Denning MR and Everleigh LJ were impressed by the need to relate this Act to the policy of the Commonwealth Immigrants Act 1962 and its successor, the Immigration Act 1971. Under the latter act, students coming only for study had a conditional leave to stay in the country limited to the purpose of study and this did not involve ordinary residence for the general purposes of everyday life. They considered that consistency with this Act requires the term 'ordinarily resident' in the Education Act to be interpreted as living as an ordinary member of the community would, which could not include residence for the limited purpose of study.

We are told in the quoted part of the text that Denning and Everleigh considered that consistency with the Education Act requires living as an ordinary member of the community and that being an ordinary member of the community does not include residence for the limited purpose of study. Accordingly, we have represented these two propositions as premises in a linked argument supporting the conclusion that conditional leave does not involve ordinary residence, as shown in Fig. 5 at the bottom right. The rightmost argument supports one premise of the argument to the left of it. It is labeled as a supporting argument labeled *+iPr* in Fig. 5. The conclusion of this argument is the opposite of the conclusion shown in Fig. 4.

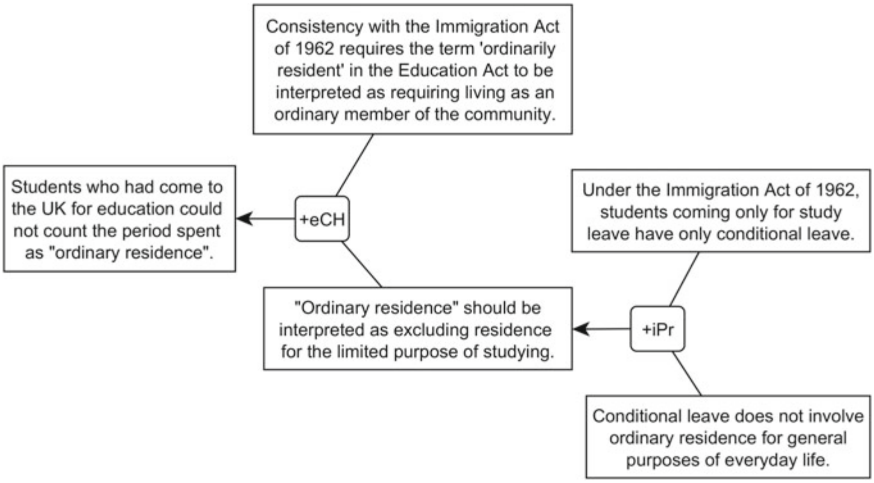

Consistency with the Immigration Act of 1962 requires the term 'ordinarily resident' in the Education Act to be interpreted as requiring living as an ordinary member of the community.
Students who had come to the UK for education could not count the period spent as "ordinary residence".
+eCH
Under the Immigration Act of 1962, students coming only for study leave have only conditional leave.
"Ordinary residence" should be interpreted as excluding residence for the limited purpose of studying.
+iPr
Conditional leave does not involve ordinary residence for general purposes of everyday life.

Fig. 5 Respondent's premise attack in the educational grants example

What we see in Fig. 5 is therefore a rebuttal because it presents an argument that attacks the ultimate conclusion of the original argument shown in Fig. 4. There is a conflict between the argument shown in Fig. 5 and the previous two arguments shown in Figs. 3 and 4.

We have chosen to use the term "interpretation" instead of "meaning," because the latter term is not only vague but is itself susceptible to many contested interpretations. Nevertheless, it can be said generally that what the interpreters of the statue are generally seeking is an interpretation that they contend that represents the genuine, true, or real meaning of the textual item they are discussing. This notion that there is what is called a real meaning underneath the vagaries in the text being examined or deconstructed has however been subject to some abuse in philosophy. For all these reasons, we generally prefer using the term "interpretation" to the term "meaning."

The evaluation system of Carneades compares the set of pro arguments against the set of con arguments if the arguments are independent of each other. However, summing the weights of arguments to check if the sum of the weights of the pro arguments outweighs the sum of the weights of the con arguments is only feasible if it be assumed that the arguments are independent of each other. This can be done with Carneades, but it requires an additional evaluation.

As with all arguments found in natural language texts, it is possible to analyze the given text in further depth by bringing out more implicit assumptions and more subtle inferences. However, building an argument map of a real argument expressed in natural language is very often a difficult interpretive task requiring learned skills and often itself providing many challenges of textual interpretation. Generally, one finds there are alternative interpretations opened up as the text of the cases is analyzed in greater depth and more implicit premises and arguments are brought out. Building an argument diagram can often raise important questions of argument interpretation and analysis that might not be initially visible to someone who is trying to deal with the argument or find out what to do with it. To illustrate some of the problems inherent in such as task, we go back to the *Dunnachie* example.

6 Fitting Interpretive Schemes to Cases

Dunnachie, following the commentary of MacCormick (2005, 128), offers an example of argument from contextual harmonization. The scheme for argument from contextual harmonization requires that a particular sentence in a statute should be interpreted considering the whole statute and any set of related statutes that are available. In line with the model of interpretive schemes introduced in Sect. 2, the scheme for contextual harmonization as applied to *Dunnachie* takes the following form.

Major premise	+*CH*: If the interpretation of *E* in *D* as *M* fits the context, then *E* in *D* should be interpreted as *M*
Minor premise	The interpretation of "*loss*" in the *Employment Relations Act* as *PecuniaryLoss* fits the context
Conclusion	"*loss*" in *Education Act* should be interpreted as *PecuniaryLoss*

The reason why this interpretation fits context is provided by the following supporting argument, which addresses the case in which the same expression occurs in different positions in the document (for simplicity's sake, we do not include in the scheme the possibility that there are multiple occurrences of the expression in the same document):

Major premise	If *E* besides occurring in position P_1 of document *D* also occurs in positions $P_1, \ldots, P_n$, where it has meaning *M*, then *E* in P_1 should also be interpreted as *M*
Minor premise	"*loss*" besides occurring in *Section* 2 of the *Employment Relations Act* also occurs in *Section* 4 where it has the meaning "*pecuniary loss*"
Conclusion	"*loss*" in *Section* 2 of the *Employment Relations Act* should be interpreted as "*pecuniary loss*"

Again following the commentary of MacCormick (2005, 128) on *Dunnachie*, the following example can be given to show how Carneades models a pro argument supporting a claim in a case where there is also a con argument attacking the same claim (Fig. 6).

The claim that "loss" should be interpreted as including both financial loss and emotional loss was partly based on a statement made in an earlier case. In this case, *Johnson Unisys Ltd*., Lord Hoffman had made the statement that an extension of the word "loss" to "emotional loss" could be made. So, it would appear, at least initially, that the argument drawn from the statement can be classified as an instance of a pro argument from precedent.

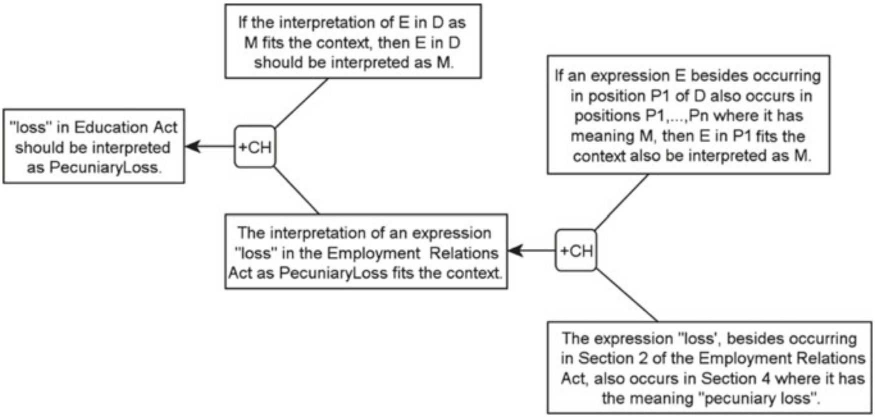

If the interpretation of E in D as M fits the context, then E in D should be interpreted as M.
"loss" in Education Act should be interpreted as PecuniaryLoss.
+CH
If an expression E besides occurring in position P1 of D also occurs in positions P1,...,Pn where it has meaning M, then E in P1 fits the context also be interpreted as M.
The interpretation of an expression "loss" in the Employment Relations Act as PecuniaryLoss fits the context.
+CH
The expression "loss', besides occurring in Section 2 of the Employment Relations Act, also occurs in Section 4 where it has the meaning "pecuniary loss".

Fig. 6 Use of the scheme for argument from contextual harmonization in *Dunnachie*

The reader will recall from the list in Sect. 2 that according to the description given by MacCormick and Summers, (1987) an interpretive argument from precedent requires that if a term has a previous judicial interpretation, it should be interpreted to fit that previous interpretation. In the previous case of *Norton Tool Co. v Tewson*, it had been ruled that "loss" was to be interpreted as signifying exclusively financial loss. Following the lines of the analysis of the structure of interpretative schemes in Sect. 3, the scheme for interpretive argument from precedent can be cast in inclusionary and exclusionary forms. Here is a exclusionary application of the argument by precedent:

Major premise	*ePr*: If the interpretation of *E* in *D* as excluding *S* fits precedents, then *E* in *D* should be interpreted as excluding *S*
Minor premise	The interpretation of an "*loss*" in the *Employment Relations Act* as excluding *EmotionalDamage* fits precedents
Conclusion	"*loss*" in *Education Act* should be interpreted as excluding *EmotionalDamage*

The supporting argument is the following:

Major premise	If *E* in *D* was understood in precedent *P* as excluding *C*, then the interpretation of *E* in *D* as excluding *C* fits precedents
Minor premise	"*loss*" in the *Employment Relations Act* was understood in Norton as excluding *EmotionalDamage*
Conclusion	The interpretation of "*loss*" in the *Employment Relations Act* as excluding *EmotionalDamage* fits precedents

In the next page, you can see is a positive application of the argument by precedent, followed by a supporting argument.

Major premise	*iPr*: If the interpretation of *E* in *D* as including *C* fits precedents, then *E* in *D* should be interpreted as *M*
Minor premise	The interpretation of "*loss*" in the *Employment Relations Act* as including *EmotionalDamage* fits precedents
Conclusion	"*loss*" in *Education Act* should be interpreted as including *EmotionalDamage*

Major premise	If *E* in *D* was understood in precedent *P* as including *C*, then the interpretation of *E* in *D* as including *C* fits precedents
Minor premise	The interpretation of an expression "*loss*" in the *Employment Relations Act* was understood in precedent *Johnson vs Unisys* as including *EmotionalDamage*
Conclusion	The interpretation of an expression "*loss*" in the *Employment Relations Act* as including *EmotionalDamage* fits precedents

These arguments could be further developed by pointing to the clues which support this understanding of the precedent, using the argument diagram in Fig. 7.

But in *Dunnachie*, in addition to this pro instance of interpretive argument from precedent, there was also a con argument against the same conclusion. The conflict between the two interpretations is shown in Fig. 8.

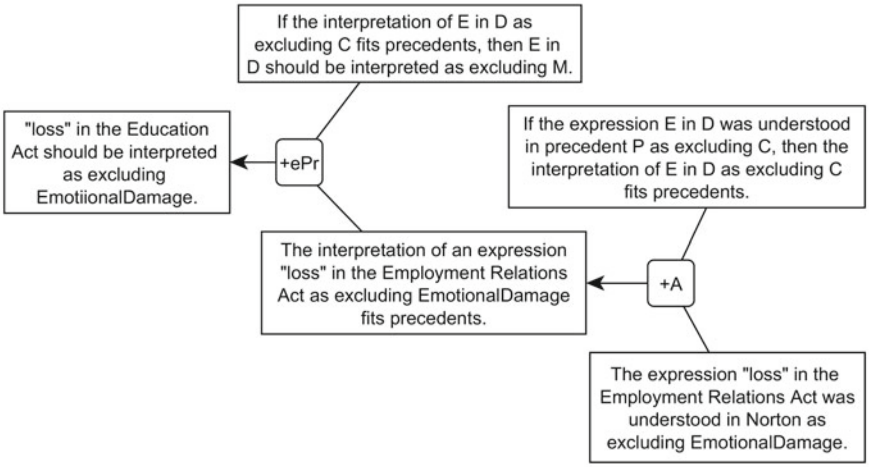

If the interpretation of E in D as excluding C fits precedents, then E in D should be interpreted as excluding M.
"loss" in the Education Act should be interpreted as excluding EmotiionalDamage.
+ePr
If the expression E in D was understood in precedent P as excluding C, then the interpretation of E in D as excluding C fits precedents.
The interpretation of an expression "loss" in the Employment Relations Act as excluding EmotionalDamage fits precedents.
+A
The expression "loss" in the Employment Relations Act was understood in Norton as excluding EmotionalDamage.

Fig. 7 Use of a prior case as a precedent supporting a textual interpretation

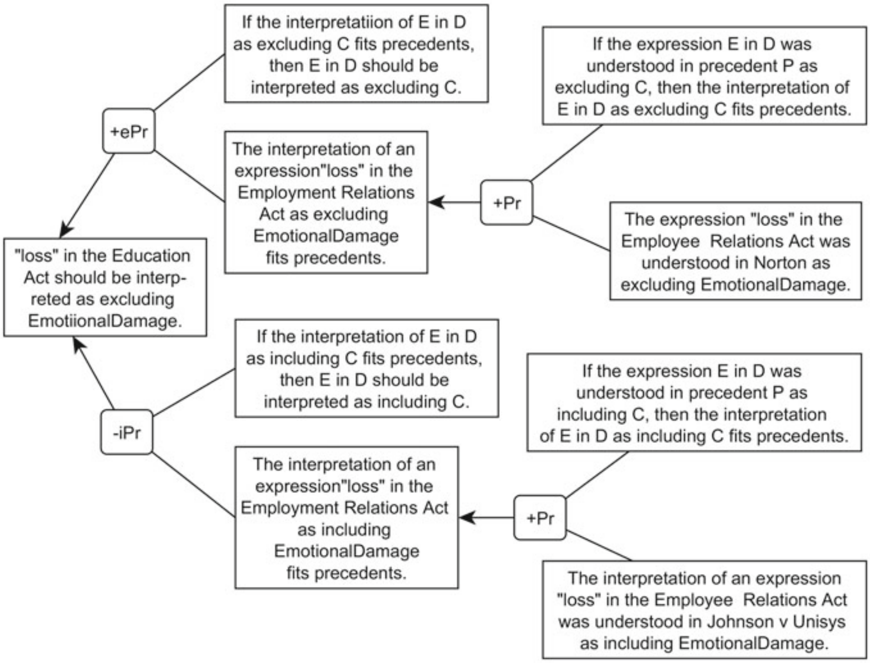

If the interpretatiion of E in D as excluding C fits precedents, then E in D should be interpreted as excluding C.
+ePr
If the expression E in D was understood in precedent P as excluding C, then the interpretation of E in D as excluding C fits precedents.
The interpretation of an expression"loss" in the Employment Relations Act as excluding EmotionalDamage fits precedents.
+Pr
The expression "loss" in the Employee Relations Act was understood in Norton as excluding EmotionalDamage.
"loss" in the Education Act should be interp-reted as excluding EmotiionalDamage.
If the interpretation of E in D as including C fits precedents, then E in D should be interpreted as including C.
If the expression E in D was understood in precedent P as including C, then the interpretation of E in D as including C fits precedents.
-iPr
The interpretation of an expression"loss" in the Employment Relations Act as including EmotionalDamage fits precedents.
+Pr
The interpretation of an expression "loss" in the Employee Relations Act was understood in Johnson v Unisys as including EmotionalDamage.

Fig. 8 Conflicting pro and con interpretive arguments from precedent

How could this conflict be resolved? The answer requires taking a closer look at the interpretive scheme for argument from precedent to see how one precedent can be stronger than another in supporting or attacking a claim about how a statute or law should be interpreted.

This way of modeling the scheme rests on the assumption that the user already has a clear idea of what a precedent is. Schauer (1987) has shown that arguments from precedent are already highly familiar in everyday conversational argumentation. This suggests that we need to begin with some intuitive understanding of what constitutes a precedent case. We could also build on the scheme for argument from precedent generally known in the argumentation literature, but there are differences of opinion on how that should be formulated (Walton 2010), in particular on the issue of how that scheme is related to the one for argument from analogy.

In his commentary on the case, MacCormick (2005, 129) made the following argument to support seeing this statement by another court as a binding premise in an argument from precedent. First, this ruling had been followed and approved many times. Second, it contained an acceptable rationale for interpreting loss exclusively as financial loss. Therefore, MacCormick concluded that it was a better guide for future rulings than the *Johnson* case.

In contrast, MacCormick put forward arguments advancing several reasons why Lord Hoffman's statement in *Johnson* might not constitute a binding precedent. First, it was not necessary to the decision reached in *Johnson*. Second, it had not been followed by other courts as a binding precedent. Third, although it was open to the House of Lords to have overruled *Norton Tool*, establishing a new ruling on the meaning of loss, this was not done. These arguments were used by MacCormick to question whether the remarks made by Lord Hoffman constitute a precedent binding on subsequent cases. These further arguments are shown in Fig. 9. For simplicity and readability's sake, we do not rigidly follow the structures illustrated above, and we omit to fully indicate the canons that are applied.

Let us say that all the propositions shown in the five rightmost rectangles are accepted by the audience. These five rectangles are shown in green backgrounds. Next, look at the pro argument from precedent at the top. Each of the two arguments supporting the proposition that *Norton Tool Co. v Tewson* is a precedent case has only one premise, and in both instances, that premise is accepted. Therefore, the proposition that *Norton tool Co. v Tewson* is a precedent case is automatically shown as accepted by Carneades. Let us also assume that the other premise of this argument is accepted. Since both premises of the argument are now accepted, the ultimate conclusion shown at the left of Fig. 9 is now automatically shown as accepted.

But now let us look at the bottom argument, the con argument from precedent. Since all three of its premises are accepted, the con argument attacking the proposition that *Johnson v Unisys* is a precedent case is successful in defeating it. Hence, this proposition is shown in a rectangle with a white background, indicating that it is not accepted. Actually, the additional evidence provided by the two pro arguments shown at the top right of Fig. 9 is not needed for the pro argument from precedent

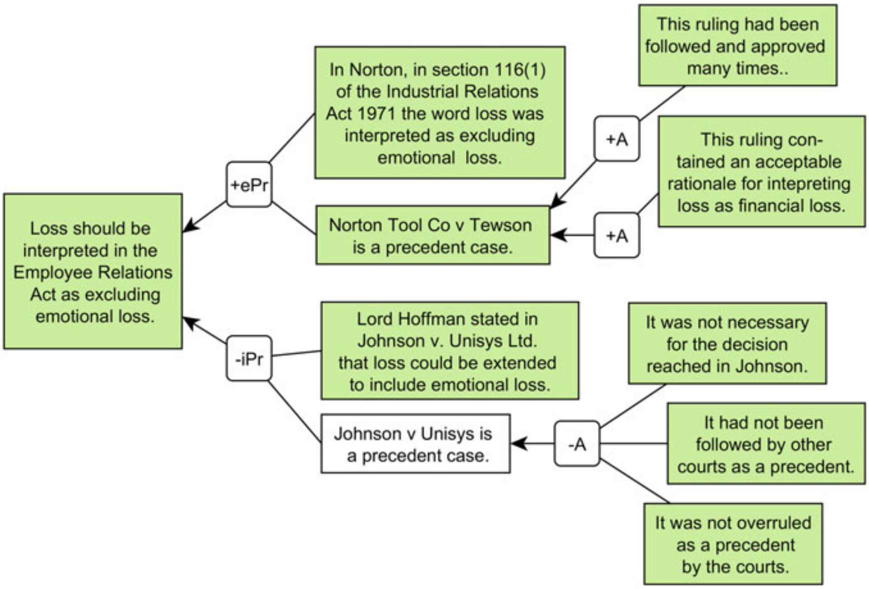

This ruling had been followed and approved many times..
In Norton, in section 116(1) of the Industrial Relations Act 1971 the word loss was interpreted as excluding emotional loss.
+A
This ruling contained an acceptable rationale for intepreting loss as financial loss.
+ePr
Loss should be interpreted in the Employee Relations Act as excluding emotional loss.
Norton Tool Co v Tewson is a precedent case.
+A
Lord Hoffman stated in Johnson v. Unisys Ltd. that loss could be extended to include emotional loss.
It was not necessary for the decision reached in Johnson.
-iPr
It had not been followed by other courts as a precedent.
Johnson v Unisys is a precedent case.
-A
It was not overruled as a precedent by the courts.

Fig. 9 Conflict resolved by taking other arguments into account

to defeat the con argument from precedent in the case. It is enough that because one premise of the con argument (shown in white at the bottom of Fig. 9) is defeated, the pro argument from precedent at the top prevails.

Summing everything up, the pro argument from precedent at the top prevails over the con argument from precedent at the bottom, because one of the premises of the con argument is unacceptable. It is shown by Carneades as not accepted because it is defeated by the applicable con argument - A. Only the pro argument is accepted, and so the conclusion is accepted. Hence, the conflict is resolved.

There is another way of modeling the conflict between the two arguments from precedent.

Using the scheme for argument from precedent put forward in Sect. 2, MacCormick's argument could be modeled as an undercutter critically questioning whether the top argument shown in Fig. 10 fits the argumentation scheme for argument from precedent. This way of interpreting MacCormick's remarks on how to model the argumentation in this instance is to take his argument above as an undercutter that attacks the argument used in the *Johnson* case by arguing that it is questionable whether the pro argument shown in Fig. 10 is a proper instantiation of the scheme for argument from precedent. Such an interpretation of MacCormick's evaluation of the argumentation is shown in Fig. 10.

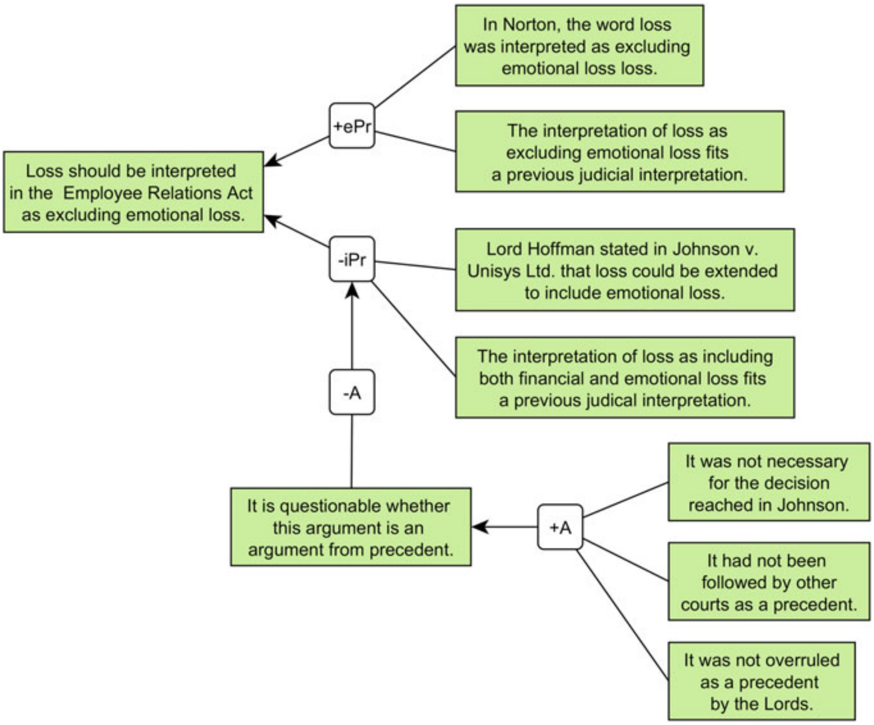

In Norton, the word loss was interpreted as excluding emotional loss loss.
+ePr
The interpretation of loss as excluding emotional loss fits a previous judicial interpretation.
Loss should be interpreted in the Employee Relations Act as excluding emotional loss.
-iPr
Lord Hoffman stated in Johnson v. Unisys Ltd. that loss could be extended to include emotional loss.
The interpretation of loss as including both financial and emotional loss fits a previous judical interpretation.
-A
It is questionable whether this argument is an argument from precedent.
+A
It was not necessary for the decision reached in Johnson.
It had not been followed by other courts as a precedent.
It was not overruled as a precedent by the Lords.

Fig. 10 Attacking an interpretive argument from precedent

The Johnson case is interesting because there is still another alternative interpretation of it that is possible, judging from MacCormick's remarks. It might be possible to argue that even though the ruling in Johnson on how to interpret loss was not binding because it was not necessary to the decision made in that case, still this ruling could be taken to have some significance. MacCormick (2005, 129) distinguishes between a binding precedent and a precedent that is persuasive but not binding. Honoring this distinction, the interpretation of the word "loss" in Johnson could be taken as a some significance. Following this line of argument, the conflict between the two arguments from precedent no longer represents a deadlock because the stronger precedent from Norton would have priority over the weaker precedent from Johnson. Carneades and ASPIC+, as well as other systems, recognize different kinds of priority orderings on rules, and so that would be another way that AI systems could model the argumentation in this case.

In Sect. 2, we only proposed schemes for some of the interpretive arguments to give the reader an idea of what these schemes should ultimately look like. However, especially with some of the schemes, the descriptions of the different kinds of interpretive arguments given by MacCormick and Summers are not enough in themselves to definitively formulate the matching schemes. In particular, the scheme for argument from precedent needs more study, a by applying it to cases, before a definitive version can be given.

7 Formalizing Interpretive Arguments—General Structure

In this section, we shall provide a general formal structure for interpretive arguments, based on the approach of interpretive arguments introduced and exemplified in the previous sections. Let us first summarize that approach.

Interpretive arguments can be distinguished along two different criteria: positive versus negative and total versus partial. The first distinction concerns whether they argue that a certain interpretation should be adopted or rather rejected. The second distinction pertains as to whether they address the whole interpretation of a term, or only the inclusion or exclusion of a subclass in the term's meaning. Correspondingly, partial interpretive arguments can be distinguished into exclusionary and inclusionary ones.

All interpretive arguments we shall consider are based on canons, namely defeasible conditionals, stating that if certain conditions are or are not met, a certain interpretive condition should or should not be adopted. Canons may be positive or negative dispending on whether their consequent is the adoption or the rejection a certain interpretation. Positive canons can also have a negative counterpart, to the extent that the absence of the condition they require leads to the rejection of an interpretation.

In this section, we shall propose appropriate formal structures for capturing all these forms of interpretive arguments.

Let us start with positive and negative total interpretive arguments. Both structures have the following elements: an expression E (word, phrase, sentence, etc.) occurs in a document D (statute, regulation, contract, etc.), interpreting this occurrence as meaning M satisfies the condition of a certain interpretive scheme (of ordinary language, technical language, purpose, etc.). Positive canons state that if all these elements are satisfied we are licensed to derive the interpretive conclusion that *E in D is interpreted as M.* Negative canons state that if an interpretation I would not fit the scheme, then *E in D is not interpreted as M.* In Sartor et al. (2014), we modeled interpretive claims as deontic claims, stating the obligation to adopt a certain interpretation. Here, we follow a different approach, focusing on the relationship between an interpretation and its justification, as a metalinguistic discourse on why a meaning is the best interpretation of an expression. In this sense, we model interpretive claims as terminological assertions concerning best interpretations of the contested or potentially contested expressions within a legal text (for a similar idea, see Araszkiewicz 2013).

All canons are modeled as defeasible rules (expressed in the form $r : \varphi_1, \ldots, \varphi_n \Rightarrow \psi$, where r is the rule name, where $\varphi_1, \ldots, \varphi_1$ and ψ are formulas in a logical language, $\varphi_1, \ldots, \varphi_1$ being the *antecedents,* and ψ being the *consequent* of the rule.

We express interpretive conclusions as claims concerning conceptual relations between a meaning M that is proposed and the outcome of the best legal interpretation of the linguistic occurrence at issue, namely expression E in document D (Bezuidenhout 1997; Carston 2002, 2013; Soames 2008; Sperber and Wilson 1986; Wilson and Sperber 2004). Such an outcome is denoted by the function expression $BestInt(E, D)$, denoting the best interpretation of expression E in document D. Conceptual relations are expressed with description logic symbols: $\equiv$ for conceptual equivalence, $\not\equiv$ for difference, $\sqsupseteq$ for inclusion. Thus *BestInt (E,D) = M* means that the best interpretation of expression E in document D is represented by meaning M.

Thus, a general pattern for positive total interpretive canons can be expressed as follows:

C: expression E occurs in document D,
the interpretation of E in D as M *satisfies the condition of positive canon C* $\Rightarrow$
$BestInt(E, D) \equiv M$

Here is an example:

OL: expression E occurs in document D,
the interpretation of E in D as M fits *ordinary language* $\Rightarrow$
$BestInt(E, D) \equiv M$

Similarly, negative canons claim that the best interpretation is not the proposed one, as in the following example, based on the non-redundancy canon:

NR: expression E occurs in document D,
the interpretation of E in D as M is redundant $\Rightarrow$
$BestInt(E, D) \not\equiv M$

Let us now provide examples for partial interpretations. For exclusionary interpretative claims. Consider the following canon:

> *eSAC*: expression E occurs in document D,
>
> the interpretation of expression E in the D as including S conflicts with usual meaning $\Rightarrow$
>
> $\mathit{BestInt}(E, D)^C \sqsupseteq S$
>
> where $\mathit{BestInt}(E, D)^C$ is the complement of $\mathit{BestInt}(E, D)$. In other terms, this canon concludes that concept (class) S is in the complement of the best interpretation of E in D, i.e., S is outside of the scope of this interpretation. For inclusionary interpretive claims, consider the following canon, which concludes for the inclusion of the concept S in the best interpretation.
>
> iSAC: expression E occurs in document D,
>
> the interpretation of E in the D as excluding S conflicts with the usual meaning $\Rightarrow$
>
> $\mathit{BestInt}(E, D) \sqsupseteq S$

We can also identify a pattern for priority arguments between different (instances of) interpretive canons (we use $\succ$ to express priority).

> C: concerning expression E in document D, the interpretation as M_1 according to canon C_1
>
> *meets the priority criterion with regard to* the interpretation as M_2 according to canon C_2 $\Rightarrow$
>
> $C_1(E, D, M) \succ C_2(E, D, M_2)$.

where *C(E,D,M)* denotes the instance of canon C which attributes meaning M to expression E in document D. Consider, for instance, Alexy and Dreier's idea that in criminal law *ordinary language* has priority over *technical language*.

> P_1: expression E in document D concerns Criminal law $\Rightarrow$
>
> $\mathit{OL}(E, D, M_1) \succ \mathit{TL}(E, D, M_2)$.

where *OL(E,D,*M_1*)* denotes the instance of canon *OL (ordinary languge)* which attributes meaning M_1 to expression E in document D, and similarly for *TL (techical languge)*. In this sense, interpretive arguments can be ordered in hierarchies depending on the specific legal context.

For reasoning about interpretation, we need an argumentation system including strict rules, defeasible rules, and preference between rules, such as the system developed by Prakken and Sartor (1996), the ASPIC+ system (Prakken 2010), or the Carneades system (Gordon and Walton 2009a). We express defeasible rules in the form $r : \varphi_1, \ldots, \varphi_n \Rightarrow \psi$ and strict rules in the form $\varphi_1, \ldots, \varphi_n \mapsto \psi$. We use arrows $\rightarrow$ and $\leftrightarrow$ for material conditional and biconditional of propositional logic. We also assume that our system includes the inferences of classical logic, namely that for any propositions of classical logic φ and ψ, if φ is derivable from ψ, then we have a strict rule $\varphi \mapsto \psi$.

Here, we assume that argument A including defeasible rules may be defeated in two ways. This first consists in successfully *rebutting* A, i.e., by contradicting the conclusion of a subargument of A, through an argument that is not weaker than the attacked subarguments (we assume that A too is a subargument of itself). More precisely, B rebuts A when (a) B's conclusion is incompatible with the conclusion of a subargument A' of A, and (b) B is not weaker than A', i.e., $A' \not\succ B$ (see Prakken

2010). Condition (b) corresponds to the idea that if A were stronger than B, it would resist B's challenge.

Regarding comparative strength, we assume that the comparison between two arguments A and B is to be assessed according to two criteria:

(a) preference for strict arguments (those only contains strict rules) over defeasible ones (those also containing defeasible rules): If A is strict and B is defeasible, then $A > B$.
(b) preference between defeasible arguments according to the last link principle: If A is preferable to B according to the last link principle, then $A > B$.

The *last link principle* assumes a partial strict ordering $\succ$ over defeasible rules and compares arguments A and B having incompatible conclusions by considering the sets of the last defeasible rules which support such conclusions in the two arguments (see for a formal characterization, Prakken and Sartor 1996; Prakken 2010).

The second way of defeating an argument A consists in *undercutting* A, i.e., in producing an argument B that rejects the application of a defeasible rule included in argument A. Let us express the claim that a rule does not apply, by denying the corresponding name of the rule: The statement $\neg r$ denies that rule named r applies. Then, we can say in general terms that argument B undercuts argument A, if B has the conclusion $\neg r$, where r is the top rule of a subargument A' of A. For instance, argument $[\rightarrow a; r_1\colon a \Rightarrow b]$ is undercut by argument $[\rightarrow c; r_2\colon c \Rightarrow \neg r_1]$. When we want to refer to the rule instance that is obtained by specifying a general rule r relatively to entities e, we use the expression $r(e)$. Thus, the expression $\neg r(e)$ expresses the claim that the rule instance $r(e)$ does not hold, or, in other words, the claim that the rule r does not apply to entities e. For instance, the proposition $\neg OL$(123(1)ERA) expresses the claim that canon OL does not apply to the text 123(1)ERA.

Semantics for an argumentation system can be based on the idea of an extension, namely a set of compatible arguments, which includes resources (arguments) that respond to all defeaters of arguments in the set. Here, we adopt the approach that consists in looking for most inclusive extensions, which are called preferred extensions (Dung 1995). An argument is then considered to be justified if it is included in all preferred extensions. It is considered defensible if it is included in some (but not necessarily in all) extensions.[1] The arguments that are defensible but not justified are only in some preferred extensions: Their status remains undecided, as their inclusion in a preferred extension depends on what other arguments are already included in the extension, different choices being possible.

Consider for instance the following set of arguments: $\{[a], [b], [a, r_1\colon a \Rightarrow c], [b, r_2\colon\ b\ \Rightarrow\ \neg c]\}$. We have two preferred extensions $E_1 = \{[a], [b], [a, r_1\colon a \Rightarrow c]\}$ and $E_2 = \{[a], [b], [b, r_2\colon b \Rightarrow \neg c]\}$. Each extension includes an argument that is defeated by, but also defeats, an argument in the other extension: $A_1 = [a, a \Rightarrow c]$ for E_1 and $A_2 = [b, b \Rightarrow \neg c]$ for E_2. So, each one of the two extensions is able to respond to all defeaters of any argument it includes. A_1

[1] In Sartor chapter 3, part II, this volume, on "Defeasibility in Law," a semantics based on labeling, which is equivalent to the extension based semantic here presented, was adopted.

and A_2 are merely defensible as they are incompatible, and we do not have, in the given set of arguments, reasons for preferring one to the other.

Assume that we add argument $[r_3 :\Rightarrow r_1 \succ r_2]$. Then, we have just one preferred extension, namely $\{[a], [b], [a, r_1\colon a \Rightarrow c], [r_3 :\Rightarrow r_1 \succ r_2]\}$, since, according to the preference $r_3 :\Rightarrow r_1 \succ r_2, A_1$ is no longer defeated by A_2.

Moving from arguments to conclusions, we have two possibilities for defining what conclusions are justified. One option is to view a conclusion as justified when it is established by a justified argument. The other option consists in viewing a conclusion as justified when it is supported in all preferred extensions, possibly through different arguments. More precisely, we get the following definition:

Definition (*Defensibility and Justifiability*).

- *Defensibility*. Claim φ is defensible with regard to argument set $\mathcal{A}$ if there exists a preferred extension S of $\mathcal{A}$ that contains an argument with conclusion φ.
- *Strong justifiability.* Claim φ is strongly justifiable with regard to argument set $\mathcal{A}$, if φ is the conclusion of an argument that is contained in all preferred extensions of $\mathcal{A}$.
- *Weak justifiability*. Claim φ is weakly justifiable with regard to argument set $\mathcal{A}$ if all preferred extensions of $\mathcal{A}$ contain arguments having conclusion φ.

Note that the weak definition of justifiability is broader than the strong, since it allows for a justifiable conclusion to be obtained through different incompatible arguments, included in different extensions. This is the notion that seems to be more appropriate to interpretation, as we shall argue in the following.

8 Interpretive Arguments—A Formalization

An interpretive argument can be constructed by combining an interpretive canon with the corresponding interpretive conditions. For instance, an argument from ordinary language can have the following form (in each argument, for conciseness sake, we put the general norm rather than its instantiation to the case at hand):

Argument A_1

1. expression "Loss" occurs in document 123(1)ERA
2. the interpretation of "Loss" in 123(1)ERA as *PecuniaryLoss* fits *ordinary language*
3. *OL*: expression E occurs in document $D \wedge$
   the interpretation of E in D as M fits *ordinary language* $\Rightarrow$
   BestInt$(E, D) \equiv M$

C.BestInt("Loss", 123(1)ERA) $\equiv$ *PecuniaryLoss*

Interpretive arguments can be attacked by counterarguments. For instance, the following counterargument based on *technical language* successfully rebuts the above argument based on *ordinary language*, by providing a different incompatible interpretation (assuming that no priority can be established and that concepts are different when denoted with a different name):

Argument A_2

1. expression “Loss” occurs in document 123(1)ERA
2. the interpretation of “Loss” in 123(1)ERA as *PecuniaryOrEmotioalLoss* fits technical language
3. *TL*: expression E occurs in document $D \wedge$

   the interpretation of E in D as M fits *technical language* $\Rightarrow$

   $BestInt(E, D) \equiv PecuniaryOrEmotionalLoss$

$BestInt$(“Loss”, 123(1)ERA) $\equiv$ *PecuniaryOrEmotionalLoss*

The interpretation based on ordinary language could also attacked by directly denying its conclusion, for instance by a non-redundancy argument claiming that “*Loss*” should not be interpreted in this way, since this would make 123(1)ERA redundant.

Argument A_3

1. expression “Loss” occurs in document 123(1)ERA
2. the interpretation of “Loss” in 123(1)ERA as *PecuniaryLoss* makes the norm redundant
3. *NR*: expression E occurs in document $D \wedge$

   the interpretation of E in D as M makes the norm redundant $\Rightarrow$

   $BestInt(E, D) \not\equiv M$

$BestInt$(“Loss”, 123(1)ERA $\not\equiv$ *PecuiaryLoss*

A rebutting attack can also be played by using partial (inclusionary or exclusionary interpretive) arguments.

Argument A_4

1. expression “Loss” occurs in document 123(1)ERA
2. the interpretation of “Loss” in 123(1)ERA as *including* $\sqsupseteq$ *EmotionalLoss* conflicts with usual meaning
3. *eAC*: expression E occurs in document D,

   the interpretation of expression E in the D as including S conflicts with usual meaning $\Rightarrow BestInt(E, D)^C \sqsupseteq S$

$BestInt$(“Loss”, 123(1)ERA)$^C \sqsupseteq$ *EmotionalLoss*

where $BestInt$(“Loss”, 123(1)ERA)^C denotes the complement of

BestInt(“Loss”, 123(1)ERA). In other words, the conclusion of this argument states that *EmotionalLoss* is included in the complement of the best interpretation of “Loss”, i.e., that it is completely excluded from this interpretation.

Given that *PecuniaryOrEmotionalLoss* on the contrary includes emotional loss, i.e.,

4. $\mathit{PecuniaryOrEmotionalLoss} \sqsupseteq \mathit{EmotionalLoss}$

we can conclude that the best interpretation of “Loss” is different from *PecuniaryOrEmotionalLoss*

5. *BestInt*(“Loss”, 123(1)ERA) $\not\equiv$ P*ecuniaryOrEmotionalLoss*

which contradicts the conclusion of the above argument A_2.

An undercutting attack against the *ordinary language* argument could be mounted by arguing that the expression “loss” in the Employment Rights Act is used in a technical context, e.g., in the context of the discipline of industrial relations, where arguments from *ordinary language* do not apply. Thus, this canon is inapplicable to the expression *Loss* in 123(1)ERA, which is expressed using the formalism above as $\neg OL$(123(1)ERA).

1. expression “Loss” occurs in document 123(1)ERA
2. 123(1)ERA is a technical context
3. *TC*: expression *E* occurs in document *D*,
   D is a technical context $\Rightarrow \neg OL(E)$

$\neg OL$(123(1)ERA)

9 Preference Arguments over Interpretive Arguments

We may have preferences over interpretive arguments. For example, in Italy, the Court of Cassation revised its interpretation of the term *Loss* (*danno*) as occurring in the Italian Civil Code (ICC) using an argument from substantive reasons (the constitutional value of health): The Court thus rejected the traditional interpretation as pecuniary damage, arguing that also damage to health should also be included in the scope of the term (and consequently compensated):

Argument A_1

1. expression “Loss” occurs in document Art2043ICC
2. the interpretation of “Loss” in Art2043ICC as *PecuniaryLoss* fits legal history
3. *OL*: expression *E* occurs in document *D*,
   the interpretation of *E* in *D* as *M* fits legal history $\Rightarrow \mathit{BestInt}(E, D) \equiv M$

BestInt(“Loss”, Art2043ICC $\equiv$ *PecuniaryLoss*)

Argument A_2

1. expression “Loss” occurs in document Art2043ICC
2. the interpretation of “Loss” in Art2043ICC as *PecuniaryLossOrDamageToHealth* contributes to substantive reasons
3. *SR*: expression E occurs in document D,

   the interpretation of E in D as M contributes to substantive reasons $\Rightarrow$

   $\mathit{BestInt}(E, D) \equiv M$

$\mathit{BestInt}$(“Loss”, Art2043ICC) $\equiv$ *PecuniaryLossOrDamageToHealth*

These two arguments conflict (rebut each other), as:

$$\mathit{PecuniaryLoss} \not\equiv \mathit{PecuniaryLossOrDamageToHealth}$$

To address the conflict, the judges argued that the second argument defeats the first, since SR in this context contributes to constitutional values.

Argument 3

1. The interpretation of expression “Loss” in Art2043ICC, as *PecuniaryLossOrDamageToHealth* according to *SR* contributes to constitutional values
2. *SR*: The interpretation of expression E in D, as M according to *SR* contributes to constitutional values$\Rightarrow$ $\mathit{SR}(E, D, M) \succ \mathit{LH}(E, D, M')$

SR(“Loss”, Art2043ICC, *PecuniaryLossOrDamageToHealth*) $\succ$

LH(“Loss”, Art2043ICC, *PecuniaryLossOrDamageToHealth*)

10 From Best Interpretations to Individual Claims

We must be able to move from interpretive claims to conclusion in individual cases, namely from conceptual assertions to individual claims. For this purpose, we can adopt general patterns for strict rules, which provide for the transition from interpretive claims to assertions concerning individuals:

1. $\mathit{BestInt}(E, D) \equiv M \mapsto \forall \boldsymbol{x}\, [E_D\, (\boldsymbol{x}) \leftrightarrow M\, (\boldsymbol{x})]$
2. $\mathit{BestInt}(E, D) \sqsupseteq M \mapsto \forall \boldsymbol{x}\, [M\, (\boldsymbol{x}) \rightarrow E_D\, (\boldsymbol{x})]$
3. $\mathit{BestInt}(E, D)^C \sqsupseteq M \mapsto \forall \boldsymbol{x}\, [M\, (\boldsymbol{x}) \rightarrow \neg E_D\, (\boldsymbol{x})]$

where $\boldsymbol{x}$ is sequence of variables which is required by concept M, $M(\boldsymbol{x})$ is the predicate corresponding to concept M, and E_D is a predicate representing the occurrence of E in D at issue. Consider for instance the above interpretive claim according to which

$$\mathit{BestInt}(\text{“loss”, 125ERA}) \equiv \mathit{PecuniaryLoss}$$

The corresponding instance of transition rule 1 would be:

$$\begin{aligned}\mathit{BestInt}(\text{“loss”, 125ERA}) &\equiv \mathit{PecuniaryLoss}\\ &\mapsto \forall \boldsymbol{x} \left[\mathit{Loss}_{\mathit{ERA}}(x, y, z) \leftrightarrow \mathit{PecuniaryLoss}(x, y, z)\right]\end{aligned}$$

To be read as: If the best interpretation of expression “loss” in document Section 125 of the Employment Relations Act is concept *PecuniaryLoss*, then a person x in an event y has a “loss” of amount z (as understood in Section 125 of the Employment Relations Act) if and only if x in y has a pecuniary loss of z.

Let us assume that John in his unfair dismissal by Tom had a pecuniary loss of Euro 100, i.e., $\mathit{PecuniaryLoss}(\mathit{John}, \mathit{DismissalByTom}, 100)$. Let us expand the ordinary language argument with the following: the latter assumption, the above instance of transition rule 1, and strict rules corresponding to an inference of classical logic. We get the following argument (where we list with the premises in the argument and with letters the intermediate conclusions).

Argument A_4

1. expression “Loss” occurs in document 123(1)ERA
2. the interpretation of “Loss” in 123(1)ERA as *PecuniaryLoss* fits *ordinary language*
3. *OL*: expression E occurs in document $D \wedge$ the interpretation of E in D as M *fits ordinary language* $\Rightarrow$ $\mathit{BestInt}(E, D) \equiv M$

---

a. $\mathit{BestInt}(\text{“Loss”, 123(1)ERA} \equiv \mathit{PecuniaryLoss}$ (from 1, 2, and 3)

4. $\mathit{BestInt}(\text{“loss”, 125ERA}) \equiv \mathit{PecuniaryLoss} \mapsto \forall \boldsymbol{x} \left[\mathit{Loss}_{\mathit{ERA}}(x, y, z) \leftrightarrow \mathit{PecuniaryLoss}(x, y, z)\right.$

---

b. $\forall \boldsymbol{x} \left[\mathit{Loss}_{\mathit{ERA}}(x, y, z) \leftrightarrow \mathit{PecuniaryLoss}(x, y, z)\right]$ (from a and 4)

5. $\mathit{PecuniaryLoss}(\mathit{John}, \mathit{DismissalByTom}, 100)$

---

c. $\mathit{Loss}_{\mathit{ERA}}(\mathit{John}, \mathit{DismissalByTom}, 100)$ (by classical logic) (from b and 5)

The mixture of interpretive and other arguments that are needed for a legal conclusion can also include additional conceptual relations. For instance, let us assume that we know that John has sustained a pecuniary loss of 100 Euros, as a consequence of his unfair dismissal. Since the concept of pecuniary loss is included in the concept of pecuniary or emotional loss, we can infer that he suffered a pecuniary or emotional loss. This conclusion would enable us to conclude that John has a loss in the sense of Section 125 ($Loss_{ERA}(John,\ DismissalByTom,\ 100)$), also on the basis of the interpretation of loss as *PecuniaryOrEmotionalLoss*, according to an argument *Argument* A_5 which includes this interpretation.

Argument A_5

1. expression "Loss" occurs in document 123(1)ERA
2. the interpretation of "Loss" in 123(1)ERA as *PecuniaryOrEmotionalLoss* fits *technical language*
3. *TL*: expression E occurs in document $D \wedge$

   the interpretation of E in D as M *fits technical language* $\Rightarrow$

   $BestInt(E, D) \equiv M$

   ---

   a. $BestInt$("Loss", 123(1)ERA $\equiv$ *PecuniaryOrEmotionalLoss* (from 1, 2, and 3)
4. $BestInt$("loss", 125ERA) $\equiv$ *PecuniaryOrEmotionalLoss* $\mapsto$

   $\forall \boldsymbol{x}[Loss_{ERA}(x, y, z) \leftrightarrow .PecuniaryOrEmotionalLoss(x, y, z)$

   ---

   b. $\forall \boldsymbol{x}\left[Loss_{ERA}(x, y, z) \leftrightarrow PecuniaryOrEmotionalLoss(x, y, z)\right]$ (from a, and 4)
5. $\forall \boldsymbol{x}\left[PecuniaryLoss(x, y, z) \rightarrow PecuniaryOrEmotionalLoss(x, y, z)\right]$
6. $PecuniaryLoss(John, DismissalByTom, 100)$

   ---

   c. $PecuniaryOrEmotionalLoss(John,\ DismissalByTom,\ 100)$ (from 5, and 6)

   ---

   d. $Loss_{ERA}(John,\ DismissalByTom,\ 100)$ (from b and c)

Arguments A_4 and A_5 are inconsistent, as they include incompatible interpretive conclusions (incompatible subarguments): According to conclusion (a) in A_4, the best interpretion of "loss" in Section 125 is *PecuniaryLoss*, while according to conclusion (a) in A_5 the best interpretation is a different concept, namely *PecuniaryOrEmotionalLoss*. However, the two arguments lead to the same conclusion in the case of John's dismissal: He suffers a loss of 100, as understood in Section 125 of the Employment Relations act.

Therefore, we may view this conclusion as legally justified, namely as weakly justified. This is the case even though we are unable to make a choice between the two incompatible interpretations (the two competing interpretive arguments are both defeasible, and neither is justified), as the conclusion follows from both such interpretations. This view corresponds to the idea that only relevant issues have to be addressed in legal decision-making: The issue of whether "loss" is limited or not to pecuniary losses is irrelevant in John's case, since he has only suffered a pecuniary loss (this issue would be relevant if he had on the contrary suffered instead, or additionally, an emotional loss).

11 Conclusions

In this chapter, our goal was to show how interpretive schemes can be formulated in such a manner that they can be incorporated into a formal and computational argumentation system such as Carneades or APSIC+ and then applied to displaying the pro–contra structure of the argumentation using argument maps applied to legal cases. To this purpose, we have analyzed the most common types of statutory arguments and brought to light their common characteristics. We have shown how canons of interpretation can be translated into argumentation schemes, and we have distinguished two general macrostructures of positive and negative, total and partial canons, under which various types of schemes and rebuttals can be classified. This preliminary classification was then used for modeling the interpretive arguments formally and integrating them into computational systems and argument maps.

The interpretive schemes can be applied initially when constructing an argument diagram to get an overview of the sequence of argumentation in a case of contested statutory interpretations. The schemes can be applied in order to help the argument analyst convey an evidential summary showing how the subarguments fit together in a lengthy sequence of argumentation in a case, as indicated in the main example of the educational grants case. The next step is to zoom in on parts of the argumentation sequence that pose a problem where critical questions need to be asked or refinements need to be considered. Here, the critical questions can be applied in order to find further weak points in an argument by bringing out implicit premises that may have been overlooked and that could be questioned.

The function of the set of critical questions matching a scheme is to give the arguer who wants to attack the prior argument some idea of the kinds of critical questions that need to be asked in replying to it. Thus, the critical questions can offer guidance as to where look for weak points that could be challenged. However, there are theoretical issues of how to structure the critical questions. If critical questions can be modeled in the argument diagrams as additional premises, ordinary premises, assumptions, or exceptions such as done in Carneades or ASPIC+, they can be modeled in argument maps as undercutting or rebutting counterarguments. The problem that always arises in attempts to fit critical questions into argument diagrams in this manner is one of burden of proof. Is merely asking a critical question enough to defeat a given argument? Or should a critical question be taken to defeat the given argument only if some evidence is given to back it up. Carneades or ASPIC+ provides a way of dealing with this problem that has been shown to be applicable to interpretative schemes.

The danger with using such schemes to construct hypotheses about the best interpretation is one of jumping to a conclusion too quickly. This danger can be overcome by asking critical questions matching the scheme and by considering possible objections to the argument fitting an interpretive scheme. For as we have seen in the example, a sequence of argumentation based on the application of interpretive argumentation schemes is defeasible and can be attacked by undercutters and rebutters in an opposed sequence of argumentation. Indeed, it is this very situation of one sequence of interpretive argumentation being used to attack another one that is characteristic of the example we studied, a standard example of statutory interpretation.

We have also provided a fresh logical formalization of reasoning with interpretive canons. Rather than modeling interpretive conclusion as deontic claims, as we did in Sartor et al. (2014), here we have modeled them as conceptual (terminological) claims concerning best interpretations.

We have then considered how interpretive arguments can be framed within argumentation systems, including defeasible and strict rules. We have argued that a semantics based on preferred extensions can provide an appropriate approach to interpretive conclusions and can be used to distinguish between defensible and justifiable interpretive claims. Regarding justification, we have argued for weak justifiability (derivation in all extensions, also through different argument) to be more appropriate to interpretive reasoning in legal contexts.

This work still is quite preliminary, but necessarily so, since AI and law research has neglected issues pertaining to statutory interpretation and more generally the issue of determining the correct meaning of authoritative sources of the law. Further research should include a more refined classification system for interpretative schemes. Also, the idea of merging argumentation with deontic logic as advanced in Sartor et al. (2014), Walton et al. (2014) needs to be reconsidered and integrated with the different framework presented in this chapter.

Acknowledgement Fabrizio Macagno would like to thank the Fundação para a Ciência e a Tecnologia for the research grants no. IF/00945/2013, PTDC/IVC-HFC/1817/2014, and PTDC/MHC-FIL/0521/2014. Douglas Walton would like to thank the Social Sciences and Humanities Research Council of Canada for Insight Grant 435-2012-0104.

References

Alexy, R., and R. Dreier. 1991. Statutory interpretation in the federal republic of Germany. In *Interpreting statutes. A comparative study*, ed. N. MacCormick, and R. Summers. Aldershot: Dartmouth.

Araszkiewicz, M. 2013. Towards systematic research on statutory interpretation in ai and law. In *Proceedings of JURIX 2014: The 27th annual conference on legal knowledge and information systems*, ed. R. Hoekstra, 15–24. Amsterdam: IOS Press.

Atlas, J.D. 2005. *Logic, meaning, and conversation*. Oxford: Oxford University Press.

Atlas, J.D. 2008. Presupposition. In *The handbook of pragmatics*, ed. L. Horn, and G. Ward, 29–52. Oxford: Blackwell.

Atlas, J.D., and S. Levinson. 1981. It-clefts, Informativeness and logical form: Radical pragmatics (Revised Standard Version). In *Radical Pragmatics*, ed. P. Cole, 1–62. New York: Academic Press.

Bezuidenhout, A. 1997. Pragmatics, semantic undetermination and the referential/attributive distinction. *Mind* 423: 375–409.

Butler, B. 2016. Law and the primacy of pragmatics. In *Pragmatics and law: Philosophical perspectives*, ed. A. Capone, and F. Poggi, 1–13. Cham: Springer.

Carston, R. 2002. *Thoughts and utterances: The pragmatics of explicit communication*. Oxford: Blackwell.

Carston, R. 2013. Legal texts and canons of construction: A view from current pragmatic theory. In *Law and language: Current legal issues*, vol. 15, ed. M. Freeman, and F. Smith, 8–33. Oxford: Oxford University Press.

Cross, R. 2005. *Statutory interpretation*, ed. J. Bell, and G. Engle. Oxford: Oxford University Press.

Damele, G. 2014. Analogia legis and analogia Iuris: An overview from a rhetorical perspective. In *Systematic approaches to argument by analogy*, ed. H. Ribeiro, 243–256. Amsterdam: Springer.

Dascal, M. 2003. *Interpretation and understanding*. Amsterdam: John Benjamins Publishing Company.

Dascal, M., and J. Wróblewski. 1988. Transparency and doubt: Understanding and interpretation in pragmatics and in law. *Law and Philosophy* 72: 203–224.

Gizbert-Studnicki, T. 1990. the burden of argumentation in legal disputes. *Ratio Juris* 31: 118–129.

Gordon, T. 2010. An overview of the carneades argumentation support system. In *Dialectics, dialogue and argumentation. An examination of douglas Walton's theories of reasoning and argument*, ed. C. Reed, and C. Tindale, 145–156. London: College Publications.

Gordon, T., and D. Walton. 2009a. Legal reasoning with argumentation schemes. In *Proceedings of the 12th international conference on artificial intelligence and law*, ed. C. D. Hafner, 137–146, New York: ACM.

Gordon, T., and D. Walton. 2009b. Proof Burdens and Standards. In *Argumentation in artificial intelligence*, ed. I. Rahwan, and G. Simari, 239–258. Berlin: Springer.

Gordon, T., and D. Walton. 2011. A formal model of legal proof standards and burdens. In *7th Conference on argumentation of the international society for the study of argumentation (ISSA 2010)*, ed. F. van Eemeren, B. Garssen, A. Blair, and G. Mitchell, 644–655. Amsterdam: Sic Sat.

Gray, C.B. 2013. *The philosophy of law: An encyclopedia, Vol. I–II*. London and New York: Routledge.

Hage, J. 1996. A theory of legal reasoning and a logic to match. *Artificial Intelligence and Law* 4: 199–273.

Hage, J. 1997. *Reasoning with rules*. Dordrecht: Kluwer.

Horn, L. 1995. Vehicles of meaning: Unconventional semantics and unbearable interpretation. *Washington University Law Quarterly* 73: 1145–1152.

Jaszczolt, K. 2005. *Default semantics*. Oxford: Oxford University Press.

Kecskes, I. 2008. Dueling contexts: A dynamic model of meaning. *Journal of Pragmatics* 3: 385–406.

Kecskes, I. 2013. *Intercultural pragmatics*. Oxford: Oxford University Press.

Kecskes, I., and F. Zhang. 2009. Activating, seeking, and creating common ground: A socio-cognitive approach. *Pragmatics and Cognition* 2: 331–355.

Macagno, F. 2015. A means-end classification of argumentation schemes. In *Reflections on theoretical issues in argumentation theory*, ed. F. van Eemeren, and B. Garssen, 183–201. Cham: Springer.

Macagno, F. 2017. Defaults and inferences in interpretation. *Journal of Pragmatics* 117: 280–290.

Macagno, F., and A. Capone. 2016. Interpretative disputes, explicatures, and argumentative reasoning. *Argumentation* 4: 399–422.

Macagno, F., G. Sartor, and D. Walton. 2012. Argumentation schemes for statutory interpretation. In *Argumentation 2012. International conference on alternative methods of argumentation in law*, eds. J. Šavelka, M. Araszkiewicz, M. Myška, T. Smejkalová, and M. Škop, 63–75. Brno: Masarykova univerzita.

Macagno, F., and D. Walton. 2011. Reasoning from paradigms and negative Evidence. *Pragmatics and Cognition* 1: 92–116.

Macagno, F., and D. Walton. 2014. *Emotive language in argumentation*. Cambridge: Cambridge University Press.

Macagno, F., and D. Walton. 2015. Classifying the patterns of natural arguments. *Philosophy and Rhetoric* 1: 26–53.

Macagno, F., and D. Walton. 2017. Arguments of statutory interpretation and argumentation schemes. *International Journal of Legal Discourse* 1: 47–83.

Macagno, F., D. Walton, and G. Sartor. 2018. Pragmatic maxims and presumptions in legal interpretation. *Law and Philosophy*. 37(1): 69–115. https://doi.org/10.1007/s10982-017-9306-4.

MacCormick, N. 2005. *Rhetoric and the rule of law*. Oxford: Oxford University Press.

MacCormick, N., and R. Summers (eds.). 1991. *Interpreting statutes: A comparative study*. Dartmouth: Aldershot.

Miller, G. 1990. Pragmatics and the maxims of interpretation. *Wisconsin Law Review*: 1179–1227.

Pollock, J. 1995. *Cognitive carpentry*. Cambridge, MA: MIT Press.

Prakken, H. 2010. An abstract framework for argumentation with structured arguments. *Argument and Computation* 2: 93–124.

Prakken, H., and G. Sartor. 1996. A dialectical model of assessing conficting arguments in legal reasoning. *Artificial Intelligence and Law* 4: 331–368.

Reiter, R. 1980. A Logic for default reasoning. *Artificial Intelligence* 1–2: 81–132.

Rotolo, A., G. Governatori, and G. Sartor. 2015. Deontic defeasible reasoning in legal interpretation: Two options for modelling interpretive arguments. In *Proceedings of the 15th international conference on artificial intelligence and law*, 99–108, New York, ACM.

Sartor, G., D. Walton, F. Macagno, and A. Rotolo. 2014. Argumentation schemes for statutory interpretation: A logical analysis. In *Frontiers in artificial intelligence and applications*, ed. R. Hoekstra, 11–20. Amsterdam: IOS Press.

Schauer, F. 1987. Precedent. *Stanford Law Review* 39: 571–605.

Sinclair, M. 1985. Law and language: The role of pragmatics in statutory interpretation. *University of Pittsburgh Law Review* 46: 373–420.

Smolka, J., and B. Pirker. 2016. International law and pragmatics. An account of interpretation in international law. *International Journal of Language and Law* 5: 1–40.

Soames, S. 2008. *Philosophical Essays, vol. 1. Natural language: What it means and how we use it*. Princeton: Princeton University Press.

Sperber, D., and D. Wilson. 1986. *Relevance: Communication and cognition*. Oxford: Blackwell.

Tarello, G. 1980. *L'interpretazione della legge*. Milan: Giuffrè.

Verheij, B. 2008. About the logical relations between cases and rules. In *Legal Knowledge and information systems. JURIX 2008: The 21th annual conference*, ed. E. Francesconi, G. Sartor, and D. Tiscornia, 21–32. Amsterdam: IOS Press.

Walton, D. 1995. *Argumentation schemes for presumptive reasoning*. Mahwah, N.J.: Routledge.

Walton, D. 2004. *Abductive reasoning*. Tuscaloosa: University of Alabama Press.

Walton, D. 2010. Similarity, precedent and argument from analogy. *Artificial Intelligence and Law* 3: 217–246.

Walton, D. 2015. *Goal-based reasoning for argumentation*. Cambridge: Cambridge University Press.

Walton, D., and T. Gordon. 2005. Critical questions in computational models of legal argument. In *Argumentation in artificial intelligence and law, IAAIL workshop series*, ed. P. Dunne, and T. Bench-Capon, 103–111. Nijmegen: Wolf Legal Publishers.

Walton, D., F. Macagno, and G. Sartor. 2014. Interpretative argumentation schemes. In *JURIX 2014: The 27th annual conference*, ed. R. Hoekstra, 21–22. New York: IOS Press.

Walton, D., C. Reed, and F. Macagno. 2008. *Argumentation schemes*. New York: Cambridge University Press.

Wilson, D. 2005. New directions for research on pragmatics and modularity. *Lingua* 8: 1129–1146.

Wilson, D., and D. Sperber. 2004. Relevance theory. In *Handbook of pragmatics*, ed. L. Horn, and G. Ward, 607–632. Oxford: Blackwell.

Varieties of Vagueness in the Law

Andrei Marmor

Vagueness in the law, as elsewhere, comes in different forms. Some of it is unavoidable, while other cases are optional and deliberately chosen by lawmakers. My main purpose in this essay is to distinguish between different types of vagueness in the legal context, and to explain their rationales. The argument proceeds in two main stages: The first part is taxonomical, mostly about the semantics of vagueness, and related linguistic indeterminacies, that we find in statutory language. The second part takes up each one of these different types of vagueness in law, suggesting some ways in which legal decision makers reason with vague language, and some of the normative considerations that apply, depending on the kind of vagueness involved.

1 Varieties of Vagueness

In philosophy of language, the term vagueness is used to designate a particular aspect of the relation between the words we use in a natural language and the objects picked out, or designated, by those words. Consider, for example, a word like "rich" applied to persons. Some people in the world are clearly and undoubtedly rich. The set of people who satisfy this condition, that they are undoubtedly rich, we call the *definite extension* of the word. Innumerable other people are clearly and undoubtedly poor, not rich; we call this set the *definite nonextension* of the word. And then, there are many borderline cases: These are people about whom, knowing all the relevant facts, there is no saying whether they are rich or not. From a semantic perspective, it would not be a mistake to say that "this person is rich," nor would it be a mistake to deny it, and say that "this person is not really, or not quite, rich." There is no answer—or,

A. Marmor (✉)
Cornell Law School, Cornell University, Ithaca, New York, NY, USA
e-mail: am2773@cornell.edu

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_19

according to some views, there is an answer, but it is not knowable[1]—to the question of whether a borderline case of "rich" is within the extension of the word or its nonextension.

Now, consider a particular person who is clearly and undoubtedly rich, if anyone is. Then, imagine that we subtract one cent from his possessions. Surely, he is still just as rich. Now subtract another cent, and he is still, undoubtedly, rich. But of course, if we continue this subtraction, at some point, we would have to doubt that the person is rich; the problem is that we cannot tell what that point is. There is no saying where exactly the borderline cases begin and where they end. This fuzziness of borderline cases is what gives rise to the famous *sorites* paradox.[2] We start with a true generalization, say, "Any person who has \$100 million is rich"; we add another true premise (called the induction step), saying that "If X is rich then X minus one cent is rich." Now the problem is that repeated applications of the induction step lead to a false conclusion. If you repeat the induction step many million times, your conclusion would have to be that a person who has very few dollars, or none at all, is rich, which is clearly false. And the problem, of course, is that we cannot tell where the induction step needs to be halted. There is no clear cutoff point in this (so-called) sorites sequence; there is no particular phase at which we can say that X is still rich, but X minus one cent is no longer rich.[3]

Lawmakers do not tend to use such obviously vague terms as "rich" (or "tall," "bald," etc.). If Congress wanted to impose higher taxes on rich people, for example, it would define the regulation much more precisely, using income figures in dollar terms.[4] And of course, the law can pick different figures for different purposes. But that does not mean that the law can avoid linguistic vagueness even when it has precise alternatives. Suppose that a law purports to impose a higher level of income tax on rich people and defines the higher tax bracket in terms of a precise dollar figure

[1] According to the epistemic theory of vagueness (mostly developed by Timothy Williamson *Vagueness* (Williamson 1994), there is a fact of the matter about the application of vague terms to what seems like borderline cases, but those facts are *not knowable*. The epistemic theory of vagueness is rather controversial, and in any case, I will not explore its possible implications in this essay. Mostly, it probably makes no difference, in the legal context, which particular theory of vagueness one works with. Soames (2012), however, argues that the ways in which we think about vagueness in the legal case may actually provide support to nonepistemic theories.

[2] What I call "fuzziness" of borderline cases is often called second-order vagueness, meaning vagueness about where borderline cases begin and where they end. As long as it is clear that there is no first-order vagueness without second-order vagueness, the terminology should not be problematic. Still, I prefer to avoid the notion of second-order vagueness because I doubt that this is a matter of hierarchy.

[3] Notably, the epistemic theories of vagueness (and some others) deny the truth of the induction step. The main motivation behind the epistemic theories, and some other theories offered in the literature, is precisely the idea that we have to avoid the sorites paradox, otherwise, we face serious problems with the principle of bivalence and the law of excluded middle in propositional logic. There is certainly no consensus in the literature about how deep the sorites paradox is and whether it is avoidable. I intend to take no stance on this complex issue.

[4] There are exceptions, of course. For example, in child support law in the USA, there are some federal guidelines, adopted by most states that provide an exemption to payors who have "an extraordinarily high income." I will discuss this case later on.

of annual income, say, at a million dollars. This would be a very precise definition, but then we might face borderline cases about what counts as "income" (e.g., a gift from a relative, even a very small one?). Even if "income" is defined by the relevant statute, the definition must use other words that are bound to have borderline cases. In short, the ordinary linguistic vagueness of general terms in a natural language cannot be avoided, thought its scope can be reduced in specific contexts.

Words like "rich," "mature," or "bald" are obviously and transparently vague. Other words are vague in exactly the same manner, but perhaps less obviously so. Consider, for example, a term like "entering" the premises, which forms part of the definition of burglary. Suppose that the defendant broke the window and had his arm through the boundary of the premises in question: Did he enter the building? And what if only his finger got through? Or only some instrument he was using to break the window?[5] Let me call these *ordinary* cases of vagueness, as opposed to words or expressions that are obviously or transparently vague, such as "rich" or "mature." Now you might think that obviousness (or transparency) is a vague criterion. That is true, of course, but in the legal case, there is a certain significance to the transparency of vagueness, in that the law typically tries to avoid it. We have countless laws using words such as "entering" or "premises," but very rarely laws using words such as "rich" or "mature." And it is an interesting question, I think, why that would be the case. After all, as we just saw, most general words we use in a natural language are vague, even if they do not carry their vagueness on their face. So why is it the case that the law strives to avoid one but not the other? Is it simply because some words like "mature" or "rich" are somehow more vague than others, or just too obviously vague? Here is a reason to suspect that this is not the only, or even the main, reason: The law does not shy away from using words that seem to be obviously very vague, even extravagantly so (to borrow a term coined by Endicott 2011, 24–25), such as "reasonable care," "due process," "neglect," "unconscionable." So why is it that we rarely, if ever, find legal norms using words such as "rich" or "mature," but we find countless legal norms that employ terms such as "reasonable" or "neglect"?

The answer resides in a very important difference between words that are *transparently* vague and those we are calling *extravagantly* vague. The essential feature of vagueness, in the strict semantic sense, consists in the fact that when a word, W, is vague, there are bound to be borderline cases of W's application to objects that are in a space between W's definite extension and definite nonextension, objects about which there is no saying whether W applies or not. In other words, if W is vague, then we are bound to have a sorites sequence.[6] This is clearly the case with words such as "rich," "mature," "bald." However, in the kind of cases Endicott calls "extravagant" vagueness, the main semantic feature is neither the obviousness nor the extent of a sorites sequence in the application of the word to concrete cases, though both would also be present, of course.

[5] See, for example, *Commonwealth v. Cotto* 52 Mass. App. Ct 225, 752 N.E.2d 768 (2001).

[6] I am not suggesting that this is the only semantic feature of words we can call vague in some sense or that there is a consensus in the philosophical literature about what vagueness really is. It is at least one standard sense of vagueness and that is how I use the term here.

To see this, lets work with the example that Endicott uses, of a UK statute making it an offense to cause a child to be "*neglected* […] in a manner likely to cause him unnecessary suffering or injury to health." The word neglect is, indeed, extravagantly vague. But it does not seem to be any more vague, so to speak, than "rich" or "mature"; it is not the case that we have more borderline cases here relative to the definite extension of the word. The main feature of extravagantly vague terms consists in the fact that they designate a *multidimensional* evaluation with (at least some) *incommensurable* constitutive elements. Neglecting a child is a very complex evaluative term. There are many potential elements that determine whether a certain case constitutes neglect. And, crucially, there is no common denominator that would allow a quantitative comparison of the various constitutive elements on a single evaluative scale. We can say, for example, that leaving a child unattended for 5 h is worse than leaving the same child unattended for 2 h or that it is worse to leave a two-year-old unattended for an hour than to leave a 6-year-old unattended for the same amount of time. But even these two simple factors (age and time) are not quite commensurable: Can we say whether it is worse to leave a two-year-old child unattended for 10 min than a 6-year-old unattended for 2 h? And of course, when you add more elements to the picture, such as the exact conditions in which the child was left unattended, the relevant environment, the child's level of maturity and the like, problems of incommensurability become obvious.[7] Needless to say, this does not mean that we are unlikely to face borderline cases. Extravagantly vague terms are also vague in the ordinary sense of vagueness. But it is the multidimensionality of such terms that makes them particularly problematic and particularly resistant to precisification.

The difference between ordinary vagueness, transparent or not, and extravagant vagueness might be a matter of degree. Even the simplest vague terms such as "bald" or "mature" are not single-dimensional; baldness, for example, might be a matter of both the number and the distribution of hair on a person's scalp. The main difference is, however, that the various elements that constitute the relevant predicate in standard cases of vagueness are not deeply and unavoidably incommensurable. Or, if some of them are, it is not typical for the incommensurability to pose a serious practical problem in determining whether an object or thing falls within the definite extension of the word or not. And this has an important consequence that explains the difference in their occurrence in law: When we have an ordinary vague term, with a sorites sequence, it is typically possible to stipulate a certain cutoff point in the sequence. Though such cutoff points are bound to be somewhat arbitrary, for the law to determine an arbitrarily chosen cutoff point is not an arbitrary decision. Consider familiar cases: We know, for example, that a certain level of maturity should be required for the exercise of certain rights, such as voting in elections. Maturity is a vague term, obviously susceptible to a sorites sequence. But the law tends to stipulate a fairly precise cutoff point, such as 18 years old for voting. Needless to

[7] I assume here that incommensurability is a relation between two (or more) items such that it is not true that one item is better or worse than the other, nor is it true that they are on a par with each other, according to the relevant evaluative dimension.

say, the exact figure of 18 years is both arbitrary and somewhat rough for the purpose.[8] But it is not arbitrary, in the sense of not being supported by reasons, to have such a cutoff point. For various obvious reasons of fairness and efficiency, it makes a lot of sense.[9] Furthermore, the cost of precisification such cases is very clear: Any reasonable cutoff point that the law chooses is bound to have some over—and some under—inclusiveness. There are going to be some persons older than 18 who are not mature enough to vote, and some persons younger than 18 who are actually mature enough to vote. Over—and under—inclusiveness is always the cost involved in such precisification. And, normally, we weigh this cost against the benefits of having a precise cutoff point. But now think about extravagantly vague terms: Here, the main problem is not the sorites sequence; the main problem in such cases is the incommensurability of the various elements constituting the multidimensional evaluation. And because the main problem is not the sorites sequence, such cases resist the stipulation of an arbitrary cutoff point. You just cannot stipulate that, say, leaving a child unattended for *n* hours would constitute neglect, even if you make *n* a variable relative to age. There are countless other factors in play, and they cannot be weighed with any precision against hours of un-attendance, the child's age, etc. And this is why the law cannot replace extravagantly vague terms with some stipulated precisification.

The kind of vagueness discussed so far is *semantic vagueness*, because it concerns the relations between the meaning of words and the objects they apply to. Vagueness, however, is not confined to the semantic aspect of language use. Expressions can be vague in the information they provide relative to a conversational context, whether they employ semantically vague terms or not. Suppose, for example, that during a political campaign, candidate *M* declares: "I did not receive any contributions from *X*, not a single dollar!" This statement does not seem particularly vague. But suppose the context is such that there is some suspicion that *X* channeled funds to the candidate's coffers indirectly, financing various organizations that are known to support *M*. Relative to *this context*, *M*'s statement might be rather vague. Or, suppose that in response to my friend's enquiries about a movie I saw last night, I express praise for one of the actors, going on and on about how well she performed her role. If my friend was interested in my opinion whether she should go and see the movie, my answer was probably too vague.

Conversational vagueness, as I will label such cases, does not have to be deliberately evasive. Expressions can be vague relative to a conversational context for a host of potential reasons, whether the speaker is deliberately evasive or not. The essential point here is that an expression that is not semantically vague can be vague relative to a specific conversation with respect to the information it contributes to the conversation. Contributions to a conversation can be more or less relevant. Some are clearly relevant and advance the common purpose of the conversation, and others are clearly irrelevant (or baffling, or conspicuously evasive, etc.), and then, there are

[8] By this roughness, I mean that we know that age is not the only dimension determining maturity, but it is the dominant one, and making the cutoff point determined by this single criterion, tough inaccurate and oversimplified, for sure, is not an obvious miss or a gross misconception.

[9] Endicott (2011) provides a very elegant account of these considerations in greater detail.

borderline cases in between. In short, the idea is that conversational vagueness is typically a function of the relevance of the speaker's contribution to the conversation in question. Relevance is a pragmatic aspect of speech, always relative to a specific conversation, its exact context, its normative framework, and various presuppositions taken for granted by the parties to the conversation. Borderline cases about relevance are, basically, what I call conversational vagueness.[10]

Notice that conversational vagueness is quite independent of semantic vagueness. Just as an expression can be conversationally vague in a given context without using vague terms, an expression can be precise even when it uses a vague term applied to a borderline case. Suppose, for example, that in responding to my wife's question, I say, "I wore the blue jacket." This may give my wife all the information she wanted to have, even if my jacket's color is actually a borderline case of blue (somewhere between very dark navy blue and black). If I only have one such jacket (say, the only other one I have is light brown), then by saying "the blue jacket" my expression picks out a singular object, relative to the conversational context that is mutually known to me and my wife. Similarly, referring to somebody in a conversation as "the tall guy," the speaker may well succeed in referring to a particular person, even if the person referred to is not particularly tall, for example, when the only other person one could have mentioned in the specific context is particularly short, and this is known to both parties to the conversation.[11]

Before I end this taxonomical section, let me mention two types of linguistic indeterminacy that should be kept separate from vagueness. I mention them here because legal cases are sometimes confused about them.

First, philosophers of language draw a sharp line between ambiguity and vagueness, and for good reasons. Normally, when we face an ambiguous expression, the assumption is that the speaker *intended* to use one of the two possible meanings; disambiguation, whether by context or other pragmatic factors, aims at figuring out the communication intention of the speaker in the specific context of the utterance. Furthermore, ambiguity, as opposed to many cases of vagueness, is typically avoidable. There are two main types of ambiguity in a natural language: syntactical and lexical. As an example of the former, consider the sentence: "I know a man who has a dog who has fleas." The sentence can be read in two ways: Either the man has fleas or the dog has them, and by itself, the sentence is indeterminate between these options. Lexical ambiguity concerns those cases in which a given word has two separate and unrelated meanings in the natural language in question, such as the word "bank" in English, meaning, in one sense, the side of a river and in a very different sense, a financial institution. This is the standard case.

It is possible, however, to extend the idea of semantic ambiguity to include expressions that have become idiomatically or colloquially ambiguous even if they are not

[10] It is possible that other conversational maxims, such as the maxim of quantity, also have borderline cases that would generate conversational vagueness in a similar way.

[11] It is not essential to my point here that the examples in the text have something to do with the distinction between referential and attributive uses of definite descriptions; other kinds of examples will be used later.

lexically so. Suppose, for example, that somebody asks me whether I use drugs. That depends, I would reply; if by "drugs" you mean hallucinatory substances, the answer is no. But if by "drugs" you mean to include medications, then yes, I regularly use prescription drugs. The word "drug" is not lexically ambiguous (like "bank"), because the two meanings are closely related, but it has come to be used, idiomatically in English, in a way that is, in effect, ambiguous. And this phenomenon is, I think, quite common, though how far it extends is not entirely clear.

Be this as it may, lexical ambiguity is a fairly special case. A much more prevalent aspect of meaning is *polysemy*. Consider, for example, the following two sets of utterances:

(1a) "I broke the window" (the window's glass)
(1b) "I opened the window" (the window's inner frame with the glass)
(1c) "I entered through the window" (the window's outer frame)

(2a) "John struggled to pull the cart out of the mud" (physical effort)
(2b) "John struggled to finish his dissertation in time" (intellectual effort).

As these examples demonstrate, the objects or features that words pick out within their definite extension—"window" in (1) and "struggle" in (2)—can vary with circumstances and contexts of expression. According to some views in semantics, polysemy is not an exception but the rule; the meaning of words, the information they encode, is very minimal, and hearers almost always work out the relevant meaning in the context of the conversation as they go along. I do not think that this minimalist (or sometimes called contextualist) semantics is quite right, but it is not my purpose to argue either way. Suffice it for our purposes to say that words often designate a particular subset of their semantic range, depending on the particular context of the conversation in which they are used. In some cases, polysemy is the opposite of vagueness; we face a sorties sequence when we need to extend the application of a word beyond its core, definite extension. Polysemy, on the other hand, often arises when the speaker refers only to a particular subset of the definite extension. (Not always, sometimes the distinction between polysemy and figurative or metaphoric use of a word is not all that clear.[12]) Both cases are prevalent in law, but they create very different interpretative challenges.

2 Vagueness in the Legal Context

Vagueness in legal language can arise in many different contexts: in legislation or agency regulations, in constitutional documents, in judicial decisions, in private con-

[12] Consider, for example, the various uses of the word "man," such as in, "Jo finally behaved like a man" (man as stereotype); "Marriage is a contract between a man and a woman" (man as adult male or gender); "Socrates is a man and therefore mortal" (man as a member of *Homo sapiens*). These kinds of examples are often given as examples of polysemy, and surely, there is a sense in which they are. But the use of "man" to stand for a stereotype can also be analyzed as a quasi-figurative use, one that goes well beyond the definite extension of the meaning of the word.

tracts and wills, etc. For simplicity's sake, I will focus on the context of statutory interpretation. Thus, the standard case I will consider here would be an act of legislation that contains some relevant expression that is, in one of the senses defined above, vague. Since the application of the law to particular cases crucially depends on what the law says, applying a legal prescription to a borderline case of a general term used in the relevant statute would seem to be a paradigmatic case of the problem of vagueness in the law.[13]

2.1 Ordinary Vagueness

Let us begin with a case of ordinary vagueness, using Hart's (1958) famous example (slightly modified). A city ordinance stipulates that "No motor vehicles are allowed in the park." Now, we know what motor vehicles are; the definite extension is pretty clear. But suppose that the question arises whether a bicycle powered by a small electric engine also counts as a "motor vehicle" for the purposes of this ordinance and thus prohibited from entering the park. Can we say whether an electric bicycle is a motor vehicle or not? The answer would seem to be that from a semantic perspective it can go either way. It would not seem to be a mistake to say that it is, nor would it be a mistake to say that it is not.[14]

We saw earlier that semantically vague expressions are not necessarily vague about the information they convey in a particular conversational setting. Speakers can use a vague term, even if applied to a semantically borderline case, to convey information that is precise enough in the specific context of the conversation. Applied to the kind of cases we are discussing here, this means that when a court faces a decision about the classification of a borderline case of a vague term, the semantic indeterminacy of the classification does not necessarily entail that the law says nothing about it. The context of the legislation, its overt purposes and similar pragmatic factors may determine an answer in some concrete cases. Suppose, for example, that in our case, the city ordinance was enacted in response to protests by residents about noise and pollution in the park. One can argue that such a context makes it clear that the ordinance was not intended to prohibit the use of vehicles that are neither noisy nor polluting, and hence it was not intended to prohibit electric bicycles.

Elsewhere, I have expressed some doubts about moving too swiftly from analysis of ordinary conversational settings to legal speech (see Marmor 2008, 2011). In an

[13] On the role of general (and very vague) concepts, as opposed to conceptions, in constitutional documents, I have written a separate paper, "Meaning and Belief in Constitutional Interpretation" (Marmor 2013). The concept v conceptions distinction raises many complicated issues that could not be dealt with here.

[14] Furthermore, it is easy to see how we get a sorites sequence here: Suppose we say that an electric bicycle is not a motor vehicle. Then, what about a small golf cart powered by an electric engine? A golf cart powered by a regular engine? A small scooter? And so on and so forth. As I explain in the text below, however, sorites sequence, which results from semantic features of words used, should not be confused with slippery-slope arguments, particularly of the causal-predictive type.

ordinary conversation, the context is usually rich enough to enable hearers to grasp the content conveyed by the speaker, even when the content asserted is somewhat different from the meaning of the words/sentences that the speaker uttered. In the legal case, however, context is often not rich enough to justify such inferences with a great deal of certainty. Perhaps the ordinance about vehicles in the park was initially motivated by the neighbors' protests about noise and pollution; perhaps it was enacted in response to such demands. But this would not necessarily entail that reducing noise and pollution are the exclusive purposes of the ordinance. Legislatures often use a particular social–political context to motivate an act of legislation, but then enact it with broader purposes, aiming to solve other problems in its vicinity as well. Truth be told, it is very difficult to generalize; sometimes the context of an act of legislation is clear enough to warrant conclusions about the assertive content of it, even if the particular case is semantically a borderline one.[15] More often, however, and perhaps in most cases, the context is just not sufficiently clear or determinate to justify such conclusions. And then, of course, semantically borderline cases remain genuine borderline cases, so to speak, and the court would need to make a reasoned decision about which way to classify the borderline case, given all the normative considerations that bear on it. Such decisions would not be an instance of applying the law but of extending it or narrowing it, that is, adding a precisification that goes beyond what the statute actually asserts. And, of course, courts often do just that.

It is, I think, quite impossible to suggest general guidelines about how courts should go about making such precisification in borderline cases; the considerations that bear on particular cases are enormously varied. But it is easy to say what kind of reasoning courts should avoid. They should avoid relying on the sorites paradox as a way of making a (type of) slippery-slope argument.

A *sorites slippery-slope argument* takes the following worry as an argument against the inclusion of a borderline case under a vague term: Let us say that the relevant expression is W, the definite extension of W is o_n, and let us assume that the court is asked to determine whether o_{n+1} is W or not. Now suppose the court reasons that o_{n+1} is W because it is very similar to o_n; it has almost all of the features that make an o W, just ever so slightly less so. So now we will have a ruling that o_{n+1} should be included under W. Then, the next case might come along, o_{n+1+1}, which is very similar to o_{n+1}, has almost all of the relevant features that make it W, just ever so slightly less so. Thus, a decision might be reached that o_{n+1+1} is also W. And then, the next case comes along … until we are bound to reach the conclusion that o_{n+m}

[15] A nice example is the case of *Garner v Burr* (1951), 1 KB 31. The British Road Traffic Act of 1930 stipulated that any "vehicle" traveling on a public highway must be fitted with pneumatic tires. Mr Burr fitted a poultry shed with iron wheels and pulled it with his tractor on a stretch of a highway. The court of appeals reasoned, quite sensibly, that even if a poultry shed fitted on wheels is not quite a vehicle, it counts as a vehicle for the purposes of the this law, because the manifest purpose of the law requiring pneumatic tires is simply to prevent damage to the asphalt roads. I have discussed this case in my "Textualism in Context" (Marmor 2012).

is also W, when clearly it is not. Therefore, the argument concludes, it would be a mistake to make the first step. Better not to decide that o_{n+1} is W from the start.[16]

Why is this a bad argument? Because it can be applied with equal force to *any* borderline case of a vague term. As we saw earlier, we get paradoxical results whenever we have a sorites sequence. An argument based on paradox is never a good argument. In other words, whenever we have a sorites sequence, we can easily construct a type of slippery-slope argument because the whole point of the sorites sequence is that there is no semantically determined cutoff point; there is no particular point at which we can say that the sequence needs to be halted, that it can go no further. Thus, pointing out that there is a kind of sorites slippery slope here is just stating the semantic feature of vagueness; no conclusion should follow from it. Any classification of a borderline case is going to be arbitrary from a linguistic point of view. The relevant question is whether there are good reasons to stipulate a certain arbitrary cutoff point or not, and if there are, what reasons would bear on the question of where the law should put it.

To be sure, suggesting that a sorites slippery slope is always a bad argument does not mean that there are no plausible versions of a slippery-slope argument that can apply to such cases. A *causal-predictive* version of a slippery-slope argument may well be relevant. The worry in the causal version is that the cutoff point stipulated by the court's decision is too far removed from the definite extension so that actual, real-life factors may cause the legal consequences to slip too far down the road to undesirable results. Notice that the nature of such an argument is empirical and predictive: The worry is that if the court includes a given borderline case under a vague term, then future decision makers, such as agencies, lower courts, or even the same court in future cases, might find it difficult to resist the temptation to go further down the road, reaching results that one finds objectionable. Now this is a matter of prediction, and such arguments tend to be rather speculative. Though the concern is empirical in nature about matters of fact, it is not unrelated to the nature of vagueness. The concern that motivates causal slippery-slope arguments derives its force from the fuzziness of borderline cases, from the fact that there is no obvious or salient cutoff point that can warn us against slipping down the road too far from the original reasons that justified the legal rule in question.

Given the speculative nature of causal slippery-slope arguments, they should always be treated with great caution. The burden of proof should be high, because it is the nature of slippery-slope arguments that they counsel against doing something that would be the right decision on the merits of the case at hand, only due to a fear that future decisions are likely to lead us astray. Thus, at the very least, the argument should provide sufficient evidence that likely errors in the future will be difficult to avoid. Gut feelings and speculations, which are mostly what one finds in such cases, should not be enough.

[16] See, for example, *Randall v. Orange County Council Boy Scouts of America*, 17 Cal. 4th 736, 952 P. 2d 261, 72 Cal. Rptr. 2d 453 (1998) where one of the main worries of the dissenting judge relies on this kind of argument.

2.2 Transparent Vagueness

Transparently vague terms, such as "tall," "mature," "rich," are rarely found in statutory language. But they are not entirely absent. In some cases, and typically in addition to a set of much more precise regulations, one finds that the law includes a transparently vague term as part of its regulatory scheme. Often the purpose of such vague additions to a regulatory scheme in a given area is preemptive: Legislatures want to safeguard against the possibility that some unpredictable, yet clearly wrong (or otherwise relevant), conduct does not fall between the cracks of the set of precise rules that purport to govern the area in question. There is an endless variety of permutations; one just cannot predict them all.[17]

Be this as it may, the most obvious aspect of legislating transparently vague standards, whether in the kind of cases discussed above or others, is that the legislature in effect delegates the decision of how to make the standard more specific to the courts or to administrative agencies. Using vague legislative language is, actually, the main technique for legislatures to delegate power to the courts without explicitly saying that this is what they are doing. Accordingly, there are two main types of reasons for opting for such transparently vague regulation: Sometimes the vague language is simply a result of a compromise between legislators enacting the bill. Legislators often have conflicting aims or intentions with respect to a bill they would want to enact, and if neither side can muster the requisite majority for their position, opposing sides may settle on wording that is sufficiently vague to let each party hope that their specific purposes might win the day in future decisions by the relevant courts or agencies that get to interpret the act in question. (Or, I presume, sometimes there is no such hope, only the attempt to conceal from the constituency that one gave up.[18])

Compromise, by its nature, is regarded by the parties to it as second best. The use of vagueness in the law, however, is not confined to such second-best choices. Sometimes there are good reasons to opt for a vague term in an act of legislation as a means of delegating the decision to the courts, and those are mostly the kind of cases I want to focus on. But let me answer an objection here before we proceed.

Some writers suggest that vaguely worded regulation in some areas is justified by the need to delegate the relevant decisions not to the courts but to the law's subjects, to those whose behavior the law purports to regulate. The idea is that in some cases, instead of telling people what it is exactly that they ought to do (or not to do), it is better to set a vague standard, leaving it for the subjects themselves to exercise their own discretion and take responsibility for the choices they make. So when the law tells drivers that, no matter what, you ought to drive *carefully*, the law imposes on the drivers themselves the responsibility to determine what careful driving is under the

[17] Traffic regulations often have some kind of a requirement to drive with reasonable attention to the conditions of the road. There are also countless such examples in US tax legislation. For instance, section 541 imposes an accumulated earnings tax on corporate-retained earnings beyond those retained "for the reasonable needs of the business." Section 535(c): Tax-free mergers are typically conditioned on the transaction having a corporate "business purpose," etc.

[18] I explained this in greater detail in Marmor (2011, 97).

circumstances. It is your responsibility, the law says to its subjects, to determine what is right or reasonable under the circumstances, and you need to bear the consequences of your own choices. And, some writers claim, this is sometimes a very good idea; it is as it should be (Waldron 2011; Endicott 2011).

Presumably, the attractive feature of this rationale for vague regulation is that it respects people's autonomy or, at least, forces people to take responsibility for their decisions. Writers who like this idea admit that in some cases, there might be a concern about chilling effect, but I think that they underestimate the normative problem here. Consider this example: My teenage daughter is going out on a Saturday night. I want to make sure that she gets back home at a reasonable hour, so I face a choice here: I can either tell her, "Make sure to be home by no later than 2 a.m.," or I can be much more vague, and tell her, "Make sure not to come home too late!" Now, we might think that the vague instruction is more respectful of my daughter's autonomy. It is more educational, in a sense, too, because it makes her more responsible for her own actions, which is generally a good attitude to foster. So far, so good. But now let us suppose that there is a sanction looming here, that is, suppose that my daughter knows that I am the one who gets to determine what would count as "too late" to come home and that if I decide that she came home too late, I get to impose a penalty. And let us further suppose that she cannot be sure, in fact she only has a vague sense of, what I would consider "too late" under the circumstances. (Notice that if my daughter knows exactly what I have in mind when I say "too late," then my instruction is no longer really vague.) Now we might begin to doubt that the vague instruction is more conducive to her autonomy. In all likelihood, it might have a chilling effect. If the sanction is not trivial, she would need to play it safe and err on the side of caution, and the more threatening the sanction is, the greater the margin of safety she would need to allow.

The legal case is, of course, in line with the latter part of the example. When the law regulates conduct with vague standards, it puts the decision about sanctions for violation in the hands of the courts, and it is the court that gets to determine, ex post, whether the subject violated the standard or not. Therefore, the real effect of such vague regulation is transferring to the subjects not the kind of decision that is respectful of their autonomy or moral agency, but the burden of trying to predict what the courts will decide. And the less information they have about it, and/or the more severe the cost of violation, the more the subjects would need to err on the side of caution. Perhaps in some cases, this legislative strategy is efficient or justified, but I do not quite see how, morally speaking, it is particularly respectful of the subjects' moral agency.[19]

None of this is meant to suggest that there are no cases in which there are good reasons to delegate the decision about precisification of vague standards to the courts. On the contrary, there are many such cases. But the rationale of delegation of power

[19] In fact, the problem is often more severe, because a serious concern about fairness also comes into the picture. The more vague a legal regulation is, in the sense discussed here, the more crucial it becomes for potential litigants to have information that enables them to predict courts' decisions, which gives repeat players, mostly large corporations, considerable advantage over ordinary citizens.

must be derived from considerations pertaining to the relative institutional competence of legislatures vis-à-vis the courts or other decision-making agencies. That is the real choice here, not the concern for the subjects' autonomy. Furthermore, I will argue that there is typically a much stronger case for the legislature to delegate to the courts decisions about specifications of extravagantly vague terms than cases in which the legislature uses a transparently vague term. In any case, different kinds of reasons apply here.

What reasons, if any, might legislatures have for using a transparently vague term as a means of delegating the precisification of a vague standard to the courts? Remember that the main problem in such cases is to set a cutoff point in the sorites sequence. Is there any reason to think that the courts will do a better job in that? Generally speaking, probably not. The relatively infrequent use of transparently vague terms in legislation suggests that it is generally recognized that legislatures are better equipped to make those kind of decisions compared with the courts. Furthermore, it is worth keeping in mind that courts' decisions, which are based on particular cases adjudicated, inevitably have a retroactive effect; unlike legislative acts and guidelines issued in advance, the decision of a court applies to conduct that has already occurred and determines a resolution to the case ex post. So there is always some cost of retroactivity involved in judicial, as opposed to legislative, decisions.[20]

I want to suggest that there is a type of cases where delegating to the courts the decision of determining the particular cutoff point in a sorites sequence makes a lot of sense, namely when the following two conditions obtain: First, the precisification is particularly context sensitive. Second, it is an area of conduct where parties concerned do not have good reasons to know in advance the exact regulatory content that applies to them. Consider, for example, the law I mentioned earlier in a note, of granting exemption from the federal guidelines concerning child support on the basis of "extraordinarily high income." Why not have the legislature stipulate a certain income figure as a cutoff point? The answer is twofold: First, the relevant considerations are very context sensitive. The rationale of the exemption has something to do with the fact that the needs of children are not unlimited that there is no reason to allow either the children or the custodial parent to have an extravagant lifestyle. But, of course, these things tend to vary a great deal with particular circumstances. The difference between comfort and luxury profoundly depends on the environment in which one lives, the kind of opportunities available to others in one's vicinity, etc. Therefore, it makes a lot of sense to avoid a generally stipulated cutoff point and allow the courts to set it on a case-by-case basis.

Furthermore, there is a distinction between the kind of regulations where it is important for the law's subjects to know, ex ante, what the law requires, and those cases in which prior knowledge of the exact legal regulation is not very important. In most cases, when we plan our conduct in a given area, knowing what the law requires

[20] In some areas, precisification requires a great deal of expertise, of the kind that legislatures typically lack. But in such cases, legislatures tend to delegate the decisions to administrative agencies, not so much to the courts. And administrative agencies tend to issue detailed general guidelines, not case-by-case decisions.

or permits is of crucial importance. But not in all cases. And child support belongs to the latter. Parents do not (and certainly should not) plan separation or divorce from their spouse on the basis of considerations about the exact amount of child support they will be required to pay. Ex post determination of such matters does not frustrate legitimate expectations, as long as those determinations are within reason, of course.[21] Thus, allowing the courts to determine, on a case-by-case basis, the exact cutoff point in a sorites sequence—which would normally have a retroactive, ex post facto element—is not normatively problematic in such cases. Notice, however, that when the rationale of opting for transparently vague terms consists in the particular context sensitivity of the relevant factors, decisions made by the courts on a case-by-case basis should have very limited precedential effect. If the whole point of letting the courts decide such matters derives from the inherent difficulties involved in ex ante generalizations, then granting courts' decisions too much precedential effect would defeat the rationale of the delegation of power to the courts.

2.3 Extravagant Vagueness

It may seem paradoxical that legislatures have much stronger reasons to delegate decisions to the courts when the relevant concept in play is extravagantly vague. But that is actually the case, and the ubiquity of such terms in legislation might attest to the fact. Let me explain why this is so. Remember that the main feature of extravagantly vague terms consists in their multidimensional aspect, not so much in the sorites sequence they entail. Thus, it might help if we focus our attention on the ways in which we make choices or decisions in cases involving multidimensional and incommensurable elements. Suppose, for example, that one of your colleagues received an offer to move to a different job in a different city, and she needs to make up her mind about whether to accept it. So let us assume that this is what she knows: The new job pays better, but it will involve a bit more teaching; she will be able to afford better housing, though the commute will be a bit longer; the faculty in the new place is probably stronger than in her current department, but the quality of the graduate students is not as high. And then, there is the fact that she will need to move from a small college town to a big city, with all the differences involved in that. And so on and so forth.

The essential point is that the various factors that your colleague needs to consider present her with a problem of incommensurability. It is very difficult to decide, on rational grounds, how much better the housing has to be relative to a given addition of commute time, and even more difficult to think about how to compare, say, better housing with more teaching or lower quality of students. So how can one make

[21] This idea is supported by the fact that most US states do not allow child support arrangements to form part of a prenuptial agreement. US tax legislation, as I mentioned earlier, is also replete with transparently vague terms. It is not all that surprising, given the fact that in US federal tax law, quite generally, retroactivity is not regarded as a major concern.

a rational decision in such cases? There seem to be two main possibilities here: In some cases, a particular factor stands out as more or less decisive. Your colleague may think, for example, that improving her housing condition is much more important to her than other considerations in play, and then, she would assign it a much greater weight in her deliberation. And this may well tilt the balance in favor of a particular decision. However, if no such decisive factor is in play, her only choice is to make an all-things-considered holistic judgment here; she would need to take everything she deems relevant to the choice into account and decide between the two packages, as it were, in a holistic manner. She would look at the whole thing, so to speak, and ask herself which one seems more attractive, overall. There is no guarantee, of course, that such a holistic method would yield a reasoned preference for one of the options, but then she might as well just flip a coin.

Now consider the example we used earlier, of an extravagantly vague legal norm such as "neglecting a child." Suppose you are presented with the particular facts of a given case that would seem to be a case of neglect. Like in the job-offer example, it is quite possible that a particular fact of the case stands out as more or less decisive. Upon hearing that the care provider left a baby in a bathtub full of water unattended for half an hour, you may not need to hear much more. But of course, many cases are not like that. In many actual cases, there is not any particular conduct that decisively counts as neglect, but the overall behavior of the care provider, over time and in varying circumstances, might well amount to criminal neglect. And you can only make this kind of judgment holistically, looking at the whole package, so to speak. And again, the whole package may not give you a decisive answer; borderline cases cannot be ruled out.[22]

I hope we can see the reason for trying to avoid ex ante specifications of how to resolve such issues. Just as it would make very little sense to decide in advance how you would react to any job offer you might receive in the indefinite future, or to try to make yourself a list of specific conditions that such an offer would have to meet (and to what extent) for you to accept it, it makes little sense for the law to try to legislate in any great detail what counts as neglecting a child. Even if one can think in advance of some factors that may stand out as decisive, often there are no such decisive factors in play, only all-things-considered, holistic judgments to make. And of course, if these kinds of decisions cannot be made ex ante, legislatures have no choice but to delegate the decisions to the courts on a case-by-case basis. And here too, for reasons we mentioned earlier, it would be a mistake to assign courts' decisions in particular cases great precedential value. The whole point of delegating such decisions to the courts is that they have to be made ex post, on the basis of the particular features of the case at hand.

Some of the more familiar examples of extravagantly vague terms in law are a bit more complex than that, because they tend to be partially defined. Consider,

[22] In some rare cases in the US vague statutory references to a child's "welfare" or a child's "neglect" have been struck down as unconstitutionally vague (see, for example, *Roe v Conn*, 417 F. Supp. 769, (1976).) Most of these cases, as *Roe v Conn* exemplifies, are entangled with problems of racial discrimination and racial bias at the enforcement level, and I am told by experts that these kinds of issues are almost always lurking in the background of void for vagueness constitutional cases.

for example, the use of the word "corruption" in the context of bribery laws: The definition of bribery under federal law (18 USC 201) defines bribery as "*corruptly* giving, offering, or promising anything of value to a public official or candidate to influence any official act." The word "corruption" is, no doubt, extravagantly vague. Very much like "neglect," in most cases the determination of whether a given set of circumstances amounts to corruption has to be made contextually and holistically. But the law does not quite leave it at that. Corruption is partly defined by various rule-like decisions, such as requiring some *quid pro quo* element, whereby merely gaining access to officials does not count as corrupt. So we end up here with a tension between two kinds of reasoning: On the one hand, we have the extravagantly vague term, aiming to allow the courts to form a holistic, all-things-considered judgment of the particular case at hand; on the other hand, we have some specific rules that are aimed to shape such decisions and determine, in advance, *some of the conditions* that the relevant conduct has to meet to count as corrupt. This compromise between different types of legal regulation, embodied in the partial definition of a transparently vague term such as corruption, reflects the fact that the law needs to set some fairly specific guidelines in advance, but that there is a limit to how specific those guidelines should be. The limit, however, is not epistemic; it does not derive from lack of knowledge or limited foresight. It derives from the multidimensionality of the evaluative elements that constitute the idea of corruption. And of course, corruption is just one example. Similar considerations apply to legal concepts such as "due process," considerations of "equity," protection of "privacy." I venture to speculate that most extravagantly vague terms deployed in the law are partially defined.

2.4 Ambiguity and Polysemy

Linguistic indeterminacy in law is not confined to vagueness. Ambiguity, and as I will argue, much more frequently, cases of polysemy, are different forms of linguistic indeterminacy that we also find in law. Standard lexical ambiguity is rarely a problem. Since the standard case of lexical ambiguity concerns words whose different meanings are unrelated, context is usually clear enough to determine which one of the two meanings of the word was intended by the legislature. Syntactical ambiguity is a bit more prevalent and typically inadvertent. Legislatures make an effort to avoid syntactical ambiguities; sometimes they fail, of course, which is typically unfortunate, as syntactical ambiguity serves no useful purpose.[23]

The problematic, and much more interesting, cases are those in which the indeterminacy is due to polysemy. Consider the famous case of *Smith v. US* (508 US 223 1993). The relevant statute mandated a much harsher punishment for drug-related crimes if the defendant was "using a firearm" during the drug-related activity. In the *Smith* case, the defendant used a firearm in a barter deal in exchange for the drugs. So the question was whether using a firearm as an object of value, not as a weapon,

[23] Scope ambiguity is the typical case of syntactical ambiguity we find in legislative language.

counts as "using a firearm" in connection with a drug deal. The majority decided affirmatively, but in a famous dissent, Justice Scalia argued that there is no linguistic indeterminacy here whatsoever. Using an object is normally understood as using it for its intended purpose or function, not for just about any use whatsoever. Hence, he argued that the expression "using a firearm" only applies to cases in which the firearm is used as a weapon, not as an object of value for a barter deal. His main argument was based on the thesis that "using a firearm" is simply not ambiguous. And, in a sense, though a somewhat different sense from what he meant, Scalia is right; this case is about polysemy, not ambiguity.

Consider again the pair of sentences:

(a) "John struggled to pull his cart out of the mud"
(b) "John struggled to finish his dissertation on time"

Two points worth noting here: The word "struggle" in (a) stands for something different from "struggle" in (b); But it is equally clear that we do not need any particular contextual information to understand the relevant extension of the word "struggle" in the sentence uttered. Our ordinary background knowledge of the world, so to speak, is sufficient to determine which one of the references is meant. The word "struggle" is polysemous but not ambiguous (the two meanings in context are closely related, well within the semantic range of the word). In this respect, I think that Justice Scalia is quite right about the fact that there is no ambiguity involved in "using a firearm." In fact, we can easily construct a similar pair of sentences about the expression "using an x":

(a*) "Jane uses a laptop"
(b*) "Jane uses a laptop to keep the door open"

We can assume that (a*) refers to using the laptop as a computer, and with equal certainly, we can infer that in (b*) the laptop is used as a doorstop. But again, we do not need any contextual knowledge of the specific speech situation to understand which one of the meanings is intended by the speaker; our ordinary background knowledge of the world is sufficient. So it is quite right that the expression "using an x" is not semantically ambiguous. The problem in *Smith* is about polysemy. The word "use" has a very wide semantic range; when we use an expression like "using an x," we may designate a specific subset of the word's definite extension. And in most cases, this is clear enough from the meaning of the relevant sentence, combined with our background knowledge of relevant aspects of the world.

Having said this, Scalia's conclusion would be correct if it is generally the case that an unqualified (without anaphora) use of "use" is normally understood in a restricted extension within the wide semantic range of the word. If it is generally true that an expression of the form "*A* uses an *x*" is understood by default in the restricted sense of using the *x* for *x*'s typical purpose or function, then Scalia is right. And that seems quite plausible.[24]

[24] Technically speaking, this is probably an example of what Grice called "generalized conversational implicature," such as the expression "an *X*," without anaphora or further clarification, normally

However, I doubt that this is generally the case with similar examples of polysemy. Notice that this does not work with words like "struggle." The expression "*A* struggled to φ" does not indicate what kind of effort "struggle" designates if we do not know the nature of φ. Or, consider the case often mentioned in parallel with *Smith*, the case of *Muscarello v. US* (524 US 125 1998): The question was whether "carrying a firearm" in relation to a drug deal applies to carrying it in the trunk of the defendant's car. Once again, Scalia used the same argument to conclude that it does not. But here I think that he was mistaken. I doubt that the expression "carrying an *x*," without anaphoric addition, by default refers to carrying it on one's body. If I ask a driver who happens to have a flat tire, "Are you carrying a spare tire?" he would respond quite sensibly by saying "Yes, it's in the trunk of my car." Or, the sentence, "The accident victim was carried to the hospital," would certainly not imply that he was carried on somebody's person. When I ask my wife whether she happens to carry some cash with her, I would refer to carrying it in her purse, of course, not in the trunk of her car. But that is so because we know that people normally carry cash in their wallet or purse. In other words, polysemy applies to words used in a given context to designate a particular subset of objects within the word's semantic range. The relevant context is typically given by our general background knowledge of how things are in the world. The context does not have to be specific to the particular conversation in question. However, it is not generally the case, as Scalia seems to assume, that by default nonanaphoric expressions of such words are normally understood in a restricted, narrow sense. It all depends on the nature of the object or instance on which the word is predicated, and things we generally know about the relevant aspects of the world. To conclude: From a semantic perspective, Scalia's reasoning was probably correct in *Smith*, but incorrect in *Muscarello*. And I suspect that most cases of polysemy are like the latter.

2.5 Conversational Vagueness

Let us recall that expressions using a vague term, or any other form of semantically indeterminate expression for that matter, can be sufficiently precise in the specific context of the conversation. In fact, though not dealing with vagueness, the majority opinion in *Smith* argued along these lines: The argument was, in effect, that even if "using a firearm" would normally be understood as using the firearm as a weapon, the context of the legal regulation here makes it clear that the word was intended in its wide semantic sense to apply to any use whatsoever. And that is so, the majority argued, because the legislature clearly wanted to act against the dangers of mixing drugs with guns, given the well-known dangers involved in the potential for deadly violence in drug deals. Hence, the mere presence of a firearm in a drug deal is precisely

implicating that the speaker has no particular knowledge about the specifics of *X* or does not deem it relevant to the utterance in question. See Grice, (1989, 37). Similarly, the expression "*A* uses an *X*" would normally implicate that *A* uses the *X* for *X*'s typical function or purpose.

what the legislature wanted to discourage. Let us assume that the majority is correct in its assumptions about the background purposes of this piece of legislation. Is it also correct to claim that the pragmatic elements of the regulation here determine the content of the legal speech with sufficient certainty to warrant the conclusion that it favors? I suspect that most readers will doubt that this is the case. Perhaps it is true that the legislature would have decided in this case, if it had considered it, as the majority ruled; but we must bear in mind that speakers often fail to convey all that they had intended to convey on an occasion of speech. For a speaker to be able to assert some content that differs from what his or her sentence semantically means, the contextual knowledge shared between speaker and hearer has to be quite rich. I doubt that the context is rich enough in this case to warrant such a conclusion with sufficient certainty, especially given the fact that the relevant legislation here is in the criminal law domain.

I want to conclude this discussion, however, with the opposite type of case, where we have a legal formulation that is semantically precise relative to a certain object or instance of application, but conversationally vague in the context of its utterance. To illustrate the (very limited) point that I want to make here, consider the case of *FDA v. Brown & Williamson Tobacco Corp* (529 US 120 2000).[25] The question in this case was whether the FDA was granted the authority to regulate tobacco products. The relevant part of the statute defining the FDA's authority to regulate drugs said that the FDA has the authority to regulate "articles (other than food) intended to affect the structure or any function of the body." Now, if you think about it from a semantic perspective, surely you would think that cigarettes and other tobacco products are intended to do just that, "affect the [...] function of the body," and hence, they are clearly within the *definite extension* of the relevant expression here. So why is this regulation conversationally vague in the context? The answer, which gave rise to this famous litigation, consists in the combination of two additional facts: First, the legal fact that if tobacco falls under the jurisdiction of the FDA, other parts of the statute render it clear that the FDA must prohibit its sale. Second, the fact that between 1965 and the time of the court's decision in 2000, Congress enacted six separate pieces of legislation regulating the sale, advertisement, etc., of tobacco products, clearly presupposing that the general sale of tobacco products is perfectly legal. Thus, the conflict between different pieces of legislation here, and their accompanying presuppositions, renders it questionable whether tobacco products fall within the ambit of the authority granted to the FDA. In the overall context of tobacco regulation, the relevant statutory expression is conversationally vague, even if it is not an instance of a semantically borderline case.[26] And there is an interesting lesson here: Contextual knowledge is often deemed helpful in determining some asserted content that would otherwise be under-determined or vague. Sometimes, however, the opposite is the

[25] I am certainly not suggesting that this is the only issue that is central to this complicated case, nor that it bears on the desirable result.

[26] In one clear sense, this is an oversimplification because I ignore the time sequence between the different pieces of legislation. But my point in the text is not to analyze the case, only to illustrate a general point.

case; an expression that is not particularly vague or indeterminate becomes pragmatically or conversationally vague precisely because the particular context of the conversation makes it doubtful that the expression applies to its ordinary semantic extension. Given the complex contextual background of legal regulations, I suspect that conversational vagueness in law is much more common than one might have thought. Sometimes context makes thing less, rather than more, clear.[27]

References

Endicott, T. 2011. The value of vagueness. In *The Philosophical Foundations of Language in the Law*, ed. A. Marmor, and S. Soames, 14–30. Oxford: Oxford University Press.

Grice, H.P. 1989. *Studies in the Way of Words*. Cambridge, Mass.: Harvard University Press.

Hart, H.L.A. 1958. Positivism and the separation of law and morals. *Harvard Law Review* 71: 593–629.

Marmor, A. 2008. The pragmatics of legal language. *Ratio Juris* 21: 423–452.

Marmor, A. 2011. Can the law imply more than it says? On some pragmatic aspects of strategic speech. In *The Philosophical Foundations of Language in the Law*, ed. A. Marmor, and S. Soames, 83–104. Oxford: Oxford University Press.

Marmor, A. 2012. Textualism in context. *USC Law Legal Studies Paper No.* 12–13 https://ssrn.com/abstract=2112384.

Marmor, A. 2013. Meaning and belief in constitutional interpretation. *Fordham Law Review* 2: 577–596. http://ir.lawnet.fordham.edu/flr/vol82/iss2/8.

Soames, S. 2012. Vagueness in the law. In *The Routledge Companion to Philosophy of Law*, ed. A. Marmor, 95–108. London: Routledge.

Waldron, J. 2011. Vagueness and the guidance of action. In *The Philosophical Foundations of Language in the Law*, ed. A. Marmor, and S. Soames, 58–82. Oxford: Oxford University Press.

Williamson, T. 1994. *Vagueness*. London: Routledge.

[27] I am grateful to the participants of the legal theory workshop at UCLA (Feb 2013), the participants of the conference on *Vagueness in Law* at the NYU Department of Philosophy (March 2013), and the participants of the legal theory workshop at the Julius Stone Institute of Jurisprudence at the University of Sydney (April 2013) and to David Enoch and Alon Harel for helpful comments on a draft of this paper.

Balancing, Proportionality and Constitutional Rights

Giorgio Bongiovanni and Chiara Valentini

1 Introduction

In the theory and practice of constitutional adjudication, proportionality review plays a crucial role. At a theoretical level, it lies at core of the debate on rights adjudication; in judicial practice, it is a widespread decision-making model that is increasingly characterizing the action of constitutional, supra-national and international courts. Despite its circulation and centrality in contemporary legal discourse, proportionality in rights adjudication is still extremely controversial with regard to its justification and limits, and also to its nature and distinctive features. As for the first aspect, proportionality raises questions of "justification," concerning the normative basis and limits of its use in rights adjudication; as for the second aspect, it raises questions of "identification," concerning its nature and distinctive features. So far, this second order of questions has remained in the background of analyses that have mostly been concerned with questions of justification and have tended to identify proportionality with its "standard" form—the prominent version in the theory and practice of rights adjudication—without dwelling on the other forms that proportionality can take. Indeed, important questions of identification are still open concerning what proportionality *is*, what forms it can take, and how these forms differ one from the other. These questions are relevant in the first place at a descriptive level, since capturing

Part I was written by Giorgio Bongiovanni and Introduction and Part II was written by Chiara Valentini. Chiara Valentini acknowledges support from the Spanish Ministry of Economy and Competitiveness ("Ramón y Cajal" fellowship—RYC-2012-12175).

G. Bongiovanni (✉)
Dipartimento di Scienze Giuridiche and CIRSFID, Università di Bologna, Bologna, Italy
e-mail: giorgio.bongiovanni@unibo.it

C. Valentini
Department of Law, Universitat Pompeu Fabra, Barcelona, Spain
e-mail: chiara.valentini@upf.edu

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_20

the distinctive features of proportionality is necessary for an adequate representation of the complex judicial practices based on proportionality review. Furthermore, these questions are relevant at a normative level because an adequate representation of proportionality review contributes to the analytical framework within which we can argue about the reasons for or against this review template.

This contribution will be divided into two parts. Part I centres on (a) the connection between the foundation of proportionality, balancing and theories of rights and (b) the critical aspects of this connection. Part II analyses the different forms of proportionality both (c) *in* review, as a template for rights adjudication, and (d) *of* review, as a way of defining the scope and limits of adjudication.

PART I

2 Rights, Balancing, Proportionality[1]

It is almost a truism in the contemporary legal debate that a link obtains between the constitutional adjudication of rights and proportionality. In this sense, as many authors have noted, it is possible to claim that "to speak of human rights in the twenty-first century is to speak of proportionality" (Huscroft et al. 2014, 1) and, in the same direction, that we can "describe" proportionality "as *the* central concept of contemporary constitutional rights law" (Gardbaum 2016, 1).[2] This centrality can be explained in relation to two main aspects, among others, that can be mentioned: on the one hand is the emergence of new theories on the nature of rights broadly understood, on the other is the resulting "shift from a culture of authority to a culture of justification" (Cohen-Eliya and Porat 2011, 463). The first aspect gives an account, from different perspectives, of the role and nature of rights and of the (necessary) processes of their implementation, while the second not only refers to a change in the legal culture but also highlights the changes that have taken place in contemporary legal systems (especially constitutional ones). These two perspectives have many points of contact, and in some respects, they engage each other: a "broad" conception of rights can be seen as one of the (necessary) conditions for developing a culture of justification (ibid.). These aspects refer to two different phenomena: while the theory of rights develops the fundamental justifications of this judicial method and the thesis of the connection between rights and proportionality, the analysis based on the culture of justification emphasizes the role of proportionality in processes by which decision-making power is legitimized. However, the theory of rights can also be used in the opposite direction, namely, to argue that there is no direct link between rights and proportionality and to highlight the shortcomings of the method. In this

[1] In this part, we use balancing and proportionality as terms that imply each other, that is as terms that are in a "close" relationship (Schlink 2012a, 721). For some authors on the contrary (*Infra* I, 2.2.), the two concepts should be separated as they express different ways of rights adjudication.

[2] Cohen-Eliya and Porat (2011, 465), note that "the spread of proportionality has been well documented, and is an undisputed fact."

Part (I), we will analyse the relation between theories of rights and proportionality, while the role of the culture of justification will be discussed in Part II.

The analysis of theories of rights will be developed in light of the relation between rights and proportionality: we will have, on the one hand, those who claim that, with different degrees of necessity, a connection does hold between these two elements, and on the other hand, those who argue that such a connection does not exist or at least is contingent, and that the judgment of proportionality maybe inadequate for the realization of rights. In the first case, we will refer to specific rights theories, while in the second reference will be made to the main problems that come with the use of a proportionality judgment.

3 Theories of the Connection Between Rights and Proportionality

3.1 Interest Theory and Balancing

One of the major aspects of the twentieth-century rights theory developed in the Anglo-Saxon context is the affirmation of the *interest theory* and of the general idea that rights are reasons. This idea marks the passage from the static theories of rights to the dynamic one and makes it possible to move beyond the so-called axiom of correlativity (Kramer 1998, 42), that is, the idea that a right exists if, and only if, there is a correlative duty.[3] On the interest theory, as is known, a right is viewed as a "cluster" of rights,[4] or as a "molecular" right (Wenar 2015),[5] and rights are crucially understood to have the function of protecting interests essential to the well-being of individuals or groups. Interest theory was developed in response to the *choice (or will) theory*,[6] in an effort to overcome the latter's limits.[7] As its name suggests, the

[3] A "dynamic aspect of rights" is described by Raz (1986, 171) as one that ascribes to rights an "ability to create new duties," an aspect that in his view "is fundamental to any understanding of their nature and function in practical thought." Kramer (1998, 41) underlines that the "the notion of strict correlativity between rights and duties does indeed obscure" this aspect. For Raz (ibid.), "unfortunately, most if not all formulations of the correlativity thesis disregards the dynamic aspect of rights. They all assume that a right can be exhaustively stated by stating those duties which it has already established."

[4] See Thomson (1992, 55), defining such clusters as "rights that contain other rights."

[5] A molecular right is a complex of Hofheldian normative positions. Wenar (2015) defines these rights as "atomic incidents [privilege, claim, power, and immunity] bond together in characteristic ways to form complex rights."

[6] Central to choice theory is the thesis that the function of rights is to ensure for the right holder, in relation to the duties of others, a choice respect to the fulfilment or not of a duty. Wenar (2015, referring to Hart 1982, 183) describes this position in this way: "Will theorists maintain that a right makes the right holder 'a small scale sovereign' […]. More specifically, a will theorist asserts that the function of a right is to give its holder control over another's duty."

[7] Its main shortcoming is that it fails to account for some situations involving rights. As noted by Besson (2005, 422), the will theory of rights does not account for the existence of rights we regard

theory identifies it as an "essential feature of the rules that attribute rights" that these rules are designed to guarantee of an interest: the aim of rights is the "protection or promotion of interests or individual goods" (Celano 2013, 58, my trans.). Raz (1986, 180), for instance, underlines that "to assert that an individual has a right is to indicate a ground for a requirement for action of a certain kind, i.e. that an aspect of his well-being is a ground for a duty on another person." As has been noted (Zanghellini 2017, 28), for Raz this interest can refer to both "an intrinsic and ultimate value" (i.e. people's well-being) and an instrumental value, that is, what may bring "beneficial consequences, either to the right-holder or others," and this includes "considerations of the general or common interest" (Campbell 2016). From this perspective, "to say that *X* is a right-holder is to say that his interests, or an aspect of them, are sufficient reason for imposing duties on others either not to interfere with *X* in the performance of some action, or to secure him in something" (ibid.). This means that "individual interests are grounds for rights, and rights are grounds for duties" (Zanghellini 2017, 26). Interest theory thus makes two claims in two steps: it shows that rights are reasons and that they are dynamic. With regard to the first claim, the existence of a right is established through a comparison of reasons, and that these reasons may therefore be different and in conflict with one another.[8] The dynamic dimension of rights, for its part, implies the possibility of realizing them progressively, meaning that a right maybe attributed even if it is not yet possible to identify all the circumstances of its application and not even the duty-holders. This implies that "the force of a right is not necessarily exhausted by any existing set of duties, etc., that follow from it, but maybe a ground for creating new duties as circumstances change" (Campbell 2016).

On this approach, as S. Besson (2005, 426) notes, "interests often conflict with one another and hence conflicts of interests lie at the foundation of rights." These conflicts, which "may arise at the level of interests, rights or duties," require extensive forms of balancing. This applies to (a) conflicts of interest in which "interests will be weighed against one another and rights will only be recognized in a limited way according to the resolution of this weighing"[9]; to (b) "*conflicts of rights stricto sensu*," which are mainly owed to the "dynamic nature of rights," when, for example, "new rights maybe derived from core rights and these rights may conflict with others;" and to (c) "*conflicts of duties*," which, in turn, depend on the "dynamic nature of rights

as inalienable and as linked to objective aspects of our well-being, and in particular it does not account for fundamental or human rights.

[8] Raz (1986, 181–182), states, "an interest is sufficient to base a right on if and only if there is a sound argument of which the conclusion is that a certain right exists and among its non-redundant premisses is a statement of some interest of the right holder, the other premises supplying grounds for attributing to it the required importance, or for holding it to be relevant to a particular person or class of persons so that they rather than others are obligated to the right holder. These premisses must be sufficient by themselves to entail that if there are no contrary considerations then the individuals concerned have the right. To these premisses one needs to add others stating or establishing that these grounds are not altogether defeated by conflicting reasons. Together they establish the existence of the right."

[9] Besson (2005, 426) underlines that "conflicts of interests are essential in determining whether one has a right in the first place, and hence whether this right can conflict with others later on."

and the possible generation of successive duties due to the diversity of the interests protected or changes of circumstances" (ibid.). As mentioned, the need to weigh reasons has been highlighted by Raz (1986, 184), arguing that in order to establish the existence of a right we have to consider the "conflicting considerations" that can "defeat" or "weaken" "the interests of the would be right holder," both in general and "on some but not on all occasions." In the example offered by Raz (ibid.), the fact that "there is a necessary conflict between free speech on the one hand and the protection of people's reputation or the need to suppress criticism of the authorities in time of a major national emergency on the other" means that "a general right is, therefore, only a prima facie ground for the existence of a particular right in circumstances to which it applies. Rights can conflict with other rights or with other duties, but if the conflicting considerations defeat the right they cannot be necessarily co-extensive in their scope." As this approach has been summarized by Zanghellini (2017, 37), "the resolution of rights disputes involves the quantitative method of balancing concrete weights."[10]

3.2 Rights as Principles: Robert Alexy's Theory

Robert Alexy's theory is undoubtedly the one that most strongly establishes a direct relation between rights and proportionality. This approach is developed on the basis of the vision of fundamental rights as principles that express values and that, "as a consequence of the principled quality of fundamental rights" (Schlink 2012a, 730), can take effect only through the balancing and use of the proportionality tool. In fact, fundamental rights as principles "require optimization of the values they express, their realization to the greatest extent possible," and they therefore "unavoidably conflict with other fundamental rights that require optimization of their own sets of values [and] with the principles that guide the state in pursuing its goals" (ibid., 730–731). The idea that fundamental rights are principles derives both from the observation of the structural complexity of contemporary legal systems (operating on the basis of several types of standards) and from the idea that the legal form of principles (as opposed of that of rules) is more appropriate for the legal formulation of rights. Structural complexity refers to the introduction of a set of values in contemporary constitutions, and the idea is that this introduction necessitates the normative form of principles. Alexy (2010a, 86ff., 92–93) argues for the structural identity between the principles and values, considering their application to be isomorphic: just like values, principles have application forms that make it necessary to look at them

[10] According to Zanghellini (2017, 37), assuming we see rights as exclusionary reasons, Raz's theory can be a useful approach to resolving rights disputes through "qualitative methods" such as the criteria of "internal relation" (as Waldron 1989 proposes) and the "harm principle" (as Möller 2012a proposes). From the opposite point of view, and proceeding from the idea of incommensurability, Verdirame (2015) refers to Raz's theory to highlight the limits of balancing and of recourse to the proportionality criterion.

in comparison (especially when the problem is to determine whether something is "good" or "best").[11]

Principles are distinguished from rules in virtue of their features under three main headings (Alexy 2010a, 44ff.)[12]: (a) their normative dimension, (b) their formulation, and (c) their collision and the process of resolving their conflict.

(a) Unlike rules, which are definitive commands (*definitive Gebote*), principles are characterized as "optimization requirements" (*Optimisierungsgebote*) (Alexy 2010a, 47–48). As an expression of (moral) values, principles are norms that prescribe that something be done to the greatest possible extent, this consistent with what is legally possible (that is, with the other principles and rules of the legal system) and with what is factually possible. From this perspective, principles can be realized to varying degrees on the basis of these possibilities, while rules can only be followed or not.

(b) Principles have an "open" formulation, meaning that their premises, and above all the conditions of their application, are sometimes not defined. This leads to principles being not only open from a semantic point of view (and principles are therefore generic and vague) but also undetermined from a deontic point of view, i.e. they do not define how they can be realized (e.g. whether through abstention or through action).[13]

(c) The conflict among principles does not affect the validity of norms (and therefore is not configured as an antinomy among standards), but has to do with the weight and importance that the various principles (or aspects of a principle) may have with respect to the specificity of the concrete case at hand (Alexy 2010a, 48ff.).[14] If a principle prohibits some behaviour and another one permits it, it will be necessary to determine which one is more relevant to the case and thus outweighs the other. Unlike conflicts between rules (solved by the criterion of formal validity and the criteria for working out antinomies), those between principles requires weighing (balancing). That one principle prevails over another does not entail that the latter is invalid, nor does it mean that an "appropriate exception" has been found: the principle that does not prevail

[11] Alexy (2010a, 162–170, 378) argues that, when it comes to application, principles and values, "are the same thing."

[12] As is well known, the distinction between rules and principles was introduced in the contemporary debate in Dworkin (1977).

[13] This aspect is clarified by Alexy (2010a, 33ff.) in relation to Art. 5 of the German Constitution (*Grundgesetz*): the different possibilities for applying this article (on freedom of science, research and teaching) depend not only on the "semantic" indeterminacy of the term *science* but also on whether this freedom is achieved by virtue of the state refraining from action (abstention) or actively intervening to protect it (action).

[14] The distinction between rules and principles is seen by Alexy (2000, 44) as "the basis for a theory of constitutional justification and a key to the solution of central problem of constitutional rights doctrine." Similarly, for Alexy, "without it can be neither an adequate theory of the limitation of rights, nor an acceptable doctrine of the conflicts of rights, nor a sufficient theory of the role of constitutional rights in the legal system."

> retains its validity and may even become prevalent in other cases. Balancing is thus the way in which principles are typically applied.

These aspects of principles, and in particular the need for balancing, are directly linked to proportionality: Alexy points out that conflicts between principles and the resulting limit on the realization of a competing principle implies the dimension of "proportionality." The conception of principles as optimization requirements leads to a conceptual relation to "proportionality": as Alexy notes (2000, 247), "the principle of proportionality [...] follows logically from the nature of the principles and is deductible from it." Viewed as optimization requirements, principles are therefore bear a "necessary" connection (Alexy 2014) to a judgment based on proportionality.

3.3 Rights and Limitations: Barak's Analysis

The relation between rights and proportionality is analysed by A. Barak on two levels: on the one hand is the distinction between rights "grounded in the constitution" and the limitations that can be posed "by a sub-constitutional norm (such as an "ordinary" statute or common law rule)"; on the other hand, is the distinction between "the scope of a constitutional right and the limitations to which it is subject" (Barak 2012a, 739). This dual distinction is intended to demonstrate the need to resort to the canon of proportionality: the first distinction—between laws on different hierarchical levels (constitutional and sub-constitutional)—establishes this need in working out the relationship between "democracy, separation of powers and constitutional rights" (ibid.), while in the second, the same need emerges from the exigency not to compromise the scope of rights and their realization. As Webber (2016) has noted, at the basis of this reflection we find the idea that rights "cannot be realized to [their] fullest extent" and that a "limitation on a constitutional right by law [...] will be constitutionally permissible if, and only if, it is proportional" (Barak 2012b, 27, 3).

For Barak, proportionality must be seen "as the standard for determining the constitutionality of a sub-constitutional norm that limits a constitutional right" (Barak 2012a, 739). In this sense, proportionality is the tool for assessing whether a sub-constitutional rule unduly covers the scope of a right, namely, the "area that it covers—its content and its boundaries—[and that] can be changed only by constitutional amendment" (ibid.). Proportionality therefore makes it possible to assess whether a "sub-constitutional (statutory or common law) norm" does not go beyond "the limitations on a constitutional right" that can be explicitly or implicitly imposed by a limitation clause establishing "the constitutional conditions under which the right maybe less than fully realized" (ibid.).[15] It is proportionality that makes it possible to determine whether these conditions are met. The distinction between scope and limitations "establishes two stages of constitutional analysis. At the first stage,

[15] Barak (2012a, 739) notes that "in some legal systems, relative rights have a core that cannot be limited; that core is absolute. That a constitutional right is relative does not mean, however, that it is a prima facie right. A relative right is still a definite right."

the inquiry pertains to whether a constitutional right is limited by a sub-constitutional norm […]. At the second stage, the inquiry considers whether the limitation on the constitutional right is proportional" (ibid., 740).[16]

This twofold assessment is articulated in the four elements of proportionality: "proper purpose," a "rational connection" between means and purpose, "necessity," and "balancing" (ibid., 742ff.). The function served by evaluating these aspects is to "ensure that a sub-constitutional norm limiting a constitutional right fulfils its four elements. If those elements are not fulfilled, the sub-constitutional norm will lack the force to limit the constitutional right, for a higher norm trumps a lower norm" (ibid., 741).

The proportionality test "is a framework that must be filled with content." For Barak, this content "will be determined by a set of considerations that are external to proportionality and that inform it. That content therefore may vary from one legal system to another" (ibid.). However, there is a definite point that highlights the decisive role of the proportionality test: this test "is not neutral with respect to human rights, and it is not indifferent to their limitation. It is grounded in the need to realize human rights" (ibid.). This applies to the higher hierarchical level of constitutional rights and their limitations. For Barack, "democracy is based on human rights, and the restriction of those rights cannot become routine." For this reason, "the limitations that proportionality imposes on the realization of constitutional rights" require "continuing justification, grounded in public reason" (ibid., 749). This also applies to the scope of rights: "the elements of proportionality reflect the idea that a sub-constitutional norm may impose limits on a constitutional right, but that those limits are themselves bounded. This is the concept of 'limits on the limitations'" (ibid., 741).

3.4 Kai Möller: The Global Model of Constitutional Rights

The global model is "a morally reconstructive theory of rights" (Möller 2014, 5) that offers an account of how rights have evolved since the second half of the twentieth century. It is not intended as philosophical or theoretical account of rights, but as a reconstructive one, since its purpose is primarily to be compatible with this evolution (which is claimed to point to the "existence" of global model): "it is a theory of the actual practice of constitutional rights law around the world" (Möller 2012a, 20). For Möller, this theory "must meet two criteria: first, it must 'fit' the global model sufficiently well to be rightly considered a theory 'of' that practice, and second, it must be morally coherent" (Möller 2014, 2). This second aspect (the

[16] Barak (2012a, 740), stresses that at the first stage "the burden of proof is on the party asserting the limitation," while at the second "the burden of proof […] is on the party asserting proportionality."

moral reconstruction) "aims at finding moral value in a practice: something which makes it worth continuing with that practice" (Möller 2012a, 21).[17]

From the first point of view, it is necessary, for Möller, to overcome the traditional theories of rights that fail to take current legal practice into account. What he calls the "dominant narrative of the philosophy of fundamental rights" supports a number of arguments that are largely overcome. These are in particular the theses for which (a) "rights cover only a *limited domain* by protecting only certain *especially important* interests of individuals"; (b) rights have only a "vertical" domain (they "operate only *between a citizen and his government*") and "impose exclusively or primarily *negative* obligations on the state"; and (c) "rights enjoy a *special normative force*, which means that they can be outweighed, if at all, only under *exceptional circumstances*" (ibid., 2). Reality and legal practice are instead totally different and are marked by "rights inflation, positive obligations and socio-economic rights, horizontal effect, and balancing and proportionality" (ibid.). In this sense, it can be said that the current legal practice—which "sees rights as protecting an extremely broad range of interests but at the same time limitable by recourse to a balancing or proportionality approach"—directly contradicts "the conceptions of rights proposed by most if not all moral and political philosophers who agree that rights protect only a limited set of especially important interests while enjoying a special, heightened, normative force" (ibid., 1). Particularly important is the fact of "rights inflation," attesting to the plurality of interests and needs that, in the contemporary context, are defended as rights.

From the second point of view, this approach aims to be "general in that it does not focus on specific issues or rights but aims at identifying features of their moral structure which are shared by many or all constitutional rights" (ibid., 2).[18] It wants to answer two main questions. The first is: What are the values protected by rights? The second is: What limits are the rights based on these values subject to.[19]

The value that rights refer to is that of autonomy. Autonomy is to be understood in a broad sense inclusive of all the activities that are "valuable *from the perspective of the agent*," which means that something is valuable if "the agent has an autonomy interest in the activity that must be protected by a right" (Möller 2013, 10). It is not possible to define such interests through a "threshold model" that selects these

[17] Möller (2012a, 21) notes that "it is of course possible that there is no such moral value, in which case this would have to be acknowledged by adopting the perspective of what Dworkin calls the internal sceptic; the consequence is that the practice ought to be discontinued."

[18] Möller (2012a, 1–2) stresses that his "theory follows a *substantive moral approach* in that it is grounded in political morality," and therefore that his approach "can be contrasted with a formal theory such as Robert Alexy's influential theory of rights as principles or optimization requirements."

[19] Möller (2012a, 1) points out other question that his theory seeks to answer: "(1) Which *theory* or *conception of rights* explains best the global model of constitutional rights, including the questions of which values are protected by rights and what are their limits? (2) How does the judicial enforcement of *this particular conception* of constitutional rights relate to the value of *democracy*? (3) How does it relate to the value of the *separation of powers*, in particular to considerations of the *relative institutional competence* of courts on the one hand and the elected branches on the other?"

interests in a "qualitative" manner, because this would lead to arbitrary choices (ibid., 17).

This does not mean that "anything goes," but simply that it is necessary to distinguish "between prima facie rights and definite rights." Whereas a "definite right to engage in a particular activity [...] grounds a duty of non-interference on the side of the state," this is not the case for "the prima facie right [that] grounds a different duty: the duty of the state to take the respective autonomy interest adequately into account" (ibid., 13). Recognition of a wide range of interests as rights highlights the link between law and proportionality: the possibility of interfering or limiting these rights can only take place if there are "sufficient, and proportionate, reasons to do so" (Huscroft et al.2014, 9).

4 Proportionality and Rights: The Critical Theses

Despite its spread, proportionality has often been seen as something that is not always congruent with rights adjudication. By making it possible to limit rights, recourse to proportionality would amount to a "weakening" and an "emptying" of rights (Pino 2014a). As has been noted, one of the strongest criticisms is that "an understanding of rights that makes the existence of a definitive right dependent on applying a proportionality test undermines the very idea of rights" (Kumm and Walen 2013, 1). Proportionality test would not represent a true protection of rights as "rights are awarded no special priority," and "are reduced to defeasible premises in reasoning about proportionality" (Webber 2013, 4, 9), that is "to defeasible interests, values, or principles" (ibid., 9). The proportionality judgment is therefore seen as tantamount to a "dilution" or a "relativization" of rights: "a right or freedom is protected only to the extent that a state does not have a legitimate interest that requires its intrusion or limitation" (Schlink 2012a, 732).[20] At best, proportionality can be seen as one of the tools that, contingently, can be used in rights adjudication.

Criticism of proportionality/balancing has been raised at different levels of generality: the most important refers to the teleological/consequential dimension of the proportionality judgment, while other criticisms concern more specific aspects of rights. Without any claim to be exhaustive, we will consider (a) the limits of proportionality as a teleological judgment and (b) the problem of specific types of rights (such as "positive" and "horizontal" rights) and the possibility of different judgments on rights.

[20] Schlink (2012a, 732) underlines that "proportionality analysis is a reasoning process in which, prima facie, everything can be argued for or against the suitability or necessity of a means and the balance of the means and the end."

4.1 Proportionality as a Teleological Approach: Rights as Fungible Goods

One of the main arguments against the use of balancing and proportionality as decision-making tools in judgments on rights relates to their teleological dimension. One of the first criticisms made in this respect is that of Jürgen Habermas (1996) who, referring to Alexy's doctrine of rights as principles,[21] and relying on part of the German constitutionalist theory, argues that in a balancing/proportionality judgment, rights would be considered as fungible goods subject to a "cost-benefit analysis" in which "functionalist arguments then gain the upper hand over normative ones" (ibid., 259). Considering rights as the object of balancing "converts such rights from deontological legal principles into teleological legal interests or goods" (ibid., 258). This means that "in cases of collision *all* reasons can assume the character of policy arguments." If this happens, "then the firewall erected in legal discourse by a deontological understanding of legal norms and principles collapses" (ibid., 258–259). This means that "as soon as rights are transformed into goods and values in any individual case, each must compete with the others at the same level for priority," and this is reflected in the fact that "every value is inherently just as particular as every other, whereas norms owe their validity to a universalization test" (ibid., 259). This amounts to reducing rights to a competition between values or interests: in balancing, "values can only be relativized by other values; this process of preferring or pursuing values, however, resists attempts at logical conceptualization" (ibid., quoting Denninger 1990, 147). Rights, for Habermas, require a "deontological" consideration, that is, they need to be treated as "norms." In this way, they base their "validity" on a "universalization test" and are therefore "universally binding" (Habermas 1996, 259). Only this perspective makes possible their "unambiguous specification," and makes it unnecessary "to decide to what extent the competing values are respectively optimized" (ibid., 260). On Habermas's analysis, balancing is therefore seen as "a consequentialist form of reasoning that does not fit the deontological nature of at least some rights" (Kumm and Walen, 2013, 4).

This criticism offered by Habermas, summarizing a number of critical points of the balancing method,[22] carries three significant corollaries.

In the first place, the relativization of rights translates into the arbitrariness of decision-making, and ultimately into the irrationality of balancing. As we have seen, balancing should be regarded as irrational, as "there are no rational standards" on which to base it, and "weighing takes place either arbitrarily or unreflectively,

[21] Reference is being made here to Alexy's identifying values and principles in his theory of rights. Habermas rejects this identification, arguing that, while legal principles are deontological, values are teleological.

[22] Following Habermas's line of reasoning, Tsakyrakis (2009, 487) argues that the "balancing approach, in the form of the principle of proportionality, appears to pervert rather than elucidate human rights adjudication. With the balancing approach, we no longer ask what is right or wrong in a human rights case but, instead, try to investigate whether something is appropriate, adequate, intensive, or far-reaching." The teleological/consequentialist dimension of balancing is often associated with a utilitarian perspective.

according to customary standards and hierarchies": the practice of balancing increases "the danger of irrational rulings" (ibid., 259).

In the second place, further specifying the problem of irrationality, the balancing/proportionality method "requires the weighing of incommensurables lacking a common metric" (Jackson 2015, 3156): this is "to compare the length of lines to the weight of stones" (Petersen 2013a),[23] or "apples and oranges" (Borowski 2013). The problem lies not only in the impossibility of comparing cardinal numerical values but also in that of an ordinal comparison: all this would result in "an unacceptable level of indeterminacy" (Jackson 2015, 3153). As has been noted (Petersen 2013b, 3), if we treat balancing as a "cost-benefit-analysis," then we need to be prepared to accept that "the limitation of an individual right passes the constitutional muster if the marginal benefit of the state measure for a public purpose outweighs the marginal restriction of the constitutional right." The problem is that this comparison "usually requires that the compared goods can be measured in one common normative currency, i.e. they are commensurable." This commensurability appears "often lacking when it comes to the resolution of conflicts of competing constitutional rights and values" (ibid.).[24]

In the third place, balancing/proportionality seems to contradict the liberal rights tradition, that is, the idea that rights have a core immune from the state's decisions. If we see "the practice of proportionality justification" in a teleological sense, it is evident that it appears "as antithetical to the very idea of constitutional rights" (Thornburn 2016, 309–310). In this sense, one can argue that "rights subject to limitation under a general proportionality test should be viewed with suspicion across the wide spectrum of liberal theories of fundamental rights that reject consequentialism" (Verdirame 2015, 349). The liberal position requires that there be a sharp distinction "between the logic of ordinary public policymaking and the logic of rights" (Thornburn 2016, 311): "rights must not be amenable to the balancing of interests. Instead, they must act as firm and impenetrable constraints [...] on the ordinary logic of state action" (ibid.).[25] As has been noted (Klatt and Meister 2012, 16), in part of the literature on rights, balancing seems to contradict the "basic liberal intuition that rights enjoy some kind of special priority, which gives them lexical priority over other considerations, in particular over any public interest."

[23] Jackson (2004, 3156–3157) attributes this phrase to Justice Scalia in *Bendix Autolite Corp. v. Midwesco Enters., Inc.*, 486 U.S. 888, 897 (1988)

[24] Endicott (2014, 9ff., 20ff.) distinguishes between "radical" and "vague" incommensurability and points out (starting, however, from the premise for which "the incommensurabilities in human rights cases [...] do not necessarily lead to arbitrary decision-making" and that "proportionality reasoning is not generally pathological in human rights cases") six potential pathologies in the judicial appraisal of incommensurabilities. Möller (2012b, 719ff.), analysing "whether [...] incommensurability poses a threat to the principle of proportionality," distinguishes between "strong and weak incommensurability."

[25] Thornburn (2016, 310) recalls the conceptions expounded by Nozick (side constraints), Dworkin (trumps), Schauer (shields), and Habermas (firewalls), among others.

4.2 Proportionality and "Positive" and "Horizontal" Rights

A less general criticism highlights how proportionality cannot be used for an important class of rights. As has been pointed out by Gardbaum (2016, 1), the fact that "the actual practice of rights adjudication around the world reveals that there are limits to the use of proportionality" seems to be linked to "two newer types of rights": "positive" rights (which include "social and economic rights") and "horizontal" ones. These two types of rights, which would be one of the examples testifying to the spread of the "global model of constitutional rights" (Möller 2012a), require a different jurisprudential treatment and an analysis not linked to the idea of proportionality. This thesis is argued on a twofold level: on the one hand, Gardbaum points out that in the "practice of Courts" having jurisdiction in cases involving constitutional rights (and this applies in particular to the European Court of Human Rights, the German *Bundesverfassungsgericht*, and the South African Constitutional Court)[26] proportionality is not used as a basis of adjudication; on the other hand, he highlights the need to more accurately distinguish between rights-conflicts and the ways in which they can be solved them. On this reconstruction, there are two main types of conflict: one that assumes a "relationship" between conflicting rights and one that instead does not assume this relationship. In the first case, what is to be considered is the relationship between the advantages that can derive from limiting a right and the disadvantages this involves (for another right),[27] while in the latter, we have to choose between "values" that do not stand in any relation to one another.[28] In the case of a conflict between values and of "positive" and "horizontal" rights, the logic of proportionality does not appear appropriate: it is linked to a "means-end relationship," which therefore assesses the proportionality of a measure limiting a right, and which does not apply in the absence of such relationship. As Gardbaum (2016, 7) points out, therefore, "not all balancing is the same," and thus "not all balancing involves proportionality; i.e. it can be, and is, done without."

[26] Referring to the ECtHR, Gardbaum (2016, 13), notes that "what we find when we look at the ECHR's leading positive rights cases is essentially no use of proportionality. Rather, the court focuses almost exclusively on the first stage issues of determining the content and scope of the right, and whether it has been infringed."

[27] Gardbaum (2016, 9) stresses that "proportionality is a relational concept."

[28] Gardbaum (2016, 7–9) exemplifies the two types in this way. On the one hand (a) we have conflicts that require us to evaluate a disproportionate act like that proposed by Schlink (2012b, 293), in which "a crippled homeowner sitting on his porch shoots a child stealing apples from his tree after the only other ways of protecting his property open to him—calling to the child to desist—fail." In this hypothetical case, "balancing the value of the child's life against that of the saved apples on any relevant metric results in the conclusion that the means used were massively disproportionate and hence the action clearly unjustified" (Gardbaum 2016, 7). On the other hand (b), we have conflicts that, by contrast, do *not* require us to make a proportionality evaluation, a case in point being a "person who has promised to meet a friend for coffee but shortly before the appointed time learns that his wife has just been admitted to an emergency room." In this case, "balancing is not used to evaluate the proportionality of the means employed as the two duties are independent values: there is no means-end relationship between them to assess the proportionality of."

The basic thesis is therefore that in the context of a general judicial task of assessing the reasonableness of legislative measures,[29] conflicts between values not linked by a means-end relationship require another type of assessment: something close to the test of the "reasonableness review" the South Africa court employs for "socio-economic rights." This is an assessment that does not involve analysing the proportionality of a measure, but "in positive rights cases at least" seeks to see, for example, if "an omission *is* a violation" of a social right (ibid., 35).[30]

PART II

5 Proportionality and Rights Adjudication

Proportionality review was introduced in Germany,[31] and spread across the rest of Europe and beyond, becoming a distinctive feature of the "post-war paradigm" of rights adjudication (Thorburn 2016, 305): a dominant justificatory practice deployed by national, supra-national and international courts in reviewing the legitimacy of acts that interfere with rights.[32]

The most prominent version of this practice structures the judicial analysis so as to "focus on the same questions in the same order" (Jackson 2015, 3094), determining whether the acts under review are justified according to criteria of (1) suitability, (2) necessity and (3) proportionality stricto sensu. More precisely, this version of proportionality review—call it the *standard version*—devises a three-step analytical sequence[33] aimed at assessing (1) the suitability of the institutional act under review relative to its (legitimate) purpose (suitability test); (2) the necessity of the act in relation to the accomplishment of that purpose (necessity test); and, finally, (3) the

[29] For Gardbaum (2016, 5) the task "of judicial review in a democracy" is that "of policing the boundaries of the reasonable."

[30] This position is, however, open to challenge: Klatt and Meister (2012, Chap. 5) argues for the full applicability of proportionality to "positive" rights, while Young (2012, 219) stresses that "the constitutional doctrine of reasonableness [in South Africa] [...] uses a form of proportionality reasoning."

[31] The first decision applying the proportionality principle is reported (for instance, see Grimm 2016, 172; Alexy 2010a) to be BVerfGE 3, 383 (1954); followed by BVerfGE 7, 377 (1958); BVerfGE 13, 97 (1961); BVerfGE 16, 194 (1963); BVerfGE 19, 342 (1965).

[32] Proportionality standards are applied in rights adjudication from Germany to Canada, from Israel to Colombia and South Africa; furthermore, the use of proportionality review has been spreading at a transnational and international level, being adopted, among others, by the European Court of Human Rights and the European Court of Justice. On the wide circulation of proportionality standards in rights adjudication, see, among many, Stone Sweet and Mathews (2008, 80) defining proportionality as a "global constitutional standard"; Barak (2012b, 175–210), Jackson (2015), Beatty (2004, 162), defining proportionality as "a universal criterion of constitutionality"; Cohen-Eliya and Porat (2013, Chap. 1).

[33] In some versions (see, for instance, Barak 2012a; Grimm 2007) there is a further, initial, analytical step, specifically devoted to verifying whether the aim pursued by the act under review is legitimate. In the standard version, however, the inquiry into the legitimacy of the aim pursued by the act comes with the suitability and necessity tests.

proportionality of the act, that is, whether the relevance of the interests protected by the act and the extent to which they are satisfied can justify sacrificing the rights at stake.[34] Under the third test, proportionality analysis enters the sphere of judicial balancing: it requires to balance the interests realized by the act under review against the rights to be sacrificed; to this end, the reasons for interfering with a right should be weighed against the reasons that underlie its protection, so as to satisfy the former without disproportionately sacrificing the latter.

This version of proportionality—the standard version—has so far had the greatest visibility and impact on the theory and practice of rights adjudication. There are, however, further versions of proportionality, which for the most part differ by the order and/or the characterization of the proportionality sub-tests—especially the balancing test—and/or the emphasis placed on one or the other of these tests.[35]

In fact, proportionality is not *a unitary* justificatory practice but a *family* of such practices that deploy a range of different argumentative paths to rights adjudication. These practices share an approach to rights adjudication—call it *proportionalism*—characterized by recourse to judicial tests by which to assess whether interferences with rights are an adequate instrument for realizing legitimate ends (that is, they do not exceed what is required to realize those ends). However, proportionalist practices diverge in designing these tests, especially when it comes to weighing the interests that underlie the act under review against the interests protected by the rights at stake.[36]

The following sections are aimed at sketching the different forms of proportionalism, taking into account a distinction between two dimensions of rights adjudication. On the one hand is an "internal" dimension, relating to the judicial assessment of acts that interfere with rights; on the other hand is an "external" dimension, relating to the grounds for, and the limits of, adjudication.

In these terms, a distinction maybe drawn between proportionalism *in* review and proportionalism *of* review.

In the first form, proportionalism is a way of structuring judicial analysis in order to test whether the acts under review are legitimate in their interfering with rights. More specifically, proportionalist models of review require that we verify that these acts pursue legitimate objectives and are proportionate means for the realization of those objectives. On some models, then, the judicial analysis should also test

[34] This is the *proportionality as such* (Jackson 2015, 3094) or *proportionality* stricto sensu stage. According to Alexy (2003, 436–437) the proportionality stricto sensu principle establishes that: "The greater the degree of non-satisfaction of, or detriment to, one right or principle, the greater must be the importance of satisfying the other." Under this sub-principle the "Law of Balancing" requires three steps by which to establish (a) the degree to which a right or principle has fallen short of being satisfied or has been sacrificed; (b) the import that can be associated with satisfying a right or principle in conflict with the one that has been sacrificed; and (c) the importance of satisfying the second principle as justification for sacrificing or failing to satisfy the first.

[35] For an overview and analysis of different versions of the proportionality review template see, among many others, Jackson (2015), Stone Sweet and Mathews (2011), Grimm (2007).

[36] See, among many, the proposals of Luterán (2014) and von Bernstoff (2014).

whether the acts under review carry a sacrifice for rights that is commensurate with their relevance, given the relevance of conflicting interests.

In the second form, proportionalism is a way of handling the institutional dimension of rights adjudication. In particular, proportionalist approaches to rights adjudication aim at determining the scope and intensity of judicial scrutiny so as to combine the protection of rights with the respect owed to political decisions, taking into account the weight they carry from time to time.

Based on this distinction, I will first discuss proportionalism *in* review as a way of getting the "measure" of the protection the different interests at stake deserve, and then I will turn to the forms taken by proportionalism *of* review as a way of getting the "measure" of judicial action.

6 Proportionalism *in* Review

Proportionalism *in* review can be defined as a set of approaches to rights adjudication that are aimed at testing the suitability and necessity of interference with rights in light of the relevance of the various rights/interests involved. Proportionalist approaches, then, mostly diverge with regard to the assessment of such relevance.

More specifically, the different versions of proportionality *in* review primarily differ in two respects: (1) the connection between proportionality analysis and balancing, and (2) the understanding of balancing as a form of practical reasoning about legal questions.

(1) On the one hand, there are versions of proportionality analysis that require a balancing test aimed at weighing the various interests and rights at stake; on the other hand, there are versions that only require a means-ends analysis guided by criteria of suitability and necessity, but that do not entail a specific balancing step in order to weigh the different rights/interests against one another.

(2) In the background, the different versions of proportionality analysis rely on different accounts of balancing as a form of legal reasoning. Some versions qualify it in strong terms, that is, as a specific, distinctive form of legal reasoning presenting features that distinguish it from other forms of legal reasoning, notably subsumption (Alexy 2003). Other versions qualify balancing in weak terms, as a merely prudential attitude in legal reasoning, one that makes judicial review sensitive to the different interests and factors that are relevant to a decision but does not call for any specific test (Luterán 2014).

Of course, the two aspects just mentioned are connected, since the relation between proportionality and balancing also depends on how the latter is conceived: when conceived in weak terms, its ties to proportionality analysis only call for a prudential attitude in review, but do not require the review to follow a specific route by performing specific tests. By contrast, when balancing is conceived in strong terms, as a *specific* kind of legal reasoning, its ties to proportionality analysis become more "stringent"—requiring that a sequence of specific analytical steps be followed—and hence more controversial (ibid., 24–26).[37]

In the first case, balancing qualifies, not as a separate stage in the decision-making process, but rather as the judicial attitude of taking into account the different interests at stake, an attitude that different review paradigms can express in different terms. Indeed, this judicial attitude may result in a cost-benefit analysis—disconnected from more complex analytical frameworks—as well as in a standard proportionality analysis, framing the balancing test as one of the three, separate, analytical steps. When balancing is conceived in weak terms, it is quite uncontroversial that it somehow characterizes judicial reasoning and does not point to the adoption of any specific argumentative template. From this perspective, the balancing attitude should characterize any and all of the tests adopted so as to assess the legitimacy of the act under review.

In the second case, by contrast, balancing is conceived as a distinctive form of legal reasoning, pointing to the establishment of a normative equilibrium among different constitutional protections under a criterion of proportionality stricto sensu. In these terms, judicial balancing is a *specific* way of solving normative conflicts that involve rights, and it marks the conclusive step in a structured analysis, placing specific argumentative constraints on the judicial decision-making process. In this case, the role of balancing in rights adjudication is more controversial: first, it is controversial that this form of legal reasoning is a necessary component of proportionality review; second, it is disputed that judicial recourse to a balancing test is legitimate and acceptable.

Concerning the first aspect, we will analyse the different role of balancing in the two main versions of proportionality review; concerning the second aspect, we will point out the problems of legal certainty raised by the judicial application of a balancing test within the proportionality review template.

6.1 Proportionality Balancing and Means-Ends Proportionality

In both the theory and the practice of rights adjudication, proportionality review tends to be identified with its standard version—the version that, as mentioned, envisages a three-step sequence of judicial tests guided by the criteria of suitability, necessity and proportionality stricto sensu.

[37] As pointed out by Luterán (2014), a conception of balancing in strong terms underlies the accounts of proportionality of Alexy (2003, 2010a), Barak (2010), Möller (2013), whereas a "weak" conception seems to underlie the analysis of Gunn (2005).

Other versions, however, characterize proportionality review in different terms. In fact, at least two fundamental versions of proportionality *in* review can be identified: *proportionality balancing* and *means-ends proportionality*.[38]

Proportionality balancing is the standard version, the prominent one in both the theory and the practice of rights adjudication. It requires, first, a means-ends analysis of the act under review, that is, a scrutiny of the suitability and necessity of the act as a means for realizing a certain end. Second, it requires a test aimed at balancing the weight of the interest(s) underlying the act under review against the weight of the right(s) at stake. Indeed, the criterion of proportionality stricto sensu governing this latter step calls for the establishment of an equilibrium among the different interests/rights that come into conflict in a given case.

The most influential model of *proportionality balancing* is the one put forward by Robert Alexy. This model, in a nutshell, constructs constitutional rights as "double aspect constitutional norms" combining the level of rules with the level of principles (Alexy 2010a, 84–86). From this perspective, as we have seen, constitutional rights are "optimization requirements," that is, principled norms "requiring that something be realized to the greatest extent possible, given the factual and legal possibilities at hand" (Alexy 2010b, 21). Rights are thus construed as prima facie principles (Alexy 2010a, 57)[39] that may conflict with other rights or interests interfering with their full realization. From this feature of rights there logically follows the principle of proportionality: it "is implied by it and vice versa" (ibid., 66).

Indeed, the proportionality principle expresses the idea of optimization and requires us to assess what conflicting principles demand in concrete cases, this by "testing" their optimization under the criteria of suitability, necessity and proportionality stricto sensu. According to the criteria of suitability and necessity, the optimization of principles is tested by what is *factually* possible. According to the criterion of proportionality in the narrow sense, the optimization is tested by what is *legally* possible, that is, by competing principles.[40] The latter test, performed according to the criterion of proportionality stricto sensu, is decisive: it is aimed at balancing the non-satisfaction of a principle with the importance of satisfying the other principle with which it is in conflict. As Alexy put it in the *Law of Balancing*: "The greater the degree of non-satisfaction of, or detriment to, one right or principle, the greater must be the importance of satisfying the other" (ibid., 102). The result is a "conditional relation of precedence" among the principles at stake.

The standard version of proportionality review is challenged by models that qualify proportionality review as a two-step *means-ends* analysis aimed at testing the

[38] This distinction draws on the analysis of Luterán (2014, 22 ff.). Proportionality in review is associated with balancing by many theories of rights adjudication (see, for instance, Alexy 2003, 2010a; Barak 2010; Webber 2010, 2013). There are, however, accounts of proportionality review that do not associate it with balancing (see Luterán 2014 and also—in different terms—Rivers 2006).

[39] On the prima facie character of principles Alexy makes reference to Ross (2002) and also K. Baier (1965) and Hare (1985).

[40] The idea that a neat distinction can be drawn between questions of fact and questions of law in this kind of review has been challenged: see Pino (2014b, 606).

acceptability of the means adopted to achieve the aims pursued by the institutional act under review. This scrutiny comprises a suitability test and a necessity test but does not include a separate balancing test proper.

Some authors defend this version of proportionality review as the authentic form of proportionality—the one closest to its "lost meaning" (Luterán 2014). The origins of proportionality, from this perspective, go back to the doctrine of "double effect," specifying the conditions under which a human behaviour is morally permissible by looking at both the positive and negative effects associated with that behaviour (ibid.).

The doctrine, in other terms, explains the permissibility of an action causing serious harm as a "side effect" (or "double effect") of bringing about a good result "even though it would not be permissible to cause such a harm as a means by which to bring about the same good end" (McIntyre 2014).

More specifically, according to the doctrine of double effect an act is morally permissible if it meets four conditions as follows: (a) the act itself must be morally good or at least indifferent; (b) the agent may not positively will the bad effect but may permit it. If he could attain the good effect without the bad effect he should do so. The bad effect is sometimes said to be indirectly voluntary; (c) the good effect must flow from the action at least as immediately (in the order of causality, though not necessarily in the order of time) as the bad effect [...]; (d) the good effect must be sufficiently desirable to compensate for the allowing of the bad effect (Connell 1967, 1021).

The standard of proportionality would be at work with regard to the latter condition, guiding the appraisal of the act's bad effects under the requirement that such effects "must not be disproportionate in the given circumstances" (Luterán 2014, 20). In legal reasoning, proportionality would perform the same function, since rights-adjudication deals with problems presenting the same structure as the problem of the double effect. The institutional actions under judicial review, on the one hand, claim to pursue legitimate ends and, on the other hand, give rise to negative effects consisting of interferences with rights. This double-effect problem "provides the intellectual bridge between the idea of proportionality in ethics and law" (ibid., 31). From this perspective, proportionality *in* review grounds a test that requires judicial reasoning to "identify as precisely as possible the intentional state action and the negative effect complained of," differentiating the kind of analytical route to be followed on the basis of the different kinds of conflict at stake (ibid., 41).

Beyond the theoretical proposals, a means-ends version of proportionality review to some extent characterized the case law of the European Court of Human Rights[41] until the 1980s, as well as the case law of the European Court of Justice.[42]

[41] This along the lines defined by the European Court of Human Rights in the *Case "Relating to Certain Aspects of the Laws on the Use of Languages in Education in Belgium" v. Belgium*, 1968, interpreting the proportionality requirement as concerning a "relationship of proportionality between the means employed and the aim sought," and, more precisely, calling for a "reasonable relationship of proportionality between the means employed and the aim sought to be realized."

[42] The Court made use of a proportionality review template for the first time in Case 11/70 *Internationale Handelsgesellschaft v Einfuhr- und Vorratsstelle* [1070] ECR 1125. In broader terms, on

Furthermore, the means-ends version is in certain respects close to the tier-scrutiny template adopted by US courts (with all the difficulties of comparing judicial frameworks adopted in different jurisdictions).[43] Indeed, the dominant approach to rights adjudication in Europe is characterized by the use of proportionality analysis, mostly in the proportionality-balancing version; in the United States, by contrast, courts do not resort to proportionality review but rather apply a *tier scrutiny* of the means-ends rationality characterizing the act under review. Indeed, in the United States, "the closest thing […] to a common rubric for reviewing claims across different substantive areas—is the set of standards that make up the 'tiers of scrutiny'" (Stone Sweet and Mathews 2011, 104). This set of standards defines a review template requiring the judicial scrutiny of (i) governmental interests, (ii) the effectiveness of the means chosen to realize these interests and (iii) the alternatives to these means (so as to determine whether less-restrictive means are available for furthering the governmental interests in question). This scrutiny, then, can be exercised at different levels of intensity: the "three tiers" of strict scrutiny, intermediate scrutiny and rational basis review.[44]

Two features of this review template need to be pointed out. First, it is not connected to the idea of proportionality, at least not explicitly. In fact, the judicial analysis deployed in tier scrutiny resembles the means-ends version of proportionality but is not explicitly tied to it. And, second, tier scrutiny includes a suitability and a necessity test but does not incorporate a specific balancing test. Balancing, indeed, stands as a judicial mechanism of its own and is not embedded in a specific structured template. On this basis, some Authors (Cohen-Eliya and Porat 2010, 2013; Stone Sweet and Mathews 2011) contrast the European approach to judicial balancing with the US approach: the first integrates balancing into a structured proportionality analysis, whereas the second frames it as a less structured, policy-oriented judicial test (Cohen-Eliya and Porat 2013, 60). From this perspective, the first approach reflects a constitutional culture—the European culture—characterized by "epistemological optimism," (Cohen-Eliya and Porat 2013, 59) thus framing judicial balancing as an interpretive mechanism bound to a "pyramidal, objective system of values" (Bomhoff 2008, 124). The second approach, by contrast, is presented as reflecting a different constitutional culture—such as the US culture—that is more pragmatic and characterized by "epistemological scepticism," and which disconnects judicial balancing from structured argumentative templates, such as the proportionality framework (Cohen-Eliya and Porat 2013).

the use of review templates guided by standards of proportionality in EU law, see Emiliou (1996), Ellis (1999), Arai-Takahashi (2002), Harbo (2010).

[43] The proportionality review template adopted by the Canadian Supreme Court is also closer to the means-ends template, rather than the proportionality-balancing template adopted by the German Constitutional Court; it contemplates a balancing test, but the analytical emphasis is on the suitability and necessity tests. According to Grimm (2007, 393), indeed, "The most striking difference between the two jurisdictions is the high relevance of the third step of the proportionality test in Germany and its more residual function in Canada. Here the German Court argues at length, whereas the Canadian Court mostly presents a 'résumé' of previous analysis."

[44] See Chemerinsky (2011, 687 ff).

The divergence between these approaches, then, ultimately draws on different constitutional traditions, representing in very different ways the role and limits of constitutional justice and, in the background, the relation between democracy, constitutionalism and rights. Proportionality analysis is grounded in a "culture of justification"[45] that characterizes European constitutional systems in contrast to a "culture of authority" that serves as the basis for non-structured balancing approaches (Cohen-Eliya and Porat 2013).

Indeed, in a culture of authority the legitimacy of governmental action is based on the fact that "the actor is authorized to act," whereas in a culture of justification this authorization is not sufficient and the essential requirement for the legitimacy of governmental action is that it be justified in substantive terms (ibid., 112).

From this perspective, the use of proportionality review in rights adjudication reflects a culture of justification and, moreover, is a *constitutive* condition for expanding and enhancing this culture. Indeed, proportionalism in review is conceived as an essential component of a constitutional culture of justification, one that is based on a fundamental (meta)-right of everyone to justification (Kumm 2010; Forst 2012; Cohen-Eliya and Porat 2013). The idea is that proportionality review institutionalizes the right to justification—a right that is linked to a conception of legitimate legal authority on which the "law's claim to legitimate authority is plausible only if the law is demonstratively justifiable to those burdened by it in terms that free and equals can accept" (Kumm 2010, 143). In other terms, a culture of justification requires that governmental actions are substantively justified "in terms of the rationality and reasonableness of every action and the trade-offs that every action necessarily involves, i.e. in terms of proportionality" (Cohen-Eliya and Porat 2011, 463).

This idea of a "constitutive" link between proportionality review and a culture of justification is not just relevant in descriptive terms, as a way to explain how the widespread of the former has contributed to the establishment of the latter. Indeed, this link is also normative in its import, serving as a basis for a non-instrumental justification of proportionalism in judicial review (Cohen-Eliya and Porat 2013): proportionality review is the institutional embodiment of—or *constitutes*—the right to justification and, in this sense, it is inherently valuable.

In these terms, we can go beyond a justification of proportionality review as an instrument for promoting "flexibility, political stability, efficiency, judicial legitimacy, or simply judicial power" (Cohen-Eliya and Porat 2013, 111). Indeed, the value of proportionality depends not only on the merits of the decisions it makes it possible to arrive at but, more fundamentally, on the role it plays in *constituting* a condition for the fulfilling the right to justification (ibid.).[46]

Although schematic, the juxtaposition just sketched out allows us to point out the traditional disconnect between the European proportionality-balancing approach and the US balancing approach, which is mostly due to a strong resistance to explicit recourse to proportionality analysis in the United States.

[45] Drawing on the idea of a "culture of justification" advanced by Mureinik (1994).

[46] More broadly, for a non-instrumental justification of judicial review based on the idea of a right to a fair hearing, see Harel and Kahana (2010).

In spite of this aversion, however, the evolution of the US case law, at least in some contexts, reveals a latent use of judicial frameworks resembling proportionality review (Cohen-Eliya and Porat 2009; Jackson 2015; Yowell 2014). As Justice Breyer noted in *District of Columbia v. Heller*, "contrary to the majority's unsupported suggestion that this kind of 'proportionality approach' is unprecedented, the Court has applied it in various constitutional contexts, including election-law cases, speech cases and due-process cases."[47] In this respect, there are also scholarly analyses that emphasize, in the first place, how proportionality review has in fact been applied by US Courts in many contexts, and, in the second place, how—in contexts in which it has *not* been applied—proportionality would have been a valid alternative to the review templates which have so far been adopted and which result in categorical rules that are "increasingly uncertain and complex" (Jackson 2015, 3129).

This emphasis on proportionality shows how it has come to be a dominant approach in the theory and practice of rights adjudication, not only in Europe but also in the United States—an essential part of a *global* constitutionalism (Stone Sweet and Mathews 2008), albeit with some differences among proportionalist models.

6.2 Proportionality Between Ad Hoc and Definitional Balancing

As noted in the previous sections, the connection between balancing and proportionality review is of crucial relevance in two main respects. First, the inclusion of a balancing test in the review template forms the basis for a distinction between two, fundamental, versions of proportionality *in* review, that is, proportionality balancing and means-ends proportionality. Second, when the proportionality-review template incorporates a balancing test, questions arise about the nature of this test and about whether it is a legitimate instrument of review. Judicial recourse to balancing raises many deeply controversial questions, not least of which whether it is compatible with legal certainty. In this respect, the main criticism is that the balancing test incorporated into proportionality review is ad hoc and therefore leads to unstable, uncertain outcomes.

Indeed, according to a widely applied distinction, balancing in adjudication can take two different forms. It can be ad hoc, that is, narrowed to the case at hand, or *definitional*, when its outcomes in a concrete case serve as the basis for a general rule aimed at covering subsequent cases of the same kind (Nimmer 1968; Aleinikoff 1987). Ad hoc balancing only provides a particular solution to the conflict arising among the principles involved in the case at hand, while definitional balancing points to a general solution applying to future conflicts among the same principles.

[47] District of Columbia v. Heller, 554 US 570, 690 (2008).

The chief difference between ad hoc and definitional balancing, then, is that from the latter "a rule emerges" and this rule can be applied in future cases "without the occasion for further weighing of interests" (Nimmer 1968, 945).[48]

The distinction between these two forms of balancing has been at the centre of a wide debate.[49] On the one hand, the judicial use of ad hoc balancing techniques is criticized for producing unstable results, whereas definitional balancing is presented and defended as the appropriate way of coping with conflicts arising between fundamental rights/interests, as in the case of conflicts between free speech and other constitutional values (ibid.).[50] On the other hand, definitional balancing is criticized for its false promises: it wouldn't be able to generate rules that apply to future cases without further weighing the interests at stake (Aleinikoff 1987). Definitional balancing, from this perspective, does not deliver the legal certainty it promises due to the fact that "new situations present new interests and different weights for old interests," and "if these are allowed to re-open the balancing process, then every case becomes one of an 'ad hoc' balance, establishing a rule for that case only." From this perspective, the distinction between definitional and ad hoc balancing is "artificial": balancing can be definitional only "if the Court stops thinking about the question" (ibid., 980).

However, this distinction is relevant for our analysis because, as mentioned, the standard version of proportionality *in* review—Proportionality balancing—is criticized for incorporating an ad hoc version of balancing, thereby yielding instability and uncertainty in judicial decision-making. The problem would be that application of the stricto sensu proportionality test narrows the scope of judicial balancing to the case at hand (Moreso 2012, 38–39; von Bersntorff 2014; Bernal Pulido 2007, 194–199). Indeed, the resulting decisions would be based on balancing that courts are called on to do from time to time, on the basis of the weight that rights carry in the specific circumstances of the case at hand: the outcomes would be neither general nor predictable.

This criticism, however, overlooks the "definitional" features of Proportionality balancing. In fact, this review template combines the assessment of the concrete weight carried by rights—and the balance struck among them in specific cases—with the production of general, defeasible rules. On this point, Alexy (2010a, 51–52) argues that the solution to a conflict among principles results in the establishment of a "conditional relation of precedence" between them, which takes the circumstances

[48] It is the existence of such a rule that "makes it more likely that the balance originally struck will continue to be observed despite new and perhaps otherwise irresistible pressures" (Nimmer 1968, 945).

[49] This dispute loomed large in the debate on the US Supreme Court's adjudication of cases regarding the protection of free speech and mainly focused on the viability, and desirability, of a "definitional" approach to balancing. See Nimmer (1968), Schauer (1981), Aleinikoff (1997).

[50] According to Nimmer (1968, 942), definitional balancing is "a third approach which avoids the all or nothing implications of absolutism versus ad hoc balancing [...]. That is, the Court employs balancing not for the purpose of determining which litigant deserves to prevail in the particular case, but only for the purpose of defining which forms of speech are to be regarded as 'speech' within the meaning of the first amendment."

of the case into account. This is defined a *conditional* relation because the decision about the conditions of precedence among rights/interests is based on the circumstances of the case, such that, when these conditions change, so does the relation of precedence. Even so, the preferential statement establishing a precedence among principles "gives rise to a rule requiring the consequences of the principle taking precedence, should the conditions of the precedence apply" (ibid., 53, 101). From this perspective, Alexy frames under a *law of competing principles* the connection that holds between conditional relations of precedence and rules: "The circumstances under which one principle takes precedence over another constitute the conditions of a rule which has the same legal consequences as the principle taking precedence" (ibid., 54).

In other terms, the results of balancing serve as the basis of a rule under which subsequent cases can be decided. And such rules are prima facie rules, in the sense that it is possible to incorporate exceptions into them, with the result that they loose their definitive character. But the prima facie character of rules is different from the prima facie character of principles. A principle is "trumped when some competing principle has a greater weight," while a rule is not automatically trumped when the underlying principle has less weight than a competing principle applying to the same case. In fact, there are also formal principles that must be weighed, like the one establishing that the "rules passed by an authority acting within its jurisdiction are to be followed" or the one under which we "should not depart from established practice without good reason." The strength of the prima facie character of rules, then, depends in part on the weight of formal principles (Alexy 2010a, 57–58). In these terms, proportionality balancing is neither ad hoc nor definitional, but seeks to combine flexibility with legal certainty; a case-by-case assessment of conflicting norms with the production of rules that can guide future assessments.

7 Proportionalism *of* Review

Proportionalism *of* review comes in two fundamental versions: *optimizing proportionality* and *state-limiting proportionality* (Rivers 2006; Young 2014). These two versions are grounded in different conceptions of judicial review and its limits.

Optimizing proportionality is a review model aimed at establishing a balance among fundamental rights and other rights/interests "in the best possible way" (Rivers 2006, 176). *State-limiting proportionality* is a model essentially aimed at protecting rights and enforcing their primacy in the constitutional system on the basis of a conception of judicial review "as a set of tests warranting judicial interference to protect rights" (ibid., 176). In the first case, the focus of judicial review is on the degree of protection deserved by rights in light of how relevant the protected interests in conflict are. In the second case, the judicial focus is on the limits on state action deriving from the primacy of rights.

The main difference between these models lies in the judicial attitude towards normative conflicts involving rights. On the optimizing model, judicial review determines, from time to time, the extent to which rights can be optimized and protected against illegitimate state interference. On the state-limiting model, judicial review expounds and sets the limits that rights place on state action, defining the scope of the latter on the basis of the scope of the former.

On the basis of this reconstruction, *optimizing proportionality* characterizes the outcomes of judicial review as "open-ended" and requires us to draw a distinction between two orders of issues: on the one hand, are the substantive issues concerning the legitimacy of interferences with rights designed to protect conflicting interests; on the other hand, are the formal issues relating to the responsibility of courts in ensuring that such interferences are justified (Rivers 2006, 177).

On the *optimizing proportionality* model, judicial balancing lies at the core of an "open-ended" review of the reasons that may justify sacrificing a right to satisfy conflicting interests/rights. There are no substantive constraints on the outcomes of this review, which may differ depending on the weights the different interests at stake are found to carry on each occasion.

The *state-limiting proportionality* model, by contrast, places substantive constraints on judicial review, since it requires judicial scrutiny to ultimately preserve the primacy of rights by identifying, and enforcing, the limits on state action that derive from it. This model on the one hand provides courts with solutions that place heavy substantive constraints and on the other hand entrusts courts with a role as state-limiting actors. In these terms, the state-limiting version of proportionality review does not come with any clear distinction between substantive and formal issues in rights adjudication: the solution to the former—namely the primacy of rights—stands as a solution to the latter—a "default" state-limiting function that courts are called on to perform (id.).

Further elaborating on the distinction between optimizing-proportionality and state-limiting proportionality, Young (2014) argues that these two versions of proportionalism are not alternative but complementary. The first version aims to determine the degree and the scope of rights protection by balancing rights against other conflicting rights/interests; the second version aims to determine the limits within which institutional action can legitimately impact on rights, given a certain relation of precedence among rights and conflicting interests. The first version therefore relies on an interest-based conception of rights, as protections of interests that may or may not prevail over interests of a different sort; the second version instead relies on conceptions of rights that ascribe a strong lexical priority to them (ibid., 51). Nonetheless, these two paradigms of rights adjudication are not incompatible, but rather allow judicial review to perform different functions that can be combined. In the case of optimizing proportionality, judicial review performs the function of contributing to the determination of the scope and degree of protection accorded to rights in constitutional systems in which they are still uncertain. In the case of state-limiting proportionality, judicial review performs the function of specifying the limits imposed on institutional action by the scope and degree of rights protection in those contexts in which they are already fixed and established.

In these terms, the two models of proportionality work at different levels of abstraction. State-limiting proportionality operates in contexts in which the degree and scope of rights protection are determined; therefore, it works at an intermediate level of abstraction as a model of adjudication requiring that rights be taken as "intermediate conclusions,"[51] without entering into the question of their basis or justification. Otherwise, in the case of optimizing proportionality, judicial reasoning unfolds at a higher level of abstraction because it operates in contexts where the degree and scope of rights protection are not yet established: it must elaborate those intermediate conclusions to determine the content and degree of rights protection (Young 2014, 63–66; Luterán 2014, 24–25).

The option for one or the other form of proportionality, and their possible combinations, depends on the role that constitutional justice is called on to play in a specific context, taking into account the extent to which the content and level of protection of rights are determined and established.

In these terms, the differentiation between optimizing proportionality and state-limiting proportionality highlights two relevant aspects.

First, it points out the distinction between the substantive dimension and the formal dimension of rights adjudication, that is, between issues relating to the acceptability of interference with rights and the issues relating to the scope and intensity of the judicial scrutiny of such interference. Second, it draws attention to the ductility of proportionalist adjudication, not only in the substantive dimension, where it grounds and combines different argumentative frameworks, but also in the formal dimension. In the latter dimension, proportionalism covers a range of different models of rights adjudication, so that the scope and intensity of judicial review can be *proportionally* adjusted—in optimizing and/or in state-limiting terms—depending on the context in which courts adjudicate interferences with rights.

From this perspective, proportionalism not only guides the review of conflicts among rights/interests at a substantive level but also provides (meta-) judicial guidelines for setting—and justifying—the scope and intensity of judicial review at a formal level. Proportionalism, more precisely, provides justificatory frameworks for judicial action, requiring courts to adjust—and account for—their action in terms of proportionality, that is, as the adequate means for the realization of a legitimate end.

Along these lines, proportionalism—with the nested balancing mechanisms—has been identified as the appropriate approach to formal questions concerning judicial review and, in particular, to the conflicts among constitutional competence norms that come about at a formal level and regard the exercise of judicial review (Klatt 2015). The idea is that such norms, like constitutional norms that protect rights, are norms of principle that can come into conflict and thus need to be optimized by means of balancing. Therefore, when a conflict arises among competence norms with respect to the exercise of judicial review, this conflict should be settled by balancing those norms against one other, to this end taking into account the weight carried by the different institutional reasons at stake in a given case.

In other terms, the "institutional problem" concerning the determination of the scope and intensity of judicial review is presented as a conflict between formal

[51] The reference is to Raz's (1986, 181) idea of rights as "intermediate conclusions in arguments from ultimate values to duties."

principles of competence, that is, between the political competence to decide on certain issues and the judicial competence to review those decisions (Klatt 2015). From this perspective, “if it is correct that the institutional problem of judicial review is a conflict of formal principles, rather than of rules, then the solution of that problem is not to be found by means of interpreting competence norms. Rather, a balancing procedure has to be employed” (ibid., 364). And this procedure brings into the formal dimension of rights adjudication those balancing mechanisms that we already analysed as the core components of the standard proportionality template that is widely applied in the substantive dimension of the judicial review.

8 Alternative Approaches

If, and how, courts should engage in proportionalist adjudication is very controversial, especially when it comes with the application of a balancing procedure.

On the one hand, proponents of balancing, and of proportionality, claim that this is a rational procedure and that, given the structure of constitutional rights, it makes it possible to appropriately deal with conflicts among such rights. From this perspective, balancing provides judicial reasoning with an argumentative structure and leads it to acceptably justified decisions (Alexy 2010a; Barak 2010; Beatty 2004; Klatt and Meister 2012; Kumm 2010). On the other hand, critics of judicial balancing argue, among other things, that it is impossible to either do it rationally or characterize its results in general and predictable terms as required by legal certainty. Therefore, the judicial balancing of constitutional rights would be an arbitrary and illegitimate exercise of judicial power (Habermas 1996; Tsakyrakis 2009; Webber 2009, 2014).

In general, proportionalist-balancing approaches characterize judicial review as aimed at expounding the normative reasons expressed by rights, weighing them in light of the factual and legal circumstances and protecting them as much as their weight requires, being balanced against the weight of conflicting/competing rights/interests. These approaches, as we saw, are grounded in a “broad” construction of constitutional rights conceived as principled norms that can come into conflict. These norms have an open-ended structure and a weight dimension, and therefore require to be weighed and balanced against one another. If we follow this route, then, balancing can take different forms, and its justification can be more or less distant from the circumstances of the case at hand.

The alternative approaches to proportionality-balancing take a different route. They ground two main adjudicatory strategies, namely categorization and specificationism, that are both based on a conception of constitutional rights norms as rule-like norms. When these norms “compete” for application to a case, the appropriate specification of their scope makes it possible to identify which among the competing norms covers the case and must be therefore applied. In other terms, both strategies rely on the idea that the proper determination of the scope of rights "dissolves" such conflicts as seem to emerge among them.[52]

The first strategy—categorization—draws attention to the content of rights and requires us to proceed by narrowing that content so as to conceptually isolate rights from other potentially conflicting interests and to grant them full protection. This strategy, then, characterizes the judicial activity as aimed at "labelling and classifying" (Sullivan 1993, 241) the many facets of rights, along with the limits they place on the institutional action: the resulting taxonomy would allow courts to rely on a conceptually predetermined scope of rights and to identify what pertains to it and what doesn't.[53]

The second strategy—specificationism—focuses on the scope of rights and requires us to progressively expound and specify its boundaries, as implied by the scope of other rights/interests (Wellman 1995; Richardson 1990; Scanlon 2000; Moreso 2012; Webber 2009; Oberdiek 2004). In these terms, specificationism sets out to "reduce the scope of principles by preserving their stringency," which "amounts to conceiving the formulation of principles as incomplete and expanding them in a manner that, conveniently hedged, the scope of principles remains different and yet they are not in conflict" (Moreso 2012, 36).

In both cases, the adjudicatory approach is to take as fully protected what is covered by the scope of rights, whether identified in positive or negative terms. As a result, we arrive at rules forbidding actions that violate such rights and permitting actions that do not.

However, it is controversial that the categorical and the specificationist approaches ground a balancing-free judicial review. Indeed, the task of specifying the content/scope of rights seems to require us to define their boundaries *in relation* to the boundaries of other rights/interests. This task can be hardly presented as free of balancing without paying the price of disguising the interaction among the different reasons that must be assessed in order to establish what falls within the scope of a right (Alexy 2010a, 208–209; Klatt and Meister 2012, 46–47).[54]

Rather, what seems to be at work is a sort of "exclusive" balancing, namely, a balancing that proceeds by isolating the reasons that fall within the scope of rights from those that fall outside, so as to hide the latter and give full visibility and protection to the former. In this way, a "residue" of reasons gets lost and there is the problem of accounting for the difference between violating a right and infringing a right (Thomson 1986; but also Oberdiek 2004). From this perspective, the outcomes of

[52] From a specificationist perspective (Wellman 1995, 279), "the only way to make sense of apparent cases of conflicting rights is to assert that, despite initial appearances, (at most) only one party in fact has a right in this relationship."

[53] According to Sullivan (1993, 241), "[c]ategorization is the taxonomist's style—a job of classification and labelling […] Once the relevant right and mode of infringement have been described, the outcome follows, without any explicit judicial balancing of the claimed right against the government's justification for the infringement." On categorical and balancing approaches to review see also Schauer (1981).

[54] "The outcome of a narrow interpretation of a fundamental right is always based on balancing, since it relies on the reasons for and reasons against the protection" (Klatt and Meister 2012, 46).

categorical and specificationist approaches to rights adjudication serve as a basis for rules that do not fully capture *all* the different reasons involved in judicial decision-making and therefore do not allow for their proper consideration, as is necessary if we are to make sense of the evolution of case law.

References

Aleinikoff, T.A. 1987. Constitutional law in the age of balancing. *The Yale Law Journal* 96: 943–1005.

Alexy, R. 2000. On the structure of legal principles. *Ratio Juris* 13: 294–304.

Alexy, R. 2003. On balancing and subsumption. A structural comparison. *Ratio Juris* 16: 433–449.

Alexy, R. 2010a. *A theory of constitutional rights*. Oxford: Oxford University Press (1st ed. in German 1985).

Alexy, R. 2010b. The construction of constitutional rights. *Law & Ethics of Human Rights* 4: 21–32.

Alexy, R. 2014. Constitutional rights and proportionality. *Revus* 22. https://philpapers.org/rec/ALECRA-2.

Arai-Takahashi, Y. 2002. *The margin of appreciation doctrine and the principle of proportionality in the jurisprudence of the ECHR*. Anrwerp: Intersentia.

Baier, K. 1965. *The moral point of view*. New York: Random House.

Barak, A. 2010. Proportionality and principled balancing. *Law & Ethics of Human Rights* 4: 1–16.

Barak, A. 2012a. Proportionality (2). In *The Oxford handbook of comparative constitutional law*, ed. M. Rosenfeld, and A. Sajó, 738–755. Oxford: Oxford University Press.

Barak, A. 2012b. *Proportionality. Constitutional rights and their limitations*. Cambridge: Cambridge University Press.

Beatty, D.M. 2004. *The ultimate rule of law*. Oxford: Oxford University Press.

Bernal Pulido, P. C. 2007. *El principio de proporcionalidad y los derechos fundamentales*. Madrid: Centro de Estudios Políticos y Constitucionales (1st ed. 2003).

von Bernstorff, J. 2014. Proportionality without balancing: Why judicial ad hoc-balancing is unnecessary and potentially detrimental to the realization of individual and collective self-determination. In *Reasoning rights: Comparative judicial engagement*, ed. L. Lazarus, C. McCrudden, and N. Bowles, 63–86. Oxford: Hart Publishing.

Besson, S. 2005. *The morality of conflict. Reasonable disagreement and the law*. Oxford: Hart Publishing.

Bomhoff, J. 2008. Luth's 50th anniversary: Some comparative observations on the German foundations of judicial balancing. *German Law Journal* 9: 121–124.

Borowski, M. 2013. On apples and oranges. Comment on Niels Petersen. *German Law Journal* 8: 1409–1418.

Campbell, K. 2016. Legal rights. In *The Stanford encyclopedia of philosophy,* ed. E. Zalta. https://plato.stanford.edu/archives/win2016/entries/legal-rights/.

Celano, B. 2013. I diritti nella jurisprudence anglosassone contemporanea. Da Hart a Raz. In Id., *I diritti nello Stato costituzionale*. Bologna: il Mulino.

Chemerinsky, E. 2011. *Constitutional law: Principles and policies*. New York: Wolters Kluwer.

Cohen-Eliya, M., and I. Porat. 2009. The hidden foreign law debate in Heller: The proportionality approach in American constitutional law. *San Diego Law Review* 46: 367–414.

Cohen-Eliya, M., and I. Porat. 2010. American balancing and German proportionality: The historical origins. *I·CON International Journal of Constitutional Law* 8: 263–286.

Cohen-Eliya, M., and I. Porat. 2011. Proportionality and the culture of justification. *The American Journal of Comparative Law* 2: 463–490.

Cohen-Eliya, M., and I. Porat. 2013. *Proportionality and constitutional culture*. Cambridge: Cambridge University Press.

Connell, F.J. 1967. Double effect. *Principle of New Catholic Encyclopedia* 4: 1020–1022.

Denninger, E. 1990. *Der gebändigte Leviathan*. Baden-Baden: Nomos.

Dworkin, R. 1977. *Taking rights seriously*. Cambridge, MA.: Harvard University Press.

Ellis, E. (ed.). 1999. *The principle of proportionality in the laws of Europe*. Oxford: Hart Publishing.

Emiliou, N. 1996. *The principle of proportionality in European law: A comparative study*. London: Kluwer Law Intl.

Endicott, T. 2014. *Proportionality and Incommensurability*. http://www.academia.edu/13177344/Proportionality_and_Incommensurability.

Forst, R. 2012. *The right to justification: elements of a constructivist theory of justice*. New York, NY: Columbia University Press.

Gardbaum, S. 2016. *Positive and horizontal rights: Proportionality's next frontier or a bridge too far?* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2726794.

Grimm, D. 2007. Proportionality in Canadian and German constitutional jurisprudence. *University of Toronto Law Journal* 57: 383–397.

Grimm, D. 2016. *Constitutionalism: Past, present, and future*. Oxford: Oxford University Press.

Gunn, T.J. 2005. Deconstructing proportionality in limitations analysis. *Emory International Law Review* 19: 465–498.

Habermas, J. 1996. *Between facts and norms. Contributions to a discourse theory of law and democracy*. Cambridge, MA: The MIT Press.

Hare, R.M. 1985. *Moral thinking*. Oxford: Oxford University Press.

Harel, A., and T. Kahana. 2010. The easy core case for judicial review. *Journal of Legal Analysis* 2: 227–256.

Harbo, T. 2010. The function of the proportionality principle in EU law. *European Law Journal* 16: 158–185.

Hart, H.L.A. 1982. *Essays on bentham: studies in jurisprudence and political theory*, Oxford: Clarendon.

Huscroft, G., B.W. Miller, and G. Webber (eds.). 2014. *Proportionality and the rule of law: Rights, reasoning, justification*. Cambridge: Cambridge University Press.

Jackson, V. 2004. Being proportional about proportionality. *Constitutional Commentary* 21: 803–859.

Jackson, V. 2015. Constitutional law in an age of proportionality. *The Yale Law Journal* 8: 3094–3196.

Klatt, M. 2015. Positive rights: Who decides? Judicial review in balance. *I·CON International Journal of Constitutional Law* 13: 354–382.

Klatt, M., and M. Meister. 2012. *The constitutional structure of proportionality*. Oxford: Oxford University Press.

Kramer, M. 1998. Rights without trimmings. In *A debate over rights. Philosophical enquiries*, ed. M. Kramer, N.E. Simmonds and H. Steiner, 7–111. Oxford: Oxford University Press.

Kumm, M. 2010. The idea of socratic contestation and the right to justification: The point of rights-based proportionality review. *Law & Ethics of Human Rights* 4: 141–175.

Kumm, M., and A.D. Walen. 2013. Human dignity and proportionality: Deontic pluralism in balancing. New York University Public Law and Legal Theory Working Papers 383. http://lsr.nellco.org/nyu_plltwp/383.

Luterán, M. 2014. The lost meaning of proportionality. In *Proportionality and the rule of law: Rights, justification, reasoning*, ed. G. Huscroft, B.W. Miller, and G. Webber, 21–42. Cambridge: Cambridge University Press.

McIntyre, A. 2014. Doctrine of Double Effect. In The Stanford Encyclopedia of Philosophy, ed. E. Zalta. https://plato.stanford.edu/archives/win2014/entries/double-effect/.

Möller, K. 2012a. *The global model of constitutional rights*. Oxford: Oxford University Press.

Möller, K. 2012b. Proportionality: Challenging the critics. *I·CON International Journal of Constitutional Law* 10: 709–731.

Möller, K. 2013. Proportionality and rights inflation. LSE Law, Society and Economy Working Papers 17. http://ssrn.com/abstract=2272979.

Möller, K. 2014. The global model of constitutional rights: A response to Afonso da Silva, Harel, and Porat. LSE Law, Society and Economy Working Papers 28. http://ssrn.com/abstract=2526685.

Moreso, J.J. 2012. Ways of solving conflicts of constitutional rights: proportionalism and specificationism. *Ratio Juris* 25: 31–46.

Mureinik, E. 1994. A bridge to where? Introducing the interim bill of rights. *South African Journal on Human Rights* 10: 31–48.

Nimmer, M.B. 1968. The right to speak from times to time: First amendment theory applied to libel and misapplied to privacy. *California Law Review* 56: 935–967.

Oberdiek, J. 2004. Lost in moral space: On the infringing/violating distinction and its place in the theory of rights. *Law and Philosophy* 23: 325–346.

Petersen, N. 2013a. How to compare the length of lines to the weight of stones: Balancing and the resolution of value conflicts in constitutional law. *German Law Journal* 8: 1387–1408.

Petersen, N. 2013b. Proportionality and the incommensurability challenge—Some lessons from the South African constitutional court. New York University Public Law and Legal Theory Working Papers. Paper 384. http://lsr.nellco.org/nyu_plltwp/384.

Pino. G. 2014a. *Diritti fondamentali e principio di proporzionalità*. http://www1.unipa.it/gpino/Pino,%20Diritti%20fondamentali%20e%20principio%20di%20proporzionalit%E0_RP.pdf.

Pino, G. 2014. Proporzionalità, diritti, democrazia. *Diritto e Società* 3: 597–628.

Raz, J. 1986. On the nature of rights. In Id., *The Morality of Freedom* 175–191. Oxford: Clarendon.

Richardson, H.S. 1990. Specifying norms as a way to resolve concrete ethical problems. *Philosophy & Public Affairs* 19: 279–310.

Rivers, J. 2006. Proportionality and variable intensity of review. *Cambridge Law Journal* 65: 174–207.

Ross, W.D. 2002. *The right and the good*. New York: Oxford University Press (1st ed. 1930).

Scanlon, T. 2000. Intention and permissibility. *Proceedings of the Aristotelian Society* 74: 301–317.

Schlink, B. 2012a. Proportionality (1). In *The Oxford handbook of comparative constitutional law*, ed. M. Rosenfeld, and A. Sajó, 718–737. Oxford: Oxford University Press.

Schlink, B. 2012b. Proportionality in constitutional law: Why everywhere but here? *Duke Journal of Comparative & International Law* 22: 291–302.

Schauer, F. 1981. Categories and the first amendment: A play in three acts. *Vanderbilt Law Review* 34: 265–307.

Stone, Sweet A., and J. Mathews. 2008. Proportionality balancing and global constitutionalism. *Columbia Journal of Transnational Law* 47: 73–165.

Stone, Sweet A., and J. Mathews. 2011. All things in proportion? American rights doctrine and the problem of balancing. *Emory Law Journal* 60: 101–169.

Sullivan, K. 1993. Categorization, balancing, and government interests. In *Public values in constitutional law*, ed. S. Gottlieb. Ann Arbor, MI: University of Michigan Press.

Thomson, J.J. 1986. Self-defense and rights. In Id. *Rights, Restitution, and Risk: Essays in Moral Theory* 33–48. Cambridge, MA: Harvard University Press.

Thomson, J.J. 1992. *The realm of rights*. Cambridge, MA: Harvard University Press.

Thornburn, M. 2016. Proportionality. In *Philosophical foundations of constitutional law*, ed. D. Dyzenhaus, and M. Thornburn, 305–322. Oxford: Oxford University Press.

Tsakyrakis, S. 2009. Proportionality: An assault on human rights? *I·CON International Journal of Constitutional Law* 7: 468–493.

Verdirame, G. 2015. Rescuing human rights from proportionality. In *Philosophical foundations of human rights*, ed. R. Cruft, S.M. Liao, and M. Renzo, 341–357. Oxford: Oxford University Press.

Waldron, J. 1989. Rights in conflict. *Ethics* 3: 503–519.

Webber, G. 2009. *The negotiable constitution: On the limitation of rights*. Cambridge: Cambridge University Press.

Webber, G. 2010. Proportionality, balancing, and the cult of constitutional rights scholarship. *Canadian Journal of Law & Jurisprudence* 23: 179–202.

Webber, G. 2013. On the loss of rights. LSE Law, Society and Economy Working Papers 16. http://ssrn.com/abstract=2272978.

Webber, G. 2014. On the loss of rights. In *Proportionality and the rule of law: Rights, justification, reasoning*, ed. G. Hushcroft, B.W. Miller, and G. Webber, 123–154. New York, NY: Cambridge University Press.

Webber, G. 2016. Proportionality and absolute rights. LSE Law, Society and Economy Working Papers 10. http://ssrn.com/abstract=2776577.

Wellman, C.H. 1995. On conflicts between rights. *Law and Philosophy* 14: 271–295.

Wenar, L. 2015. Rights. In *The Stanford encyclopedia of philosophy*, ed. E. Zalta. https://plato.stanford.edu/archives/fall2015/entries/rights/.

Young, A.L. 2014. Proportionality is dead: Long live proportionality. In *Proportionality and the rule of law: Rights, justification, reasoning*, ed. G. Hushcroft, B.W. Miller, and G. Webber, 43–66. Cambridge: Cambridge University Press.

Young K.G. 2012. *Constituting economic and social rights*. Oxford: Oxford University Press.

Yowell, P. 2014. Proportionality in United States constitutional law. In *Reasoning rights: Comparative judicial engagement*, ed. L. Lazarus, C. McCrudden, and N. Bowles, 87–116. Oxford: Hart Publishing.

Zanghellini, A. 2017. Raz on rights: Human rights, fundamental rights, and balancing. *Ratio Juris* 30: 25–40.

A Quantitative Approach to Proportionality

Giovanni Sartor

1 Introduction

The present contribution discusses the extent to which value-based reasoning, as deployed in proportionality arguments, may include quantitative reasoning and constraints.[1] It will therefore address from a different angle the proportionality debate, which has been examined from a legal and philosophical perspectives in chapter "Balancing, Proportionality and Constitutional Rights" of part III, by Giorgio Bongiovanni and Chiara Valentini, and link proportionality to teleological reasoning, as addressed in chapter "Choosing Ends and Choosing Means: Teleological Reasoning in Law" of part II by Lewis Kornhauser. Assessing the merit of alternative choices relatively to values involves considering the magnitudes that quantify the impacts of such choices on the realisation of these values and the weights of the affected values. Even though this reasoning does not use numerical symbols, it still deals with quantities (it subtracts, multiplies divides, etc., such quantities) and is therefore subject to the basic laws of arithmetic. In fact, according to research on cognitive and evolutionary psychology, processing non-symbolic approximate magnitudes is a fundamental cognitive capacity, which seems to be deployed also when we are reasoning with values. This capacity needs to be integrated with logic and argumentation to provide a comprehensive account of value-based reasoning.

On the basis of this assumption, a conceptual framework is developed for reasoning with values. Ways to determine the impacts of a choice on single values are presented, and ways to determine the associated utilities and merge the utilities into

[1] On the approach presented in this chapter, see Sartor (2010, 2013).

G. Sartor (✉)
Dipartimento di Scienze giuridiche, Università di Bologna, Bologna, Italy
e-mail: giovanni.sartor@gmail.com

G. Sartor
European University Institute, Florence, Italy

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_21

a single measure of the merit of that choice are introduced. Some issues pertaining to the comparison of alternative choices are addressed.

It is shown how the standards of proportionality assessments—suitability, necessity and proportionality in a strict sense—fit in the proposed framework.

Finally, it is considered how value-based assessments may be constrained by the requirement of consistency with precedent assessments of the same kind.

This framework is applied to examples of legal reasoning in judicial decision-making.

2 Quantitative Reasoning Without (Symbolically Expressed) Numbers

When we are to assess whether a decision α duly realises some values, we need to compare the extent to which these values are implemented by α and the extent they would be implemented if a different choice β were made instead of α (where β may consist in not interfering with the status quo or in changing it in a different way). On this basis, as we shall see, we need to determine differential merit of making choice α.

This raises the issue of how we are going to determine the impact of a choice on all values at stake and aggregate such impacts into a determination of the overall benefit or loss that is provided by that choice, as compared with different possible choices. If we could obtain appropriate numbers,[2] it seems that some mathematics should provide the answer on the merit of a choice. For this purpose, we should need numbers expressing the different impacts of our choices (in all possible scenarios) on the implementation of the values at stake and functions determining, for each such impact, the corresponding gain or loss in the overall benefit being delivered. However, in most legal cases (at least when constitutional adjudication is at issue), we do not have sensible ways for assigning numbers and constructing the corresponding functions. Nor have we an exhaustive set of preferences between all possible combinations of the different impacts on values, which may be represented as a utility function, in accordance with the so-called representation theorems used in economics.[3] This makes quantitative methods used in decision theory and cost-

[2] I use the term "number" to refer only to the cases where a quantity is expressed with the symbols (the numerals) or a particular number system. When a quantity is represented (e.g. graphically, or mentally) without the use of such symbols, I use the term "magnitude."

[3] According to the so-called Morgenstern-Von Neumann representation theorem, if we have a set of preferences among alternatives, and these preferences are complete, transitive, independent and continuous, then we can build a utility function assigning a numerical utility to each alternative, in such a way that any alternative (strictly) preferred to another would have a higher utility than the latter.

benefit analysis not directly applicable to many legal contexts, and in particular, to constitutional decisions involving impacts on different values.[4]

One explanation of our ability to assess the impacts of our choices on the relevant values—though we cannot sensibly not express these impacts through numbers—is that people possess some, inborn or acquired, capacity to reason with non-numerical quantities.[5] In fact, experiments have shown that we can make computations with quantities without associating numerical symbols to such quantities (see Gallistel and Gelman 2005). This capacity is not limited to ordinal comparisons, namely to assessing whether a certain object is more or less than another (with regard to dimensions such as length, volume, weight, speed). It also covers cardinal measures: even without numbers, we are able to assess, though in a very approximate way, the size (the cardinal measure) of an object or the extent of its difference from another. To express such non-numerical cardinal evaluations, we often refine our ordinal assessment with adverbs. For instance, we may say that this object is a little, fairly, a lot larger, or smaller, or quicker, than that object. We can sometimes map such approximate cardinal evaluations into quantitative proportions and relations without referring to a general unit of measure and without engaging in explicit numerical computations. For instance, we may say that an entity is about a half, two times, three times larger, or smaller, or quicker than another. Thus, not only can we compare two lines and establish which one is longer, but we can say that one line is twice longer than the other, or that a line is the sum of the two lines of different sizes, without making numerical calculations.

Apparently, this kind of mathematical competence is quite widespread in the animal kingdom. Animals not only are able to order objects according to their size, but they can also perform tasks that involve processing magnitudes: they compute distances by summing up the extent of successive displacements, they make visits to caches according to the difference between the time when the food was stored and its expected rotting time, and they remain in different locations according to ratios between time spent and rewards obtained, etc.

> Research with vertebrates, some of which have not shared a common ancestor with man since before the rise of the dinosaurs, implies that they represent both countable and uncountable quantity by means of mental magnitudes […] The system of arithmetical reasoning with these mental magnitudes is closed under the basic operations of arithmetic, that is, mental magnitudes may be mentally added, subtracted, multiplied, and divided without restriction. (Gallistel et al. 2006, 259)

Thus, it seems that there exists an inborn ability to represent and mathematically process mental magnitudes, which is deployed without translating these magnitudes into the linguistic symbols (the numerals) of a number system. Contrary to a famous

[4] This does not exclude that the methods of decision theory and cost-benefit analysis can be usefully deployed in many cases; for a technical account of multi-criteria decision-making, see Keeney and Raiffa (1993).

[5] This assumption is not meant to exclude that other ways of reasoning may also be significant for this purpose, such as the capacity of making analogies out of cases, or of building arguments. We rather integrate these different skills in complex value assessment.

statement by the mathematician Leopold Kronecker ("God made the integers; all else is the work of man"), it seems nature has endowed us, and other animals, with the primitive ability to store and process continuous (though approximated or noisy) mental magnitudes,[6] quantities that are only mappable into real numbers (since they include also negative magnitudes, fractions, and even irrational magnitudes, such square roots). On the top of this ability, humans have the additional possibility of using symbols for expressing such quantities and making them more precise. Our mind, however, continues to map numerical values into analogical magnitudes, and we do when making quick, unreflected, judgments. We may possibly say, using the terminology of Kahneman (2011) that reasoning with analogical magnitudes pertains to our fast (parallel, intuitive and apparently effortless) thinking, while the corresponding numerical processes pertain to our slow (sequential, reflective and demanding) thinking.

According to Pollock (2006, Ch. 3), this capacity for intuitive cardinal assessment of quantities, which he calls "analogical quantitative cognition," applies not only to lengths, weights or volumes, but also to our likes and dislikes and to the realisation of our values.

I shall accept the assumption that humans have a basic (and largely inborn, though improvable by training and experience) intuitive capacity for non-symbolic quantitative reasoning, a capacity that includes not only assessing and comparing magnitudes, but also performing on such magnitudes approximate mathematical operations: sums, subtractions, proportions, multiplications and divisions (and even approximate differentiation and integration). I shall argue that exactly this capacity is involved in assessing impacts on values according to proportionality. We can deploy it in choices concerning our private life (choosing a car or a computer by balancing design, performance and cost; choosing a restaurant by considering quality of food, service and price, choosing a course of studies balancing interest and work-opportunities, etc.) but also when public choices have to be taken or assessed. For engaging in this kind of intuitive, or "analogical" quantitative reasoning, we do not need to translate quantities into numbers through measurement (which is an ability that only humans possess and in many domains only after adequate schooling): we just rely on our intuitive appreciation of the quantities involved and of their relations. When more precision is needed, and numerical quantification makes sense, we may move to symbolically expressed numbers, to test and refine our intuitions.

The nature of this mathematical capacity entails that mathematical relationships do not hold only among symbolically expressed numbers: they also constrain the

[6] Thus, apparently, these findings of contemporary cognitive science seem to validate Leibniz's principle of continuity (often expressed by the saying *natura non facit saltus*), at least with regard to the mental processing of quantitative information: "I also take it for granted that every created being is subject to change […] and even that this change is continuous in each." (G. W. Leibniz. Monadology. 1720, Section 10, translation by N. Rescher. G.W. Leibniz's Monadology: An Edition for Students. Pittsburgh: University of Pittsburgh Press, 1991). There are various methods for dealing with approximate quantities, but here I cannot even attempt at discussing them, and moreover, the general account here provided is meant to be neutral in this regard as much as possible. For a review of methods for reasoning for uncertainty, see, for instance, Parsons (2001).

process of our intuitive-analogical quantitative reasoning. Thus, such relationships can be used as a standard of rationality for that reasoning, and for facilitating the transition to numerical quantification, when possible and convenient. Finally, note that the assumption that we can reason with approximate quantities does not entail that we can precisely assess such quantities, nor that we can always determine with certainty whether one object's magnitude is bigger than another's. For instance, we may sometimes (though not in most cases) remain uncertain when comparing the lengths of two twisted lines or the volumes of two solid objects. Similarly, we may sometimes (though not in most cases) remain in doubt concerning the impacts of our choices on our values, and the comparative merits of such choices.

In the following section, I shall examine teleological reasoning in law as an instance of non-numerical quantitative reasoning and I shall derive some implications of this idea.

3 Basic Concepts

I shall specify certain notions that are needed in order to proceed in the analysis. First of all, I assume that we can quantify the level of the realisation of a value in a particular situation (where a situation is an actual or possible set of circumstances, including social and institutional arrangements).

Definition 1 (*Realisation-quantity of a value*) The realisation-quantity of a value v in a particular situation is the extent up to which v is realised in case that situation obtains. Let us write $Real_v\ (s)$ to denote the realisation-quantity of value v in situation s. We correspondingly denote as $Real_v\ (s_c)$, or simply $Real_v$, the level of realisation of v in the current situation s_c (the present state of affairs). Thus, $Real_v = q$ means that in the current situation the value v is realised in quantity q.

We can express our assessment of the realisation-quantity of a value in non-numerical term (e.g., we may say that privacy is protected to a sufficient extent in Country x, while its protection is low in Country y, that a large freedom of speech is enjoyed by the citizen of Country w, etc.) or in numerical term, when appropriate numerical indicators are available (as for GDP per head, employment rate, etc.). For some values (transparency, democracy, economic freedom, equality, non-discrimination, etc.), proxies are available according to various measurements, such as those that are used for ranking countries according to their levels of welfare or protection of human rights. However, even when no such proxies are available, we still engage in quantitative assessments, while being aware that such assessments are inevitable noisy, approximate and revisable.

Such assessments may be different according to different conceptions of the values at issue and individual attitudes and experience, but different people would show usually some consistency in making them. For instance, I think that very few people would disagree that a 50% increase in the revenue per head (with the same distribution) would provide a higher welfare, or that that storing personal data for a longer time would involve an additional limitation to privacy, or that extending the time for detention without judicial authorisation would additionally restrain individual liberty.

We may wonder, however, if it really possible to compare situations where values are realised in different ways. Assume for instance that we have to compare a situation where privacy is well protected against governmental interference but much less protected against commercial interference, and a situation where privacy is well protected again commercial interference, and much less protected against governmental interference. In case there really is a competition between two different aspects of value (we can increase privacy towards governmental bodies only by reducing privacy towards private bodies, or, more plausibly, given a fixed amount of resources available for the welfare of dependant people, we can increase the welfare of old people only by decreasing welfare for children), we could see the two aspects as distinct values, to be comparatively assessed, as it were a conflict between different values.

When a value is realised up to a certain extent in a certain situation, a certain amount of benefit or utility is delivered.

Definition 2 (*Utility-quantity concerning a value*) The utility-quantity concerning a value v is the amount of utility provided by the realisation of v. We write $Ut_v s$ to denote the utility which is obtained with regard to value v in situation s, where v is realised up to the extent $Real_v s$. Thus, $Ut_v s = q$ means that in situation s, consequently to the realisation of v, utility q is obtained.

Note that here I use *utility* as a "neutral" term denoting the amount of goodness (or badness, when the utility is negative) that is provided by a choice, without making any assumption on the nature or distribution of such goodness. Thus, the "utility" of a choice includes the assessment all of its aspects and consequences that affect the relevant values, increasing or decreasing their realisation.[7]

In the following, I shall consider how to move from the realisation $Real_v s$ of a value v to the utility $Ut_v s$ that is provided by that realisation. Obviously, people's assessment of the utility of the realisation of a value may be quite variable, and in particular, more variable than their assessment of the realisation-quantity of a value. However, some relations between such assessments may be considered to be invariant.

[7] In particular, I do not assume a utilitarian approach, according to which utility is to be viewed as happiness or preference satisfaction. On the contrary, here "utility" refers to the sum of all impacts on all (legally relevant) communal and individual values, since I assume that such impacts are independent. In principle, such utility might also be specified in such a way that the distribution of individual opportunities is subject to some fairness requirements. Then, such opportunities, as they follow from the realisation of the relevant values, according to their importance, would need to be allocated according to a scheme that is fair enough, in the sense that it balances the requirement of distributive fairness against the importance of increasing the total realisation of the concerned values. Here however, I shall not examine whether including fairness requirement in the model here proposed would require a relaxation of some of its assumption, such as the assumption of the independence of the utilities obtained by realising different values.

First of all, since values are by definition good things, we can assume that the utility provided by the realisation of a value, increases as the realisation of that value increases. Thus, we assume that relation between the realisation of a value and the corresponding utility is a monotonic function and indeed a strictly increasing one. We take this as a defeasible assumption, which expresses what is usually the case, and does not exclude that in certain cases over-realisation of a value can be counterproductive.

Assumption 1 (*Increasing utility from values*) A higher realisation of a value provides a higher utility. In other words, when the realisation-quantity of value v increases, also v's utility-quantity increases: if the realisation-quantity of v in a situation s_2 is higher that the realisation-quantity of v in s_1, than also the utility-quantity of v in situation s_2 is higher than its utility-quantity in s_1. In other words, if $Real_v(s_1) < Real_v(s_2)$ then $Ut_v(s_1) < Ut_v(s_2)$.

Thus, the utility resulting from the realisation of a value v will increase progressively, when v's realisation-quantity increases. For instance, a higher level of a value such as health, or environmental quality, or privacy, or freedom of speech, gives more utility than a lower level of the same value, all the rest (the realisation-quantity of the other values) remaining equal. Moreover, as we shall see in the following, the extent of this increase progressively diminishes as the realisation-quantity of the value gets higher (there is a diminishing marginal utility), but the relationship above still holds.

On the basis of the notions introduced, we can address impacts of actions (choices) on the realisation of values. We use Greek letters $\alpha, \beta, \ldots$ as variables ranging over actions. We assume that actions have outcomes, namely they make a change in the status quo, with one exception: the null action nil consists in letting things as they are (letting the status quo), or better, letting things evolve without our intervention

To simplify things, let us assume a deterministic framework, where each action has only one outcome (otherwise, we have to expand the current framework with probabilities): the unique outcome of an action α is the situation out (α) that would result from performing α, in the current situation. We are now able to specify the impact of an action α to a value v, namely the change the action α can make to the realisation of v. This is the difference between the extent up to which v would be realised by α, and the extent up to which it would be realised by not doing anything, i.e. by the null action nil.

Definition 3 (*Realisation impact*) The realisation impact of an action α on a value v, denoted as $\Delta Real_v(\alpha)$, is the difference between the realisation-quantities of v resulting from α and from nil. Let us denote the outcome of action α, namely the situation resulting from its performance, in the current situation, as $out(\alpha)$ and the realisation impact (the differential realisation) of an action α on a value v, as $\Delta Real_v(\alpha)$. Then, $\Delta Real_v(\alpha) = Real_v out(\alpha) - Real_v out(nil)$.

For instance, if α is a law prohibiting the use of a polluting substance which is currently in use in industrial processes, the realisation impact of α on health is the increased level of health that results from not having any longer the pollution caused by that substance, while α's realisation impact on productivity, is the decreased level of productivity which results from not using the substance in production processes.

The notion of realisation impact allows us to define what it means to promote or demote a value: promoting means increasing (having a positive impact on) the value's level of realisation and demoting means decreasing (having a negative impact on) it, as compared to nil.

Definition 4 (*Promotion and demotion of a value*) An action α promotes a value v if its realisation impact on v is positive ($\Delta Real_v\,(\alpha) > 0$); it demotes v if its realisation impact on v is negative ($\Delta Real_v\,(\alpha) < 0$).

Thus, a legislative choice which prohibits the use of a polluting substance may promote health and demote productivity; a legislative measure that makes Internet providers liable for violations of data protection by their subscribers may promote data protection and demote freedom of speech, etc. We can also characterise the utility of an action with regard to a value, as the differential utility-impact provided by that action with regard to that value: this is a measure of the difference in utility provided by the fact that the value is realised to a higher or lower extent.

Definition 5 (*Utility-impact of an action on a value*) The utility-impact of an action α on a value v is the difference between the utility-quantity by v resulting from α and from nil: $\Delta Ut_v\,(\alpha) = Ut_v out\,(\alpha) - Ut_v out\,(nil)$

Thus, in the above case of the prohibition α of the use of a polluting substance, we can say that since α, as compared with the status quo, promotes the value of health, while demoting the value of productivity, it increases the health-related utility concerning and decreases the productivity-related utility: $\Delta Ut_v\,(nil) = Ut_v out\,(nil) - Ut_v out\,(nil) = 0$

Corollary 1 *(Realisation- and utility-impact of nil) The above definitions entail that the realisation impact of nil on any value is 0 and so is nil's utility-impact.*

4 Impacts on Single Values

The notions we have described enable us to compare the impact of different choices on different values. First of all, we need to introduce a way to express that a choice α is superior to β with regard to its aggregate impact on a set of values.

Definition 6 (*Superiority with regard to a set of values*) We say that choice α is superior to choice β with regard to a set of values $\{v_1, \ldots, v_n\}$ and write $\alpha \succ_{\{v_1,\ldots,v_n\}} \beta$, if α's utility-impact on this set is higher than β's utility-impact on the same set. In other words, $\alpha \succ_{\{v_1,\ldots,v_n\}} \beta$ if and only if $\Delta Ut_{\{v_1,\ldots,v_n\}}\alpha > \Delta Ut_{\{v_1,\ldots,v_n\}}\beta$.

Note that since the utility-impact of *nil* is null (0), then a choice α is superior to *nil* with regard to a set of value, wherever the choice has a positive utility-impact on that set.

Consider, for instance, an environmental-protection measure α that prohibits the use of polluting substance and in this way promotes health and demotes productivity. Measure α is superior to *nil*, in case its utility-impact with regard to the combination of health and productivity is positive. In such a case, we would write: $\alpha \succ_{\{health,\, productivity\}} nil$.

We will come back later on how to establish superior utility with regard to a set of values. Let us first address impacts on a single value. When we are considering just one value, we can say that whenever the realisation impact on that value is positive, then the utility-impact on it is positive, given Assumption 1 (higher realisation of a value provides a higher utility by that value). A higher utility by a value entails superiority with regard to that value. In other words, since (a) $\Delta\mathrm{Real}_v\alpha > \Delta\mathrm{Real}_\mathrm{v}\beta$ entails $\Delta\mathrm{Ut}_v\alpha > \Delta\mathrm{Ut}_v\beta$ and (b) the latter entails $\alpha \succ_v \beta$, we can conclude (c) that $\Delta\mathrm{Real}_v\alpha > \Delta\mathrm{Real}_\mathrm{v}\beta$ entails $\alpha \succ_v \beta$. This leads us to the following corollary.

Corollary 2 *(Superiority (with regard to a value), according to contribution) Whenever α's realisation impact on value v is higher than β's, then α is superior to β with regard to v* ($\alpha \succ_v \beta$)*. In other words,* $\Delta Real_v\alpha > \Delta Real_v\beta$ *entails* $\alpha \succ_v \beta$*.*

Note that this corollary also applies to the comparison of a choice α with nil. Since nil provides 0 differential contribution to the realisation of any value, any choice giving a positive marginal contribution would be better than nil, and any choice giving a negative marginal contribution would be worse than it.

Consider, for instance, the enactment α of a law allowing wiretapping only on the basis of a judicial warrant, when the current regulation (the status quo, i.e. *nil*), allows police authorities to wiretap any communication in their criminal investigations. Under such circumstances, enactment α has a positive impact on privacy and a negative impact on crime prevention: $\alpha \succ_{privacy}$ nil, while nil $\succ_{crime\_prevention}$ α.

The utility-impact of a choice α on a set of values is just the sum of the utility-impacts that α delivers by affecting of each of these values.

Corollary 3 *(Utilities from different values) Given a choice α having an impact on values* $v_1, \ldots, v_n$*, the utility-impact of α with regard to the set of those values is the sum* $i_1 + \cdots + i_n$ *of the utility-impacts* $i_1, \ldots, i_n$ *of α with regard to each of such values. In other words,* $\Delta Ut_{\{v_1,\ldots,v_n\}}\alpha = \Delta Ut_{v_1}\alpha + \cdots + \Delta Ut_{v_n}\alpha$*.*

For instance, consider a law exempting host providers from liability for the privacy violations committed by their users, as compared to a situation where providers are considered to be liable for such a violation. The total utility provided by such a law results from the sum of the utility-impacts it provides on the different values involved, its positive utility-impact on freedom of expression, freedom of information and economic efficiency, and its negative utility-impact on privacy.

5 Pareto-Superiority

Let us now extend our analysis to choices having an impact on multiple values. The easy case is when α, as compared to β, provides a higher realisation of some values and does not provide a lower realisation of any other value. In this case, we say that α is Pareto-superior to β.

Definition 7 (*Pareto-superiority*) We say that choice α is Pareto-superior to β if there exists a value v_1 such that the utility-impact of α on v_1 is higher than β's and for no value v_2, the utility-impact of β on v_2 is higher than α's. In other words, α is Pareto-superior to β with regard to $\{v_1, ..., v_n\}$ if (a) there exists a $v_i \in \{v_1, \ldots, v_n\}$ such that $\Delta Ut_{v_i}\alpha > \Delta Ut_{v_i}\beta$ and (b) there exists no $v_j \in \{v_1, \ldots, v_n\}$ such that $\Delta Ut_{v_j}\beta > \Delta Ut_{v_j}\alpha$. In this case, we also say that β is Pareto-inferior to α.

Given that a sum $x_1 + \cdots + x_n$ is bigger than a sum $y_1 + \cdots + y_n$, whenever some x_i is bigger than y_i and no x_j is smaller than y_j, and that a higher realisation of a value is assumed to provide a higher utility, we get the following corollary.

Corollary 4 *(Pareto-superiority entails overall superiority) If α is Pareto-superior to β with regard to a set of values, then α is superior tout court to β with regard to the same set. In other words, if α is Pareto-superior to β with regard to* $\{v_1, \ldots v_n\}$ *then* $\alpha \succ_{\{v_1,\ldots v_n\}} \beta$.

Establishing Pareto-superiority involves comparing utilities resulting from the impacts of the alternative choices on each of value at stake. However, this assessment can be simplified according to Assumption 1 above, namely the idea that the increased realisation of a value always involves, also across different choices, an increased utility resulting from the realisation of that value. Then, to determine Pareto-superiority, it is sufficient to examine the extent to the same values are realised through the alternative choices.

Consider, for instance, that a legislator is discussing whether to raise the length of copyright from the status 70 years (the status quo) to 90 years after the death of the author and assume that the two lengths are equivalent with regard to the incentive to produce new works, but the shorter term contributes more to the value of knowledge. In such a case, we can say that the shorter term is Pareto-superior, and thus superior tout court to the longer one. A legislator's choice which, like this one, is Pareto-inferior to *nil* (to the status quo) is particularly condemnable: it makes things worse in some regards, while providing no advantage in any other regards. Such choices may, however, take place, as a consequence of mistakes in appreciating the social impacts of a new regulation or as a consequence of the fact that the legislators are being pushed by private interests representing no public value.

6 Comparative Evaluations Without Pareto-Superiority

In many cases, however, legislative choices are not Pareto-inferior to the status quo: they promote some value and demote some other values. For instance, a regulation increasing privacy protection may likely decrease freedom of speech, or a regulation increasing environmental protection may decrease productivity and economic freedom.

To evaluate choices having such impacts, we need to find a way of adding up gains and losses, providing a single outcome, on the basis of which to evaluate each choice as a whole. This means that the utilities provided by impacts on distinct values must somehow and subject to elementary arithmetical operations (sum, subtraction, comparison).

Let us assume, as above, that we have an approximate way for assessing the current realisation-quantity of a value v (such as privacy, freedom of speech, welfare, environmental quality, transparency, political freedom), which may or not be expressed in numbers, and a way of assessing the impact of a particular action α on the realisation of v. Given this information, we want to establish the utility-impact of α on v, namely we want to assess the utility-impact of the fact that v's realisation has been increased or decreased by α. And we want to express this utility-assessment in an absolute cardinal quantity, namely a quantity that is homogenous to the quantities through which we express the utility-impacts of this choice on other values at stake, so that these quantities can be added to make up the overall assessment of the utility generated or destroyed by this choice. We want to find a way of accomplishing this task that not only makes sense in principle, but is also psychologically plausible, as a way to perform intuitive quantitative reasoning.

Let us distinguish two steps in the determination of the utility-impact (gain or loss) of a choice on a value. First, we assess the impact of that choice on the realisation of the value, in a way that is independent of any particular unit of measure. Second, we determine the change in the utility that corresponds to a change in the realisation of the value. We intuitively express an assessment of the extent to which a value is affected by a choice, when we say that the choice would provide a (very) big or a (very) small gain or loss concerning the value.

Two different frames of reference seem to be usable for such a judgment. On the one hand, we could quantify increases or decreases as proportions of what the full realisation of that value would be. On the other hand, we could quantify the same increases or decreases as proportions of what is the current realisation level of that value. We use both frames, when using numbers, but also when deploying analogical magnitudes. Thus, we may say that the GDP per head in a poor county has increased a little in absolute terms (viewing the increase as a fraction of what is the GDP per head in the richest countries), but that it has increased a lot relatively its previous level. Similarly, we may say that a liberalisation measure in an authoritarian regime provides a little increase in freedom of the press in absolute terms, but a huge increase relatively to the previous level.

I would argue that in practical situation, an intermediate position can be taken. We assess the level of realisation of a value as a proportion of what might seem the maximum realisation that is concretely available under the existing conditions, within the constraints that we see as unsurpassable (the maximum realisation resulting from actions we view as practicable). As a common-sense example, consider a person who is considering what career to undertake and is considering what kind of revenue and work satisfaction he or she may obtain from different professions. The range of revenue-quantities and satisfaction-quantities the person is considering would probably end at the top of the levels of revenue and satisfaction that person considers to be reasonably achievable.

The same takes place also with regard to public choices, whose impact on the relevant values is to be considered within this feasibility horizon: changes in the GDP of a country would be assessed with reference to the maximal achievable GDP for that country and similarly changes in privacy or freedom of speech.

Thus, an action α's proportional impact on the realisation level of value v could be defined as the proportion between the increase or decrease in the realisation of v brought about by α's and the maximum amount of such realisation that is viewed as realisable by the agent.

Definition 8 (*Proportional impact on the realisation of a value*) The proportional impact of an action α on the realisation of a value v is the proportion between α's realisation impact on v and the reasonably achievable maximum level of v, denoted as $MaxReal_v\alpha$. In other words, $\Delta PropReal_v\alpha = \frac{\Delta Real_v\alpha}{MaxReal_v\alpha}$.

Similarly, we need to define the proportional contribution an action to the utility deriving from the realisation of a value, as a proportion of the utility that can be obtained by the maximal feasible realisation of that value.

Definition 9 (*Proportional impact on the utility by a value*) The proportional impact of an action α on the utility provided by the realisation of value v is the proportion between α's utility-impact on v and the utility provided by the maximal, reasonably achievable, realisation of v, denoted as $MaxUt_v\alpha$. In other words, $\Delta PropUt_v\alpha = \frac{\Delta Ut_v\alpha}{MaxUt_v\alpha}$.

The second step consists connecting a change in the proportional realisation of a value determines to the corresponding change in proportional utility. The relation between the two changes is not constant, since the realisation of a value has decreasing marginal utility: this means that the same change in the realisation of a value will provide less (more) utility the higher (the lower) the position of the realisation interval at issue.

Assumption 2 (*Decreasing marginal utility of the realisation of a value*) A change in the realisation-quantity of value v from quantity q_i to quantity q_j (the difference between q_i and q_j being constant) provides a smaller utility-difference the higher is the position of interval $\left[q_i, q_j\right]$.[8]

[8] In mathematical terms, we would say that the function connecting a value to its utility is such that its second derivative is negative. This too, however, has to be taken as what happens in most of the cases, namely as a defeasible assumption.

Thus, for instance, a proportional loss in the realisation of revenue (or of privacy) of 1/10 determines a higher utility loss if it is the passage from 5/10 to 4/10 than if it is the passage from 9/10 to 8/10.

Corollary 6 *(From decreasing marginal utility) The hypothesis of the decreasing marginal utility has the following implications:*

- *The utility loss resulting from a diminution in the realisation of a value is higher than the utility gain which is provided by an equal increase in the realisation of the same value.*
- *A greater decrease in the realisation of a value causes a proportionally greater decrease in the utility generated by the value; a greater increase in the realisation of a value causes a proportionally smaller increase in the utility by that value*

After establishing the proportional contribution of a choice to the utility provided by a value (note that the approximate magnitudes would be located in a range from 0 to 1, being proportions of the maximum achievable utility), we need to find a way of having homogeneous quantities for the utilities provided by the realisation of different values. For this purpose, we need to assign weighs to values.

Definition 10 (*Weight of a value*) The weight of value v, denoted as w_v, is a quantity expressing the importance of value v relatively to the other values.

Obviously, more important values, such as personal freedom or freedom of speech, will have a higher weight, while less important values, such as privacy or transparency, will have a lower weight. The idea of assigning weights to values may introduce arbitrariness in balancing, due to the difficulty of comparing different values. However, when we engage in such comparisons, we often come to determinations (approximate quantities) that are sufficient to support our choices and even to sharing them. As Sen (2009, 297) observes, the "reasonable variations (or inescapable ambiguities) in the choice of relative weights" do not exclude that a shared assessment, with a sufficient precision, can be made under many circumstances.

We are now in a condition to provide a quantitative characterisation of the absolute utility of an action with regard to a value.

Definition 11 (*Absolute utility-impact on a value*) The absolute utility-impact of action α on value v is the proportional impact of α on the utility concerning v, multiplied by the weight of v. In other words, $\Delta Ut_v\alpha = \Delta PropUt_v\alpha * w_v$.

This allows us to give content to the idea that the utility of a choice is the sum of its impacts on all relevant values at stake. The elements to be summed up consist in the absolute utility-impacts concerning each value, which are obtained by multiplying the proportional utility-impact on that value, for the weight of the value.

Definition 12 (*Utility of an action*) The utility of action α with regard to a set of values $\{v_1,\ldots, v_n\}$ is the sum $i_1 + \cdots + i_n$ of the absolute utility-impacts of α on each of such values. In this sum, each element i_j is the differential utility of α's impact on value v_j multiplied by the weight of v_j. In other words, $\Delta Ut_{\{v_1 \cdots v_n\}}\alpha = \Delta Ut_{v_1}\alpha + \cdots + \Delta Ut_{v_n}\alpha$

By separating positive and negative elements, in the set of the utility-impacts of α*,* we get the notion of outweighing: the positive impacts of α outweigh its negative impacts, if their sum is higher than the sum of the negative elements.

Definition 13 (*Positive impact, negative impact and outweighing,*) The positive impact of action α on value set $\{v_1,\ldots, v_n\}$ is the sum of its impacts on the values whose realisation it increases; α's negative impact on $\{v_1,\ldots, v_n\}$ is the sum of its impacts on the values whose realisation it decreases. In other words, the positive impact can be expressed as: $\Delta PosUt_{\{v_1 \cdots v_n\}}\alpha = \sum_{1 \leq i \leq n | Ut_{v_i} > 0} \Delta Ut_{v_i}$. The negative impact is correspondingly: $\Delta NegUt_{\{v_1 \cdots v_n\}}\alpha = \sum_{1 \leq i \leq n | Ut_{v_i} < 0} \left| \Delta Ut_{v_i} \right|$.

We use positive quantities for negative impacts (given that the absolute value $|-x|$ of a negative number $-x$ is the positive number x), since we want to express the negative impact through a positive quantity, which can be compared with the quantity of the positive impact.

Corollary 6 *(From the notion of outweighing) The following statements are equivalent:*

- *α's utility is larger (smaller) than 0, i.e. $\Delta Ut_{\{v_1,\cdots,v_n\}}\alpha > 0$;*
- *α's positive (negative) utility-impact on values in $\{v_1,\ldots, v_n\}$ is larger than α's negative (positive) utility-impact on values in $\{v_1,\ldots, v_n\}$, i.e. $\Delta PosUt_{\{v_1,\cdots,v_n\}}\alpha > \Delta NegUt_{\{v_1,\cdots,v_n\}}\alpha$;*
- *α's positive (negative) utility-impact on values in $\{v_1,\ldots, v_n\}$ outweighs α's negative (positive) utility-impact on values in $\{v_1,\ldots, v_n\}$;*
- *the proportion between α's positive (negative) utility-impact on values in $\{v_1,\ldots, v_n\}$ and α's negative (positive) utility-impact on values in $\{v_1,\ldots, v_n\}$ is bigger than 1, i.e. $\frac{\Delta PosUt_{\{v_1,\cdots,v_n\}}\alpha}{\Delta NegUt_{\{v_1,\cdots,v_n\}}\alpha} > 1$*

The last item of Corollary 6 in its negative form provides a generalisation of the so-called weight formula proposed by Alexy (2003a, b, 43). In fact, Alexy's formula, which provides the proportion between negative and positive impacts, has the form: $W_{[v_i, v_j]}\alpha = \frac{I_{v_i,\alpha} * W_{v_i}}{I_{v_j,\alpha} * W_{v_j}}$. In our terms $W_{[v_i, v_j]}\alpha$, which Alexy calls the concrete weight of the (demoted) value v_i as opposed to the (promoted) value v_j, in case α, corresponds to the proportion between the negative impact of α on v_i and its positive impact on v_j (which are obtained by multiplying the importance of the impact on the value for its weight), namely to $\frac{|\Delta \mathrm{Ut}_{v_i}\alpha|}{\Delta \mathrm{Ut}_{v_j}\alpha}$, which amounts to $\frac{\left|\Delta \mathrm{PropUt}_{v_i}\alpha\right| * w_{v_i}}{\Delta \mathrm{PropUt}_{v_j}\alpha * w_{v_j}}$. According to Alexy, a choice is wrong when the proportion between its negative impacts and its positive impacts is higher than 1, i.e. when $\frac{\Delta \mathrm{NegUt}_{\{v_1 \cdots v_n\}}\alpha}{\Delta \mathrm{PosUt}_{\{v_1 \cdots v_n\}}\alpha} > 1$.

Finally, we can define the utility of an action α relatively to an alternative action β.

Definition 14 (*Utility of an action relatively to another action*) The utility of action α relatively to action β, with regard to a set of values $\{v_1, \ldots, v_n\}$, is the difference between the absolute utility of α and β with regard to those values. In other words, $\Delta Ut_{\{v_1,\cdots,v_n\}}(\alpha, \beta) = \Delta Ut_{\{v_1,\cdots,v_n\}}\alpha - \Delta Ut_{\{v_1,\cdots,v_n\}}\beta$

This entails that superiority can also be specified on the basis of relative utility.

Corollary 7 *Action α is superior to action β when the utility of α relatively to β is positive. In other words, $\alpha \succ_{\{v_1,\ldots,v_n\}} \beta$ if and only if $\Delta Ut_{\{v_1,\cdots,v_n\}}(\alpha,\beta) > 0$.*

Another interesting corollary is that it may happen that given a set of actions, the action that is superior to all actions in the set is not superior to all of them with regard to any single value.

Corollary 8 *Superiority does not necessarily require maximality with regard to a single value, when at least three choices are compared with regard to at least two values. More precisely, given an option set $\{o_1, o_2, \ldots, o_m\}$ and a value set $\{v_1, \ldots, v_n\}$, it is possible that there is an option $o^* \in \{o_1, o_2, \ldots, o_m\}$ such that $o^* \succ_{\{v_1,\ldots,v_n\}} o_i$ for every $o_i \neq o_k$, but there is no $v_j \in \{v_1, \ldots, v_n\}$ such that for every o_i $o^* \succ_{\{v_i\}} o_i$.*

For instance, given three possible choices $\alpha, \beta, \gamma,$, it maybe the case that γ is superior to both α and β with regard to value set $\{v_1, v_2\}$ while being inferior to α with regard to v_1 and to β with regard to v_2. In this case, γ represents an adequate compromise between the values that are best promoted α and β: (a) γ outweighs α relatively to v_2 more than α outweighs γ relatively v_1 and (b) γ outweighs β relatively to v_1 more than β outweighs γ relatively to v_2.[9] Consider for instance how, with regard to the conflict between privacy and security, the best choice maybe one that maximises neither of the two values, but rather provides a compromise between them. For instance, the intermediate choice of keeping DNA data from suspects only for a short time, with appropriate warranties, may be preferable, all things considered, to both the most privacy favourable option (not storing the data at all) and the most security favourable option (keeping the data indefinitely).

7 Assessing Compliance with Value Norms

We can now deploy the concepts just defined in order to assess compliance with norms dealing with values, or principles, in the terminology of Alexy (2002), e.g., the norms that establish certain constitutional right and collective goals in today's constitutions (see Stone Sweet and Mathews 2008, and the Chapter Balancing, Proportionality and Constitutional Rights by Bongiovanni and Valentini). We shall focus on norms requiring the respect of a value, though the analysis can easily be extended to norms requiring the promotion or the irrelevance of a value.

A norm requiring the respect of a value is satisfied if the agent never chooses a course of action that sacrifices the value, unless the sacrifice is needed for obtaining a more significant increase in the satisfaction of other values. This means that such a norm is violated if the agent makes a choice that demotes the value, and the overall

[9] $\Delta Ut_{\{v_1\}}(\alpha,\gamma) < \Delta Ut_{\{v_2\}}(\gamma,\alpha)$ and $\Delta Ut_{\{v_2\}}(\beta,\gamma) < \Delta Ut_{\{v_2\}}(\gamma,\beta)$.

utility sum—considering all impacts in all relevant values, included the negative impact on the value at issue—is negative. In this sum, we have to include all values whose consideration is prescribed by the legal system, plus those values that have been chosen by the decision maker, to the exclusion of the values whose consideration is prohibited. The weight to be attributed to such values is the weight that is prescribed by the legal system, and for the permissible values chosen by the decision maker, the importance that is given to them by the decision maker, within the boundaries established by the legal system.

This idea can be expressed by the following two conditions. A value norm prescribing the respect of value v is violated by legislative measure α in case that:

- α demotes v, and
- the total utility-impact of α, with regard to all relevant values is negative, relatively to the null action *nil*, or in some cases with regard to an alternative measure β.

Let us see how this idea can be matched with the traditional proportionality texts. According to the reconstruction proposed by Alexy (2003a, b, 135), a legislative norm interfering with rights protected through a constitutional value-norm—a principle, in Alexy's terminology—is only legitimate when it meets the following tests:

- Suitability, which excludes "the adoption of means obstructing the realisation of at least one principle without promoting any principle or goal for which they were adopted";
- Necessity, which requires, with regard to principles P_1 and P_2, "that of two means promoting P_1 that are, broadly speaking, equally suitable, the one that interferes less intensively in P_2 ought to be chosen";
- Balancing in strict sense, which requires that "the greater the degree of non-satisfaction of, or detriment to, one principle, the greater the importance of satisfying the other."

The three tests provide independently necessary and jointly sufficient test for teleological correctness. For instance, as Alexy observes, a legislative norm requiring tobacco producers to place health warnings in their products passed the proportionality test, since the German Constitutional Court considered that (1) this norm served a suitable end, i.e. health, (2) there were no alternative measures achieving that end that would be less interfering upon the economic freedom of tobacco producers; (3) the advantage this measure provided with regard to health outweighed the minor interference it caused on economic freedom.

Let us specify the three tests using the concept introduced above, starting with choices affecting only two values: the goal value v_g pursued by the agent and the prescribed value v_p to be respected according to a value-norm.

- Suitable choice. A choice α is suitable if it has a positive realisation impact on a permissible goal value v_g. In other words, α is suitable iff $\alpha \succcurlyeq_{v_g} nil$.
- Necessary choice. A choice α having a negative impact on a prescribed value v_p is necessary if it has a positive impact on a permissible goal value v_g and there exists no alternative choice β, having a non-inferior impact of on the goal value v_g and

a better impact on the prescribed goal v_p. In other words, choice α is necessary iff $\alpha \succ_{v_g} nil$ and there exist no β such that $\beta \succcurlyeq_{v_g} \alpha$ and $\beta \succ_{v_p} \alpha$.
- Balanced choice (first definition). A choice α having a negative impact on a prescribed value v_p and a positive impact on a goal value v_g is balanced in a strict sense if the positive utility-impact on the goal value v_g is not outweighed by the negative utility-impact on the prescribed value v_p. In other words, choice α is balanced in a strict sense iff implementing it is better than not doing anything: $\alpha \succcurlyeq_{\{v_g, v_p\}} nil$.
- Balanced choice (second definition). A choice α having negative impact on a prescribed value v_p and a positive impact on the goal value v_g is balanced if there exists no alternative β such that β would have a smaller negative impact on the prescribed value v_p, and would at the same time provide an higher overall utility—the overall utility being the difference between the utility of the impact on v_g and the disutility provided by the negative impact on the v_p. In other words, choice α is balanced iff there exists no β such that $\beta \succ_{\{v_p\}} \alpha$ and $\beta \succ_{\{v_p, v_g\}} \alpha$.

I have distinguished two notions of a balanced choice, because the second one provides a much stricter standard then the first: It requires that the decision having a negative impact on a prescribed value should provide a superior overall utility to any possible decision being preferable relatively to the prescribed value. If brought to the extreme, it would almost completely undercut the possibility for a decision maker to adopt a decision that has a negative impact on a prescribed value and escape censorship. The reviewer would be free to imagine possible alternative decisions which the decision maker did not consider, speculate on their possible effects and merits, and to condemn the decision maker as soon as the latter's decision could be shown to be suboptimal. Thus, I believe, this kind of review needs to be strongly constrained, for instance, by requiring that the adoption of the chosen alternative α appears to have been an unreasonable mistake, given the evidence available when α was adopted.

By denying the conditions above, we get three conditions under which a choice infringes a value-norm.

- Unsuitable choice. A choice α having a negative impact on a prescribed value is unsuitable if it has no a positive impact on a permissible goal value. Thus, the unsuitable choice is Pareto-inferior to the status quo, the null action nil. This may depend on the fact that the chosen action is incapable of reaching that goal—its adoption is based on mistaken factual assumptions—or on the fact that the pursued goal is impermissible and thus irrelevant according to a value-norm.
- Unnecessary choice. A choice α having a negative impact on a prescribed value is unnecessary if there exists an alternative choice β, which is better than α with regard to the prescribed value and a non-inferior with regard to the goal value. Thus, the unnecessary α is Pareto-inferior to the alternative β.
- Unbalanced choice (first definition). A choice α having a negative impact on a prescribed value is unbalanced if the positive utility of its impact on the goal value is outweighed by the disutility of its impact on the prescribed value

- Unbalanced choice (second definition). A choice α having a negative impact on a prescribed value and a positive impact on the goal value is unbalanced if there exists an alternative β such that β has a smaller negative impact on the prescribed value and provides an higher overall utility than α.

We can generalise these notions to the case of decisions affecting more than two values and introduce further refinements and specifications of the notions just introduced. For instance, a different notion of necessity is needed for covering cases where a choice is qualified as "necessary" even though it has a small negative impact on the goal value, while having a much more significant negative impact on the value to be respected. This will be left to further research.

As these definitions should have been made clear, what is at issue in a proportionality assessment concerning a decision α affecting values v_1 and v_2 is not a comparison of the weights of v_1 and v_2, but rather a comparison of α's impacts on such values. Consequently, the fact that value v_1 is more important, i.e. has a higher weight than value v_2, does not necessarily entail than that α's utility-impact on v_1 is larger than its utility-impact on v_2: The utility-impact on a value depends on both (1) the proportional utility-impact on that value—the extent to which the benefit deriving from the realisation of the value is increased or decreased—and (2) the weight the value.[10] This is affirmed with particular clarity by the Israeli judge Aharon Barak:

> [T]he comparison is not between the advantages gained by realizing the goal in contrast to the effect brought by limiting the right. Nor is it between security and liberty. The comparison is between the marginal benefit to security and the marginal harm to the right caused by the restricting law and as such, the comparison is concerned with the marginal and the incremental (Barak 2010, 8).

Thus, it may happen that in a certain case, the impact of a measure α on value v_1 outweighs α's impact on v_2, while in another case, the impact of a different measure β on v_2 outweighs β's impact on v_1. To explain this, it is not necessary to assume that the weights of v_1 and v_2 have changed, being "context dependent." A more plausible explanation may be provided by the fact that the proportional impacts on v_1 and on v_2 were different in the two cases, the weights remaining the same, namely that in the first case v_1 was affected by α more than it was affected by β in the second case, or that v_2 was affected by β in the second case more than it was affected by α in the first case.

[10] More exactly, in the terminology use used before, the absolute utility-impact of a decision α on two values is the sum of its utility-impacts on each of them, where each utility-impact is the result of the proportional impact on a value for the weight of that value: $\Delta\mathrm{Ut}_{\{v_1,v_2\}}\alpha = \Delta\mathrm{PropUt}_{v_1}\alpha * w_{v_1} + \Delta\mathrm{PropUt}_{v_2}\alpha * w_{v_2}$. Thus, the condition for the first term (in hypothesis, the negative impact) to be higher than the second is that $\left|\Delta\mathrm{PropUt}_{v_1}\alpha * w_{v_1}\right| > \Delta\mathrm{PropUt}_{v_2}\alpha * w_{v_2}$ which is equivalent to $\frac{w_{v_1}}{w_{v_2}} > \frac{\Delta\mathrm{PropUt}_{v_2}\alpha}{\left|\Delta\mathrm{PropUt}_{v_1}\alpha\right|}$. This inequality can be falsified even when the weight w_{v_1} of value v_1 is much larger than the weight w_{v_2} of v_2. This happens when the proportion between $\Delta\mathrm{PropUt}_{v_2}\alpha$ and $\Delta\mathrm{PropUt}_{v_1}\alpha$ is larger than the proportion between w_{v_1} and w_{v_2}.

8 Teleological Reasoning and the Choice of Rules

The evaluation of value-impacts pertains not only to the adoption of individual decisions but also to the adoption of general rules.

A value-based choice of rules takes place in the teleological interpretation of legislative texts. Teleology requires choosing the interpretation that most realises the legislator's goals and the legal values at stake. Within constitutional review, a similar reasoning pattern is used in the so-called definitional balancing: a court does not only affirm that a certain law is disproportionate, but explains this statement considering that any law having a certain kind of content would be disproportionate and would therefore violate the constitution (Aleinikoff 1987; Alexy 2002, 80ff). It seems that two teleological arguments are involved in this reasoning—first, the teleological assessment of a specific legislative choice, according to its impacts on the values at issue in the individual case, and second, the teleological choice of a rule that generalises the outcome of the case.

Let me give you an example to explain how a court may engage in definitional balancing or on the contrary refrain from it. The European Court of Human Rights has recently addressed a case concerning a man and a woman who were both unaffected carriers of mucoviscidosis, a very serious genetic disease, and thus had a high risk (1/4) of having children affected by this illness (Costa and Pavan v. Italy, application no. 54270/10). The claimants, who already had generated an affected child, attacked an Italian law which prohibited pre–implantation tests (Law 40/2007). They argued that this law impeded them from avoiding the risk of having an affected child, through a medical procedure involving the in vitro production of embryos and the implantation of a non-affected one. The Court affirmed that this law violates art. 8 of the Charter of Human Rights by disproportionately affecting the right to private life of the claimants, in comparison with its alleged benefits to other interests at stake, such as protecting the life of embryos and preventing eugenic practices.

In this case, the judges did not state any general rule to explain or justify their decision, even though they had many possible rules available to them. Such possible rules include the following ones: (1) any prohibition of pre–implantation testing for mucoviscidosis is disproportionate, with regard to couples who have already generated an affected child, (2) any prohibition of pre-implantation testing for mucoviscidosis is disproportionate for carriers of the disease, including those not having already generated an affected child, (3) any prohibition of pre-implantation testing for any genetic disease is disproportionate, even when concerning genetic problems different from mucoviscidosis, (4) any prohibition of pre-implantation testing is disproportionate, even when the test serves non-therapeutic purposes, such as sex selection, (5) any prohibition of pre-implantation interventions is disproportionate, even when the intervention goes beyond mere testing, as for cloning or genetic engineering.

The adoption of any one of these "definitional" rules by the competent court would have enabled subsequent judges to decide similar cases through rule-based reasoning, rather than through balancing. The subsequent judge, if a definitional rule had been established, could then simply check whether a prohibition of pre-implantation testing has the properties which would make it disproportionate according to the rule and decide the case accordingly.

Since disproportionateness of a legislative measure entails that it should not be taken, such definitional rules could be re-expressed as prohibitions of adopting laws having the indicated content. For instance, assume that the judges in this case, rather than being silent, had ruled any prohibition of pre-implantation testing for genetic diseases is disproportionate. This ruling, in combination with the fact that legislator should not adopt any disproportionate law, entails the following rule: "the legislator should not adopt any law which prohibits pre-implantation testing for genetic diseases."[11]

It seems that the judges—having established that a law produces a disproportionate outcome in a particular case—face a choice concerning how to frame their opinion. Their options include choosing one of many available rule-based explanations or choosing not to provide any such explanations. On what grounds should they make such a choice? The answer, I think, requires another appeal to teleological reasoning. They should adopt the generalisation that could best realise the legal values at issue, through the subsequent application of the generalisation by judges, legislators and citizens, in the given institutional framework—as characterised by applicable norms, judicial powers, legislative competences, interpretive practices, existing precedents and social norms. However, on the basis of the same teleological reasoning, judges could also conclude that perhaps stating no rule is the best solution, for instance, given the high uncertainty of the matter at stake, which prevents them from stating with sufficient confidence even a highly defeasible general rule.

Thus, the proportionality review of a legislative measure may involve two teleological assessments.

The first assessment concerns establishing that the legislative decision α in the current case has a negative overall impact on the values at stake and is therefore disproportionate. In our example, this is the assessment that the prohibition of pre-implantation testing in the case of a man and a woman both carrier of mucoviscidosis, who already had an affected child, has a disproportionate negative impact on their private life.

The second assessment involves two steps. The first step consists in developing, through abductive reasoning, a set of possible explanations as to why α's utility-impact is negative in that case, each explanation appealing to a (different) ruling stating that legislative actions having certain features are disproportionate. The second step consists in selecting, for the rule-based justification of the decision, the rule whose future adoption by courts, legislators and citizens, in the given institutional and socio-economic context, is likely to provide the highest utility-impact or choosing to provide no rule-based justification in case no advantageous rule can be found.

[11] This is an instance of logical inference according to which premises $A \rightarrow B$ and $B \rightarrow C$ entail $A \rightarrow C$ (instantiating A with "the law prohibits pre-implantation tests for genetic diseases," B with "the law is disproportionate" and C with "the law is forbidden"). This inference may be assumed to hold, though only defeasibly, also for defeasible conditionals.

Thus, this second teleological assessment takes place at a meta-level, since it concerns the choice between alternative patterns for decision in future cases. It compares the utility-impacts provided by the future application of different rules and also the utility-impact of not having any such rule (thus entrusting future decisions to case-by-case assessments).

The formulation of rules based on proportionality assessments is also at the basis of the attempts to specify the essential content, or the core, of a right. The essential core of a right is indeed identified by indefeasible prohibitions and obligations of performing certain actions, for the sake of the right at issue. Whether a certain rule protecting a right could be viewed as defeasible or rather indefeasible can also be established through teleological reasoning. It should be preferable that the rule is viewed as indefeasible—i.e. as concerning the essential core of the right at issue—when the following conditions hold: (1) the action it prohibits, e.g., torture, has a negative impact on a value, this impact being so significant that it is very unlikely that it will be compensated by positive impacts on other values, and (2) the costs of mistaken exceptions to the rule, i.e. of failing to apply it when its application would provide a higher benefit, are presumably smaller-than the costs of its overreaching applications, i.e. of applying it when its non–application would be more beneficial. Under such conditions, we should indeed accept the view that the rule should be treated as being indefeasible and we should be ready to pay the costs of the rare disutility of its counterproductive application in extreme cases for the benefit of preventing the disutility of its counterproductive disapplication in a larger or more important set of cases. In other terms, proportionality analysis would not be used at the stage of the application of the rule, but rather at the stage of the justification of its indefeasibility.

9 Consistency in Balancing

The idea that quantitative reasoning with non-numerical magnitudes has a valuable function in the application of the law can be challenged by pointing to the arbitrariness of the inputs of such a reasoning: even though balancing is constrained by arithmetic, it operates on magnitudes that are idiosyncratic contents of the minds of individual decision makers (or reviewers of their decisions). How can there be convergence in the outcomes of such reasoning, and how can any social control over such outcomes be effective if any outcome would be possible by changing subjective input quantities?

A certain degree of convergence may be explained by our natural inclinations, social environment and cognitive capacity. We learn the magnitudes that are associated to our values (the proportional utilities of their realisation and their weights) by processing the inputs we get from our inborn attitudes, education and personal experience, possibly though inductive/abductive patterns of reasoning which deliver both adjustments and explanations of our intuitive assessments. Moreover, we may consider what reasons support or attack the quantitative assessments we are inclined to make, and thus subject them to critical review, though monological or dialogical argumentation. When we are assumed to adopt a single and shared point of view (the legal point of view), our assessments are additionally constrained by the need that they fit with the past expressions—value-norms contained in constitutions and legislative documents, other explicit statements on the absolute and comparative importance of impacts and values, decisions of individual cases involving impacts on such values, legislative rules addressing value conflicts—of the point of view we are adopting.

I cannot here provide an analysis of how we can determine the measure of fit of a new assessment with a certain past history of teleological reasoning, which may involve also incompatible decisions, giving conflicting clues.[12]

Let us first mention two basic cases, where reasoning *a fortiori* on the basis of previous assessments may give clear indications. Assume that α, involving a demotion of value v_d and the promotion of value v_p, was assessed as being proportionate (v_d and v_p, being the only values at stake). Now consider a new decision β involving a smaller (in absolute number) demotion of v_d and an equal or greater promotion of v_p: clearly, β must be considered proportionate as well.

Assume on the contrary that choice α was assessed as non-proportionate. Consider a decision β involving a greater demotion of v_p and a smaller or equal promotion of v_p: clearly, β must be considered disproportionate as well.

These ideas can be further generalised, as this simple example will show. Assume that I have to assess the proportionality of the choice β to store DNA samples of all citizens for 20 years, which demotes their privacy (v_1) and promotes their security (v_2). Assume that in the past, a choice α which involved storing DNA samples of all citizens accused of a crime for 10 years was considered to be unacceptable, since its negative impact on privacy outweighed the gain in security. Assume also that it is agreed that by increasing the conservation time (by adopting β rather than α), the damage to privacy is proportionally increased to a larger extent than the gain in security: $\frac{\Delta\text{PropUt}_{v_1}\beta}{\Delta\text{PropUt}_{v_1}\alpha} > \frac{\Delta\text{PropUt}_{v_2}\beta}{\Delta\text{PropUt}_{v_2}\alpha}$. Given such premises, any assessment according to which the new law would provide a positive balance (by subtracting losses and adding gains) would be inconsistent with the previous decision. In fact, in our example, any assignments of weights to v_1 and v_2 that would satisfactorily explain the disproportionality of the 10-year term would also determine the disproportionality a longer term.

10 Conclusions

I have argued that teleological reasoning includes the assessment of impacts upon relevant values, which may be viewed as a kind of approximate quantitative reasoning, even when we are unable to assign symbolic numerals to the concerned magnitudes. We engage in this reasoning both when making common-sense private choices and

[12] On the idea of fit, see Dworkin (1986, Ch. 7). On the connection between value-based reasoning and the interpretation of rules or the determination of their priorities, see in particular Bench-Capon and Sartor (2003, 97–142), Prakken (2000, 49–57).

when participating in public decision-making. Non-numerical quantitative reasoning involves certain rationality conditions, and first of all it should normally respect the usual arithmetical relationships, which indicate general constraints for processing of quantitative information. Thus, arithmetical relationships can also be viewed as default standards of rationality to be applied by legal reasoners (legislators, interpreters and judges) when engaging in proportionality assessments.

These quantitative assessments express intuitive appreciations of the importance of positive or negative impacts on the values at stake, but can be supported or attacked through arguments. These arguments may address all legally relevant aspects at stake (the identification of values at stake, the determination of impacts on their realisation, the assessment of their weights, etc.), and they can appeal to consistency with precedents. Thus, intuitive quantitative assessments are subject to some degree of discursive control.

References

Aleinikoff, T. 1987. Constitutional law in the age of balancing. *Yale Law Journal* 96: 943–1005.

Alexy, R. 2002. *A theory of constitutional rights*. Oxford University Press.

Alexy, R. 2003a. On balancing and subsumption: A structural comparison. *Ratio Juris* 16: 33–49.

Alexy, R. 2003b. Constitutional rights, balancing, and rationality. *Ratio Juris* 16: 131–140.

Barak, A. 2010. Human rights and their limitations: The role of proportionality. *Law and Ethics of Human Rights* 4: 1–18.

Bench-Capon, T.J.M., and G. Sartor. 2003. A model of legal reasoning with cases incorporating theories and values. *Artificial Intelligence* 150: 97–142.

Dworkin, R.M. 1986. *Law's empire*. London: Kermode.

Gallistel, C.R., R. Gelman, and S. Cordes. 2006. The cultural and evolutionary history of the real numbers. In *Evolution and culture*, ed. S. Levinson, and P. Jaisson, 247–274. Cambridge, MA: MIT Press.

Gallistel, C.R., and R. Gelman. 2005. Mathematical Cognition. In *The Cambridge handbook of thinking and reasoning*, ed. K.J. Holyoak, and R.G. Morrison. Cambridge: Cambridge University Press.

Kahneman, D. 2011. *Thinking: Fast and slow*. London: Allen Lane.

Keeney, R., and H. Raiffa. 1993. *Decisions with multiple objectives: Preferences and trade offs*. Cambridge, MA: Cambridge University Press.

Parsons, S. 2001. *Qualitative methods for reasoning under uncertainty*. Cambridge, MA: MIT Press.

Pollock, J.L. 2006. *Thinking about acting: Logical foundations for rational decision making*. Oxford: Oxford University Press.

Prakken, H. 2000. An exercise in formalising teleological case-based reasoning. In *Proceedings of the thirteenth annual conference on legal knowledge and information systems (JURIX)*, ed. H. Prakken, and R. Winkels. Amsterdam: IOS, 49–57.

Rescher, N. 1991. *G.W. Leibniz's monadology: An edition for students*. Pittsburgh: University of Pittsburgh Press.

Sartor, G. 2010. Doing justice to rights and values: Teleological reasoning and proportionality. *Artificial Intelligence and Law* 18: 175–215.

Sartor, G. 2013. The logic of proportionality: Reasoning with non-numerical magnitudes. *German Law Journal* 14: 1419–1457.

Sen, A. 2009. *The idea of justice*. Belknap: Cambridge, MA.

Stone Sweet, A., and J. Mathews. 2008. Proportionality balancing and global constitutionalism. *Columbia Journal of Transnational Law* 47: 72–164.

Coherence and Systematization in Law

Amalia Amaya

1 Introduction

Coherence theories of law and adjudication have occupied a prominent position in legal theory in the last decades.[1] This is in tune with the emergence of coherentism across a number of different domains. Coherentism has been proposed as an alternative to the foundationalist account of the structure of knowledge and justification (BonJour 1985; Lehrer 2000). Coherence is also a main topic in the new, emerging, field of formal epistemology (Olsson 1997, 1998). In philosophy of science, explanatory coherentism has been advocated as a main alternative to the dominant Bayesian approach to theory choice (Thagard 1989). Coherentist views of the nature of truth are far more controversial than coherentist approaches to the justification of empirical and scientific beliefs. However, in the last decades, new forms of the coherence theory of truth have been developed and this theory is still advanced as a main competitor to the traditional view of truth as correspondence (Walker 1989; Alcoff 2001; Young 2001). In the domain of practical, rather than theoretical, reason coherence also plays a prevalent role. Important accounts of practical deliberation give coherence a place of privilege (Hurley 1989; Richardson 1994; Millgram and Thagard 1996), and moral reasoning is widely regarded as a coherentist kind of reasoning (Rawls 1999; Goldman 1988; DePaul 1993). Coherence features not only in philosophical approaches to reasoning and rationality but also in psychological approaches to the subject. Cognitive psychologists and linguists have used the concept of coherence to

[1] See, among others, Dworkin (1986), MacCormick (1984, 2005), Alexy and Peczenik (1990), Peczenik (2009), Peczenik and Hage (2004), Günther (1989, 1993), Wintgens (1993, 2000), Hage (2004, 2013). For some earlier coherence theories of legal justification, see Sartorious (1968, 1971), and Hoffmaster (1980).

A. Amaya (✉)
Instituto de Investigaciones Filosóficas, Universidad Nacional Autónoma de México,
Mexico City, Mexico
e-mail: amaya@filosoficas.unam.mx

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_22

give an account of problems as diverse as impression formation, decision-making, discourse processing, and analogical mapping (Hellman 1995; Simon et al. 2001).[2]

Coherence has been appealed to in legal theory with a view to advancing two different projects: the development of a theory of law and the development of a theory of legal reasoning. Thus, two main perspectives on coherentism in law may be distinguished: a systemic one, according to which coherence is a feature of the legal system, and an argumentative one, according to which the coherence of a particular ruling or interpretation serves as a justification of the ruling or interpretation in question. In this chapter, I shall mainly focus on normative coherence as a standard of justification, that is to say, I will analyze coherence from an argumentative, rather than a systemic, perspective.[3] Another important restriction on the scope of this chapter should be stated at the outset. Coherence has been claimed to be an important criterion for the justification not only of normative judgments in law but also of evidentiary statements in law. Here, however, I will focus exclusively on theories of normative coherence.[4]

The structure of this chapter is as follows. In the first section, I provide a survey of the main approaches to normative coherence defended in the literature on legal coherentism. In the second section, I discuss the principal objections that threaten to undermine the coherence theory of legal justification. One problem with coherentism, namely the problem of the coherence bias, has not, however, received enough attention in the literature. I state this problem in detail in the third section and argue that a modified version of legal coherentism—namely virtue coherentism—has the resources to address this problem. The fourth section engages in a second-order debate about the relevance of coherence in justification by inquiring into the reasons why coherence is worth pursuing when reasoning in law. I conclude this chapter by assessing the value and limits of coherentist reasoning in the legal domain.

2 Theories of Normative Coherence

Three main kinds of theories of normative coherence have been advanced in the literature: principle-based theories, case-based theories, and constraint-satisfaction theories. These three versions of legal coherentism take a different stance on the issue of how coherence emerges over the course of legal decision-making. Principle-based theories take a top-down perspective on coherence: Coherence, on this view,

[2] A summary of the main coherence theories across a number of domains is provided in Amaya (2015, Chaps. 3 through 9).

[3] For analyses of coherence from a systemic perspective, see Alonso (2006), Pastore (1991), Pérez-Bermejo (2006, 2007), Ratti (2007), and Schröter (2006). On the distinction between systemic and argumentative coherence, see Moral (2003) and Bertea 'The Argument from Coherence,' *IVR Encyclopaedia of Jurisprudence, Legal Theory and Philosophy of Law*, available at http://www.ive-enc.

[4] A discussion of evidential reasoning in law is provided by Di Bello and Verheji, Chap. 1, part III, this volume, on 'Evidential Reasoning.'

is mainly a matter of principle. A given set of elements is coherent if it can be brought together under some unifying principles. Case-based theories, to the contrary, endorse a bottom-up approach and claim that coherence is built mainly by establishing relations of analogy across cases.[5] According to constraint-satisfaction theories, coherence is a matter of maximizing satisfaction of a number of positive and negative constraints among the relevant elements.

2.1 Principle-Based Theories

Principle-based theories are the most popular version of coherentism among legal theorists. Three theories are particularly influential, to wit, MacCormick's theory of normative coherence, Peczenik's weigh-and-balance model of coherence, and Dworkin's theory of law as integrity. According to MacCormick (1984, 1993, 1994, 2005), normative coherence is the property of a set of norms that is explained by general principles which delineate a satisfactory form of life. Normative coherence, MacCormick claims, plays an important but limited role in the justification of decisions in hard cases. Coherence arguments allow us to determine a set of 'justifiable' decisions, but it is consequentialist arguments that are the final clincher of justification. In this sense, MacCormick's coherence theory is a 'weak' theory, in that coherence is claimed to be a necessary but not sufficient condition of justification.

In Peczenik's view (1990, 1994, 1998, 2004), coherence is achieved in the course of legal decision-making by means of an operation of weigh and balance. Legal justification, he claims, results from the coherent weighing of a set of relevant reasons, more specifically, legal and moral reasons. Peczenik's theory of coherence is thus a 'strong' theory, insofar as it takes coherence to be a necessary and sufficient condition of justification. Nonetheless, given that, in Peczenik's view, the last step of weighing is ultimately based on personal and intuitive preferences, the potential of coherence to generate justification is significantly limited within this theory.

Last, Dworkin holds a (strong) coherence theory of justification according to which a legal decision is justified if it coheres with a structure of principle that best fits and justifies the legal practice in light of background political theory. Dworkin's interpretative theory of law also involves a commitment to a coherence theory of legal truth according to which whether a particular proposition in law is true depends on whether it belongs to the most coherent theory that fits and justifies the settled law. Thus, in Dworkin's view (1977, 1985, 1986, 1996), coherence with a set of interpretative beliefs about law and political morality yields both justification and legal truth.

[5] The difference is not a sharp one, as principles and analogies are very much related; there is, nonetheless, a difference in emphasis that is worth noting and that makes this classification useful.

2.2 Case-Based Theories

Some coherence theories of practical reasoning endorse a case-based approach to coherence-building in the course of decision-making. Two such theories, Goldman's and Hurley's, explicitly aim to provide an account of legal reasoning as well. According to Goldman, the correct answer to a controversial moral or legal issue is that which is most coherent with a base of settled judgments (Goldman 1988, 1989, 2002). Coherence is a matter of satisfying what Goldman calls 'the Kantian constraint.' This constraint requires that we do not judge cases differently without being able to cite a relevant and generalizable difference between them. Hence, in Goldman's view, legal reasoning consists in reasoning from analogy and difference under the Kantian constraint. When this constraint is satisfied, our moral judgments are coherent and therefore justified.

Hurley (1989) advances an account of case-based deliberation that gives coherence a central role. According to Hurley, deliberation is first and foremost a process whereby one builds a theory that best displays as coherent the relationships among the several values that apply in the particular case. The fundamental claim of Hurley's coherentist account of legal reasoning is that there is a conceptual relation between the reasons that are relevant in a specific case and judgments about what to do all-things-considered. More specifically, the relationship in question is that of subject matter to theory. That is to say, a judgment about what to do all-things-considered is right if it is favored by the theory that gives the most coherent account of the relationship among the specific reasons (such as moral values, legal doctrines, and precedents) that are relevant in the particular case.

2.3 Constraint-Satisfaction Theories

Constraint satisfaction theories aim to model legal coherence by applying the general framework of coherence as constraint satisfaction developed by Paul Thagard. According to Thagard (2000), the coherence of a set of elements is a matter of the satisfaction of a number of positive and negative constraints. These constraints establish relations of coherence—positive constraints—and incoherence—negative constraints—among the elements of a set. A coherence problem consists in dividing a set of elements into accepted and rejected in a way that maximizes the satisfaction of the constraints. A positive constraint between two elements can be satisfied either by accepting both of the elements or by rejecting both of them. A negative constraint between two elements can be satisfied only by accepting one element and rejecting the other. Thus, the idea is that we turn a set of elements into as coherent a whole as possible by taking into account the coherence and incoherence relations that hold between pairs of elements of this set. Thagard has developed a legal application of this theory exclusively in the context of legal fact-finding (Thagard 1989, 2006). However, several modifications and extensions of this theory have been proposed in legal scholarship to give an account of normative reasoning in law in terms of constraint satisfaction.[6]

In these three versions of legal coherentism, coherence, variously defined, features in the analysis of justification (as a necessary or a necessary and sufficient condition of justification). However, it is important to emphasize that coherence may also contribute to a theory of legal justification without it being a condition of justification. Coherence may be given an enhancing role—that is to say, it may be claimed that coherence augments the degree of justification enjoyed by a normative proposition even if it does not generate such justification. Coherence may also be assigned a negative, rather than a positive, role, i.e., incoherence defeats justification (Audi 1988, 1993). Thus, theories of legal justification can assign a role to coherence without committing themselves to the thesis, which characterizes coherentism, that coherence is a source of justification.

3 Objections to Legal Coherentism

Coherentism is as popular an account of justification as it is controversial. A number of objections have been traditionally raised against coherence theories of justification. I proceed now to examine the main objections that may be directed against coherence theories of law and adjudication.[7]

3.1 The Vagueness Problem

Coherence theories of justification take justification to be a matter of coherence. However, the notion of coherence is notoriously difficult to define, and this sheds doubts over whether the coherence theory is in a position to provide a usable standard of legal justification. The coherence theory needs to specify in detail when a legal decision is coherent with a body of legal rules and standards, and thus, justified, as well as the conditions of coherence that a set of norms must satisfy for it to be justification-conferring.

The objection seems to be appropriately raised against some versions of legal coherentism. Although MacCormick's and Dworkin's principle-based approaches to coherence identify some of the core ingredients of coherence, they do not provide a detailed conception of normative coherence. An important effort at precisely

[6] See Bench-Capon and Sartor (2001b), Joseph and Prakken (2009), Amaya (2011, 2015), Araszkiewicz (2010, 2012, 2013), Araszkiewicz and Sãvelka (2012), and Sãvelka (2013).

[7] There is another objection that has been raised in the coherentist literature but that I shall not deal with here, to wit, 'the retroactivity objection.' The objection is that coherence theories are path-dependent and that this leads to retroactive application of the law. See Kress 1984, 377–388. For responses to Kress's retroactivity objection, see Hurley (1990, 235–251) and Habermas (1996, 219–220).

determining what coherence requires is Alexy and Peczenik's list of criteria of coherence (Alexy and Peczenik 1990). According to Alexy and Peczenik, the degree of coherence of a theory is a matter of the degree of perfection of its supportive structure, which depends, in its turn, on the extent to which a number of criteria of coherence are fulfilled, such as the number and strength of the supportive relations between the statements belonging to a theory, the generality of the concepts applied by the theory and their conceptual connections with those used by other theories, or the number and variety of cases covered by the theory. This proposal does much in the way of identifying the different ingredients of coherence; however, it remains silent as to how the different criteria of coherence should be balanced against each other when they come into conflict.

Constraint-satisfaction approaches significantly mitigate the force of the objection from vagueness in that they provide a semi-formal account of the criteria of coherence against which the coherence of a given ruling may be evaluated. Normative coherence, on one such view, requires the integrated assessment of explanatory, analogical, conceptual, interpretative, and deliberative coherence.[8] Each kind of coherence is further specified by a number of principles. For instance, according to the principles of interpretative coherence, coherence is a symmetrical relation, positive constraints (relations of coherence) arise from analogy and explanation and negative constraints (relations of incoherence) arise from relations of incompatibility and contradiction, priority is given to propositions that describe normative elements (i.e., precedents, principles), and the acceptance of any proposition depends on its coherence with the rest of propositions of the system (Amaya 2015). More work, however, needs to be done to explicate how the different kinds of coherence relate to each other in order to give a solution to a problem of justification.

Formal models of coherence have also been recently developed. These approaches use a variety of formal theories to articulate a precise conception of coherence. Some of these approaches have been deployed to analyze the kind of coherence that is relevant to the justification of normative statements in law.[9] Although these approaches are necessarily limited, as not all aspects of coherence seem to be amenable to formal analysis, they are extremely helpful in rendering precise some central aspects of the notion of normative coherence and thereby in meeting the objection from vagueness.

3.2 What Is Coherence-Driven Inference?

Coherence theories in law are vague in another respect as well. These theories do not specify the process of reasoning whereby one may arrive at the legal decision that is best justified. If all the coherence theory has to say about how to reason in law is

[8] This contrasts with case-based approaches to coherence, which reduce normative coherence to analogical one.

[9] For a discussion of formal approaches to coherence that use belief revision formalisms, see Amaya (2007); for formal approaches to normative coherence within the field of A.I., see Bench-Capon and Sartor (2001a, b, 2003), Hage (2001), Joseph and Prakken (2009); for probabilistic analyses of coherence, see Shogenji (1999) and Olsson (2002, 2005).

that one should search for the most coherent decision—but it provides no guidelines as to how one should do so—then it amounts to little more than a recommendation to pick the decision that seems intuitively best. Hence, it is imperative for coherence theories that they give a full account of the reasoning patterns whereby coherence may be established over the course of legal decision-making.

A promising way, I would argue, to specify the patterns of reasoning whereby coherence may be constructed over the course of legal decision-making appeals to the model of inference to the best explanation. Inference to the best explanation is a pattern of reasoning whereby explanatory hypotheses are formed and evaluated. Some proponents of models of inference to the best explanation have developed coherentist interpretations of inference to the best explanation according to which explanatory inference aims at maximizing coherence (Psillos 2002; Harman 1980, 1986; Lycan 1988). Thus, we may describe coherence-driven inference in law as a kind of explanatory inference and use the resources of inference to the best explanation models to unfold the structure of coherence-based legal inference.

Now, what is the structure of explanatory inference? Lipton (2004, 148–151) has elaborated a detailed account of the structure of inference to the best explanation according to which the mechanism we use to settle on which explanation to infer has two stages, namely a process of generation and a process of selection. Lipton's idea is that we infer to the best explanation by first generating a short list of plausible candidates, and then by selecting the best from that list. We never consider the set of all possible explanations, for it would be too large to generate and handle. Instead, he argues, we use some sort of short-list mechanism whereby we generate a small number of live candidates, from which we then choose. Hence, inference to the best explanation includes 'two filters,' one that selects the plausible candidates and another that selects from among them. There is, however, I would argue, an intermediate stage that is extremely important, to wit, a 'context of pursuit,' in which working hypotheses are subjected to a preliminary assessment and developed in further detail.[10]

More specifically, as argued elsewhere (Amaya 2015, 512), coherence-driven inference in law may be described as a kind of explanatory inference that has the following structure: 1. The specification of a base of coherence, that is, the set of normative elements that provides the input to coherentist reasoning; 2. the construction of a contrast set that contains a number of relevant alternative theories about what the law requires from which the most coherent one is to be selected; 3. the pursuit and refinement of these alternative theories by means of a number of coherence-making mechanisms; 4. The evaluation of these theories against the criteria of normative coherence; and 5. The selection as justified of the theory that best satisfies the criteria of normative coherence.

[10] The label 'context of pursuit' is Laudan's (1977, 110). On this context, see Sintonen and Kikeri (2004, 214–218).

3.3 The Circularity Objection

A problem with a coherence-based view of legal inference, it is claimed, is that it involves a vicious circularity. Coherence theories seemingly license an inference from *p* to *q* on grounds of coherence, and then from *q* to *p*. If nothing but coherence generates justification, then any chain of arguments that one may construct would eventually bite its tail. Following a coherentist approach, we may justify, for instance, a particular principle because it properly coheres with a body of precedent, and eventually justify a decision because it coheres with such a body of precedent, which we accepted as justified by virtue of its coherence with this very same principle.

The problem of circularity, however, only arises if one accepts a linear account of the structure of justification according to which justification involves a chain of beliefs along which justification is transferred from one element to another down the chain. However, as soon as one replaces (as coherence theories do) this linear conception of justification by a holistic one, i.e., the view that justification is a property of a coherent set of elements each of which is justified by virtue of its belonging to such a set, the circularity involved may be shown to be benign (BonJour 1985, 89–92). Nonetheless, the justification-conferring set must satisfy some additional constraints to meet the circularity objection. That it be sufficiently big is one common consideration. I would like to suggest, though, that it is not so much the size of the set as the complexity of its structure that is crucial for a successful response to the circularity objection (Rabinowicz 1998, 19–20). Insofar as the justification-conferring system is divided up in several subsystems that also exhibit a sophisticated substructure of their own, the risks of a chain of justification 'biting up' its tail are minimal. For the justification of any element of the system depends on its coherence with sets of elements the coherence of which in turn depends on their connections with yet other sets of elements. Thus, a highly structured set of beliefs is necessary for the circularity of coherentist justification to be unproblematic.[11]

Constraint-satisfaction models of coherence, in addition to assuming a holistic view of justification—in that, as the principle of acceptance says, the justificatory status of any element depends on its coherence with the rest of the elements within the set, have an additional safeguard against the perils of circularity. In these models, relations of coherence are symmetrical so that if *p* coheres with *q*, then *q* coheres with *p*. Thus, these coherentist approaches do not justify the acceptance of a particular principle on the grounds that it coheres best with a set of rules and then justify a decision based on these rules because it coheres best with this very principle. Rather, on these views, the acceptance of a decision depends on its coherence with the rest of relevant normative elements so that particular decisions, on the one hand, and

[11] Dworkin's reply to the so-called objection from 'theory-dependence,' an objection which, at bottom, is a version of the problem of circularity, also appeals to the relevance of complexity. On the relation between the objection from circularity and the objection from theory-dependence, see Marmor (1991, 79). For a statement of the objection from theory-dependence, see Dworkin (1985, 169). For Dworkin's response, see Dworkin (1983b, 293).

normative elements, on the other, which cohere with one another are claimed to mutually support each other, rather than one being inferred from another.

3.4 Coherence Versus Authority

One traditional objection against coherence theories of justification is the ‘isolation’ or input objection. The objection is that since coherence is exclusively a matter of internal relations within a system of beliefs, it cannot allow for any input from the external world and, therefore, it isolates justification from reality. For example, a coherent system of beliefs that were the product of a visionary mad man or literary fiction would be, according to this theory, justified.[12] In the context of law, this objection translates into the worry that coherence theories of justification cut off justification from the body of rules and standards accepted as authoritative. The critique is that coherence theories are inconsistent with the authoritative nature of law, as they fail to give an account of the pivotal role that legal sources ought to play in determining what the law is and how cases should be decided.[13]

A popular response to the isolation (or input) objection against epistemic coherentism consists in adding some material constraints on coherentist justification to ensure that one’s system of beliefs takes good notice of empirical input and is thus properly connected with the external world. This is the strategy deployed by Lawrence BonJour, who appeals to the Observation Requirement, i.e., the requirement that a system of beliefs must contain beliefs to the effect that a reasonably variety of kinds of observational beliefs are likely to be true (BonJour 1985, 141). An alternative way in which the isolation objection may be counteracted is suggested by another leading proponent of coherence theories of epistemic justification, namely Keith Lehrer. In contrast to BonJour, Lehrer does not directly impose any material constraints on the justification-conferring system. His suggestion is that these constraints indirectly follow from the very conception of justification. According to Lehrer, a person is justified in accepting that *p* if and only if all objections to *p* are either beaten or neutralized, including the objection that one is not properly connected with the external world. Thus, what is needed for meeting the objections against one’s claims also suffices, in Lehrer’s view, to meet the isolation (or input) objection (Lehrer 2000).

Principle-based theories of normative coherence have deployed both of these strategies to meet the isolation objection as this objection is raised against coherence theories of legal justification. Peczenick claims, following Lehrer, that the isolation objection is just a skeptical claim that, like any other challenge, ought to be beaten on the basis of one’s system of acceptances and preferences (Peczenik 1998, 12; 2000a, 163; 1999, 199; and 2000b, 293). The recourse to second-order beliefs that is critical

[12] For a presentation of the isolation objection, see Pollock (1974, 27–28). Pollock replies to this very same objection in Pollock (1986, 76–77).

[13] See Raz (1985, esp. 305–310). For a discussion of Raz’s critique, see Michelon (2011) and Rodriguez-Blanco (2001).

to BonJour's response to the isolation objection seems to be also quintessential, I would argue, to the response given by some leading coherentist approaches to legal justification to the objection that coherentism severs justification from the body of authoritative rules and standards. The reply is the following one: A coherence approach to theory construction in law demands that one make cohere a set of beliefs that includes not only first-order beliefs about law and political morality, but also beliefs about the relative weight and relevance of first-order beliefs. Thus, a theory about what the law requires that is unconnected to authoritative sources will be ruled out as unjustified because it would fail to cohere with one's beliefs about the proper role of authoritative sources in theory construction in law.[14]

In a similar fashion, constraint-satisfaction theories of coherence overcome the isolation objection insofar as they give to reasons from authority a priority in being accepted. Constraint-satisfaction approaches to coherence are discriminating, that is to say, unlike pure coherence theories, they give to a set of elements a degree of acceptability on their own, even if their final acceptance depends, like any other element, on their coherence with all the elements within the set. In the case of explanatory coherence, for example, the theory—through the 'principle of data priority'—favors the acceptance of propositions that describe the results of observation (Thagard 2000, 43). Likewise, a constraint-satisfaction approach to normative coherence gives to propositions that describe the reasons of legal authorities a degree of acceptability on their own (Amaya 2015, 499). Thus, propositions describing reasons from authority play a role in computations of normative coherence analogous to the role that propositions describing observations play in computations of explanatory coherence. The discriminating nature of constraint-satisfaction approaches to justification, and more specifically, to legal justification provides the main mechanism whereby the risk of isolating justification from the set of authoritative rules and standards is avoided.

3.5 Coherence and Legal Conservatism

Against coherence theories, it has been argued that they have a tendency to favor the status quo and that they prevent genuine revisions of beliefs.[15] In particular, coherence theories of law urge legal decision-makers to stick to previously held views and decide cases accordingly. In short, the charge is that coherence theories, insofar as they take justification to be a matter of coherence with the settled law, are a roadblock to legal change. That this objection is a serious one becomes clear as soon as one considers the implications of coherence-based legal decision-making in morally wicked legal systems, if such theories—as the objector claims—have an inbuilt conservative tendency. In such systems, coherence theories, as Raz puts it, 'require further injus-

[14] Arguably, this is the kind of response that is suggested by Hage and Pecenick (2004, 337), Hage (2004, 97–99), and Dworkin (1983b, 312).

[15] For a statement of this objection, see Williams (1980, 249).

tices to be perpetuated in "hard cases" in the name of coherence.'[16] This objection may be viewed as the flip side of the objection from isolation. Both objections call into question the capacity of coherence theories to give reasons from authority their due. Coherence theories may give such reasons either too little (the objection from isolation) or too large (the objection from conservatism) a role in justification.

What are the responses to this objection available in the coherentist literature? Dworkin (1986, 220) has given a response to the objection from conservatism by arguing that once 'we grasp the difference between integrity and narrow consistency' we may come to see that 'integrity is a more dynamic and radical standard than it first seemed, because it encourages a judge to be wide-ranging and imaginative in his search for coherence with fundamental principle.' That is, coherence theories enable judges to justify their decisions by virtue of their coherence with fundamental principles that are necessary to justify the law as a whole. This reply, however, would fail to avoid the undesirable conservative implications in systems in which even fundamental principles are morally outrageous. For consistency in principle, as opposed to bare consistency, would still prevent judges from mitigating, rather than propagating, the injustices of such regimes.

Another response to the problem of conservatism, I would argue, is available for (strong) coherence theories in law. A coherence theory of legal justification—such as Dworkin's or Peczenik's—which includes moral reasons into the base of coherence, properly developed, has the resources for meeting this objection.[17] It is possible to downplay the impact of unjust source-based law in the outcome of a coherence calculation (i.e., in the result of applying a test of coherence to a particular base) by requiring that a justified legal decision must cohere as well with sound moral principles. Hence, a coherence theory will not perpetuate injustice in systems in which fundamental principles of law are morally indefensible if it takes the kind of coherence that is relevant to legal justification to be coherence among both legal reasons and moral reasons. In such a theory, a decision may cohere with a set of legal and moral reasons even if it fails to cohere with (morally wicked) fundamental principles of the settled law.

However, there must be limits to the foregoing coherentist strategy for meeting the objection from conservatism or else decisions that are justified according to a coherentist standard of justification could not be regarded, in any sense, as the result of interpreting the legal materials rather than simply as a result of following the requirements of morality. That is to say, to use Dworkin's terminology, one cannot make the set of beliefs about fit with settled law and political morality coherent by loosening the requirement of fit in such a dramatic way as to make a legal decision, one that is justified by virtue of its coherence with such a set, unrecognizable as a 'legal' decision, rather than a moral one. Otherwise, requiring coherence with both moral and legal reasons for legal justification would solve the problem of conservatism only

[16] Raz (1986, 111). Raz's critique is directed specifically against Dworkin's theory of coherence. See also Wacks (1984).

[17] The line of response I suggest is thus not available to weak coherence theories insofar as these theories take coherence with the settled law to be a necessary condition of justification.

at the cost of making the coherence theory fall pray of the isolation objection—which I have just discussed.

Constraint-satisfaction theories of coherence in so far as they give a priority in being accepted to propositions describing reasons from authority clearly sanction a conservative tendency. This, however, does not make these theories vulnerable, I would argue, to the objection that coherence inference is unduly conservative. First, that reasons from authority enjoy a privileged status is part and parcel of what it means to be engaged in 'legal' reasoning in the first place. Thus, it cannot be a problem for a coherence theory that it acknowledges the prominent role that such reasons play when justifying normative conclusions in law. Nonetheless, it is also desirable that a theory of legal reasoning allow for normative change when necessary. The constraint-satisfaction approach to normative coherence recognizes this need in that it permits that reasons from authority be rejected when doing so significantly increases the coherence of the overall set of reasons. The acceptance of any proposition, including the propositions that describe reasons from authority, depends (as the 'principle of acceptance' dictates) on their coherence with all the elements within the system. Thus, this theory exhibits a moderate kind of conservatism which is a distinctive characteristic of our practices of legal justification.

3.6 The Alternative Coherent Systems Objection

A standard objection against coherence theories of justification is the so-called 'alternative coherent systems objection.' There will always be many systems of beliefs that are equally coherent and between which the coherence theory will be unable to choose in a non-arbitrary way (BonJour 1985, 25). In the context of normative reasoning in law, the objection is that coherence theories fail to provide a criterion for choosing between decisions that cohere with the settled law equally well.[18]

The force of the alternative coherent systems objection varies depending on the kind of coherence theory of legal justification that is being proposed. This objection has no force against coherence theories, such as MacCormick's, which take coherence to be a necessary, albeit insufficient, condition for legal justification. These theories are ready to admit that there may be cases in which the coherence theory does not provide determinate guidance, and among which one should choose on the basis of moral considerations. Under this view, moral merit acts as a tiebreaker between equally coherent theories. But the need to appeal to moral considerations does not show any flaw of this version of the coherence theory, for they never claimed that coherence is a sufficient condition for justification in the first place.

The objection mainly affects those coherence theories that claim that coherence with the settled law is only one among other conditions of legal justification, but do

[18] For statements of this objection, see Kress (1996, 538–539) and Raz (1992, 299 and 309 n. 64). Some authors have given yet other reasons why coherence theories undermine the determinacy of law. See Edmunson (1996) and Mackie (1983, 168–169).

not give coherence lexical priority over other values. Under this view, the justified decision is the one that results from the best combination of coherentist and other considerations. This version of the coherence theory gives rise to legal indeterminacy, for as Raz (1992, 299) puts it, 'there is no way of deciding which mix of coherence and other values is best.' Several 'mixes' of coherence and other values may be equally good—so the objection goes—and coherence theories leave us with no guidance as to how to select one as best.

Before examining what coherence theories have to say against this objection, it is important to notice that the objection so stated misrepresents the coherence theories that are defended in legal scholarship in an important respect. For the conflict among values is not one between coherence and moral value/s but, as Alexy (1998, 46) puts it, one 'inside coherence.' Coherence theories, such as Dworkin's or Peczenik's, do not take coherence to be one value among many. They do not ask legal decision-makers to balance coherence with the settled law against other values, most importantly, moral values. Instead, what they propose is that we broaden the base of coherence so as to include not only authoritative reasons, but also moral reasons. These coherentist views are best described as urging decision-makers to take the decision that coheres best with a base broadened in this way. In other words, what the 'best' mix consists of is pretty clear for coherence theories, namely the 'most coherent' one. The problem that the objection correctly understood raises is whether there may be different 'mixes' that are equally coherent, so that the coherence theory fails to provide guidance for choosing among two decisions which equally cohere with a set of beliefs which comprises both beliefs about the law and beliefs about morality.

Now, what are the responses to this objection available in the current state of the coherence theory in law? The responses heavily depend on the different views that coherence theories have on the scope of reason within the domain of morality. To recall, Peczenik contends that we achieve coherence in legal reasoning by means of weighing and balancing the relevant values, but the last step of this operation is a matter of 'personal feeling' or 'individual preference.' Given that individuals have different preferences, a weighing operation may yield different justified outcomes. Thus, according to Peczenik, the fact that coherence methods yield different but equally justified decisions in hard cases is only a natural consequence of the subjective nature of moral values. It is only if we (incorrectly) assume that there needs to be only one correct answer that the alternative system objection should be taken to be an objection in the first place. This position, I would argue, is deeply unsatisfactory. It just disposes of the problem by endorsing a highly implausible subjectivist view of moral and legal values.

Dworkin's response to the objection we are considering is markedly different from the one given by Peczenik. According to Dworkin, if law—as he claims—is an enterprise in which propositions are assertable as true if they provide a better fit than their negation with propositions already established, then the question of whether in a hard case the judge ought to decide for one side or the other almost certainly has a right answer. For it is very unlikely that, in modern legal systems, one answer will not provide a better fit than the other (Dworkin 1972, 75–76 and 83–84). Dworkin concedes that there is a theoretical possibility of a tie between cases in which the argument from one side may be as good as the argument from the other. However, he claims that in complex legal systems such ties must be rare. Thus, Dworkin's reply to the objection that coherence fails to provide guidance in cases in which there are equally coherent decisions is that the situation the objector envisages is just too rare to have any relevance for practical purposes. That it is unlikely that in modern legal system such ties occur presupposes, he admits, 'a conception of morality other than some conception according to which different moral theories are frequently incommensurate' (Dworkin 1983a, 272). Thus, Dworkin's response to the objection is based on the assumption that values are commensurable (or at least that incommensurability is a rare possibility).[19]

Although this response does not solve the alternative systems objection, it does mitigate its force by diminishing its practical import. However, whatever success this line of response has depends on the (controversial) view that values are in some sense commensurable.[20] There is, however, no reason why coherentism needs to be committed to the view that moral and legal values are commensurable. In fact, prominent proponents of coherentists approaches to practical reasoning have defended the view that coherentism is neutral as far as commensurability/incommensurability is concerned (Hurley 1989) and others have even explicitly rejected commensurability as well as resisted the supposed dilemma between commensurability and irrationality (Richardson 1994). Thus, it is an open possibility for the coherentist to show that one may reason about conflicting values in law in a way that does not assume value commensurability.

3.7 The Problems of Holistic Coherentism

Coherence theories in law are often committed to holism about justification. This is certainly the case of theories that embrace a 'global' version of coherentism according to which it is the whole system of beliefs about the law and political morality that is relevant to determining the coherence, and thus, the justificatory status, of any particular legal decision (Dworkin 1986; Peczenik and Hage 2000, 2004; Hage 2004, 2013). This unrestricted holism is deeply problematic. First, holism about justification is psychologically implausible. No judge—save Hercules—has the memory resources and cognitive capacities required to establish coherence within the whole domain of the law—let alone within the whole domain of law and morality.[21] Second, holistic models of justification offer a poor description of the processes whereby

[19] In a similar vein, Alexy (1998, 46–48) claims that a coherence approach is based on a non-skeptical view about the possibility of weighing values and on some kind of commensurability among values.

[20] See Finnis (1987, 374–375), Mackie (1983, 165), and Raz (1992, 309 and 312), for critiques of Dworkin's coherence theory on the grounds that it assumes value commensurability.

[21] In fact, Dworkin seems to be well aware of this problem. See Dworkin (1993, 144, and 1986, 245).

legal decision-makers justify their decisions, for judges and other decision-makers do not typically engage in the kind of global justification that holism requires. And, lastly, holistic models of legal justification have undesirable normative consequences in that they make the justification of any single decision defeasible on grounds of the incoherence of any part of the legal system. While it is highly plausible that this is so if the incoherence arises between beliefs that are in the 'near neighborhood,'[22] it is problematical that we should rule out as unjustified decisions that fail to cohere with largely unrelated areas of law.[23]

Global models of coherence, insofar as they identify the domain of coherence—that is to say, the set of elements coherence with which yields justification—with the whole body of beliefs about the law and political morality, inherit all the problems associated with holism about justification. However, coherentism does not need to be tied up with holism. There are versions of coherentism that are not holistic and therefore not vulnerable to the aforementioned problems. One could defend a 'local' version of coherence, thereby restricting the domain of coherence to a specific area of the law. Alternatively, one could also defend a contextualized version of coherentism (Amaya 2015, 525–531). In this view, the domain of coherence shifts with context so that the set of beliefs that is relevant to assessing the justification of a legal decision depends on a number of contextual factors such as the costs of being wrong, one's institutional role or the resources available. Thus, both local and contextual models of coherence specify the domain of coherence in ways that avoid the objections raised against holism about justification.

3.8 Value Pluralism, Conflict, and Coherence

Coherence theories, some scholars have argued, are based on the wrong assumption that legal systems are coherent (or that they may be reconstructed so as to be coherent in a way that is compatible with their authoritative nature).[24] This assumption is untenable given the impact of political contest in the shape and development of the law as well as the pluralism of values that the law is likely to reflect. This objection has been most forcefully stated by Raz.[25] According to Raz (1992, 295), while we may expect the law to be coherent 'in bits—in areas relatively unaffected by continuous political struggles,' there is no reason to expect the law as a whole to be coherent. Raz's argument is that (1) the content of the law must be determined by reference to the intentions of legal authorities; (2) there is a plurality of intentions, or, as he puts it, 'there is no spirit to the law, only different spirits' (ibid., 302); therefore, (3) there

[22] The phrase is Plantinga's. See Plantinga (1993, 112).

[23] The point is persuasively argued for by Schauer (1986–1987, 858).

[24] This is the problem of authority, which I have discussed above.

[25] This objection has also been raised by Critical Legal Studies, specifically against Dworkin's version of the coherence theory. See Kennedy (1997). For a discussion of the critique advanced by Critical Legal Studies, see Waldron (2008).

is no reason to expect the law to be coherent. By assuming that the law is coherent (or by imposing coherence on the whole of the law) global coherence accounts err in two ways. First, they underestimate the degree and implications of value pluralism. And second, they wrongly attempt to idealize the law out of the concreteness of politics. But in countries with decent constitutions the untidiness of politics is sanctioned by the morality of authoritative institutions, and thus there is no reason why the effects of politics ought to be minimized (ibid., 309ff.).

To what extent does this critique undermine coherence theories? First, one should notice that the critique is directed against global versions of coherence, and thus that it leaves non-global accounts intact. Hence, whatever the force of the objection is, a properly developed local or contextual version of the coherence theory is immune to this objection. In addition, it is worth pointing out that Raz's critique is most effective if coherence, as he suggests, is understood as 'unity' (ibid., 286). However, there is no reason why coherence should be so defined. In fact, most coherence theories in law do not endorse such a view of coherence.[26] Raz saddles the coherence theory with a conception of coherence that makes these theories more vulnerable to the objection we are considering.

That said, the most important line of response to Raz's critique is to deny the assumption from which it proceeds, to wit, that coherence methods aim at explaining away the value conflict that is inherent in law. In contrast, I would argue coherence rather than explaining conflict away provides a way to proceed in the face of conflict.[27] It is precisely because 'morality is a plurality of irreducibly independent principles' and because 'the reality of politics leaves the law untidy' that we need coherence methods in the first place. Thus, coherence accounts, far from being incompatible with the value pluralism that pervades our legal systems, can arguably be seen as providing a method for guiding legal decision-making, given such value pluralism. Some coherence theories of practical reasoning provide illuminating accounts (briefly discussed below) as to how one may reason in a coherentist way in cases of value conflict. These approaches may prove to be useful for the purposes of showing how coherence methods may help us reason about the different and oftentimes conflicting values that inform the law in democratic societies.

3.9 Coherence and Truth in Law

Arguably, the main objection raised against coherence theories of legal justification is that there is a questionable relationship between coherence and truth. Do we have any reason to believe that coherentist standards of legal justification are truth-conducive?

[26] See Kress (1996, 539–546), for an analysis of the different ways in which coherence and unity relate to each other in various conceptions of coherence. Most versions of the coherence theory in law do not define coherence as unity. A notable exception is Weinrib's theory. See Weinrib (1988, 1994). For a critique, see Kress (1994).

[27] For an extremely illuminating account of the relationship between coherence and conflict in moral deliberation, see Hurley (1989), part III, to whom this response to Raz's critique owes much.

Do coherence methods lead us to accept as justified propositions about the law that are also likely to be true?

To begin with, it is important to notice that the problem of how coherence and truth connect up does not arise if one holds, as MacCormick does, a weak coherence theory of legal justification. Because weak versions of the coherence theory do not claim that coherence is all there is to legal justification, they do not need to address the issue of how endorsing as justified beliefs about the law by virtue of their coherence leads one to accept true propositions about the law. The connection between justification and truth in law may be established by means other than coherence. The problem arises for 'strong' theories of legal justification, according to which coherence is both a necessary and a sufficient condition of justification.

Among (strong) coherence theories of justification, we may distinguish between two different positions. One may hold a coherence theory of legal justification while endorsing a realist view about legal truth.[28] Moore's version of natural law theory provides an example of such a view.[29] The truth objection against these views is particularly forceful. In order to meet the objection, these theories need to show that coherence, which is a matter of internal relations, yields beliefs that are likely to be true, in the sense that they correspond with a realm of mind-independent legal facts.[30] The problem of the truth-conduciveness of coherence, when truth is understood along realist lines, is not, however, insoluble. There is an array of different strategies in the coherentist literature that purport to show that the appropriate connection obtains between coherence (as correspondence) and truth. BonJour (1985) and Thagard (2007, 2012) have employed an inference to the best explanation to meet the truth objection. The connection between coherence and truth is effected in other theories by assuming an externalist epistemology (Lehrer 2000). Davidson (2001) has argued for the truth-conduciveness of coherence by forging a conceptual link between coherence and truth through the concept of belief. There are also responses to the truth objection that appeal to probability theory (Shogenji 1999). These strategies, I would argue, may provide a useful starting point for mounting an argument to the effect that coherence connects up with a realist conception of legal truth in the right way.

Alternatively, some coherence theories of justification take an antirealist stance toward legal truth. More specifically, these theories endorse a coherence theory of truth, i.e., the view according to which the nature of truth is constituted by a relation of coherence between the belief being assessed and other beliefs (Lynch 2001, 97–198; Walker 1989; and Alcoff 1996). Truth, on this view, is not discovered but constructed

[28] It is crucial to note that by 'realism' about truth, I am referring to any theory of truth that claims that truth-makers are mind-independent facts. This is a kind of 'metaphysical' realism, importantly distinguished from 'legal' realism, as a theory of legal decision-making. In fact, legal realists typically reject realism about truth (at least in the normative domain).

[29] See Moore (1985, 2004). See also Brink(1989), for a defense of a realist *cum* coherentist moral epistemology. See BonJour (1985), Lehrer (2000), and Thagard (2000), who defend a coherentist theory of epistemic justification and realism about the truth of empirical beliefs.

[30] For an analysis of the problems that arise from combining a coherence theory of legal justification with a realist view about legal facts, see Coleman and Leiter (1993, 612–616).

through a process of coherentist justification. It is thus coherence with a specified set of beliefs that yields legal truth. These coherence theories provide a direct response to the truth objection insofar as they establish a conceptual relation between coherence and truth.

Three main versions of the coherence theory of legal truth may be distinguished. A first version claims that whether propositions about the law are true is a matter that partially depends on their coherence with personal morality. In this sense, the content of the law is relative to one's views about morality (Hage and Peczenik 2004). The problem with this position is that it injects a subjectivist element into the coherence theory of legal truth, which is very unsatisfactory. A similar objection may be raised against coherence theories of truth in law that claim that legal truth is partially dependent on moral truth understood as coherence with the mores or social conventions. In these views, a connection between truth and coherence is established at the price of adhering to moral relativism.[31]

A second version claims that it is coherence among beliefs about the law and morality that would be held under ideal conditions that yields legal truth.[32] This is the view that, according to some interpretations, Dworkin may be seen as defending.[33] This view would have to face severe objections against 'ideal' views about what morality requires, the most significant of which is their seeming inability to guide conduct in a less-than-ideal world.[34]

Last, one may claim legal truth to consist in coherence with the best theory of law and political morality that may be devised under real-life constraints—as distinct from the one that we would elaborate under ideal conditions. Dworkin's writings on truth and objectivity in law may also be interpreted as arguing for this view.[35] A problem for this version of the coherence theory of truth is that it is uncertain whether it articulates a sufficiently stringent standard for assessing the truth of legal claims. It seems that some constraints ought to be imposed upon the relevant set of alternative theories so as to ensure that the best out of this set has some legitimate claim to truth. In addition, this view also makes truth opaque, in that it is uncertain

[31] For a defense of a relativist coherentist theory of the truth of moral judgments, see Goldman (1988). See also Young (2001), for a defense of a coherence theory of truth according to which truth amounts to coherence with the largest consistent set of propositions currently believed.

[32] There are also important theories about the truth of empirical statements that define truth as coherence by resorting to some idealized perspective. See Putnam (1981) (defining truth as coherence with the system of beliefs that we would hold at the limit of inquiry) and Rescher (1985) (defining truth in terms of ideal coherence, i.e., coherence with a perfected database).

[33] This view was defended by Coleman and Leiter (1993: 633–635). Coleman (2001: 165) later rejected this interpretation.

[34] This problem is closely related to the 'problem of access,' as Coleman and Leiter (1993, 629) term it. The problem of access is as follows. If, according to views of legal truth as 'ideal' justification, true propositions about the law are those that one would justifiably hold under ideal conditions, and given that ideal conditions do not obtain (by definition), then it follows that legal truth would be inaccessible to real judges operating in less-than-ideal conditions. Such a theory fails to guide judges in their decision-making task, for it makes legal justification and legal truth unattainable by real judges.

[35] I have argued for this view in Amaya (2015, 49–50 and 69).

how we can be sure that we have achieved the best possible theory (MacCormick 1983, 188).

To conclude, the problem of the truth-conduciveness of coherence remains an important one for current coherence theories in law, in both its realist and antirealist varieties. However, the problem is not, as argued, insuperable. Realist approaches may avail themselves of different strategies that aim at connecting coherence with truth—even if admittedly the challenge of showing coherence to be truth-topic when truth is interpreted as correspondence are considerable. Although different problems arise depending on the specific version of the coherence theory of truth being endorsed, constructivist approaches do succeed in meeting the truth objection. This significantly diminishes the import of this objection given that constructivism is generally regarded as a fairly plausible account of the nature of legal truth.

4 The Coherence Bias: A Plea for Responsibilist Coherentism

In the previous section, I have reviewed some of the main objections that have been directed against legal coherentism. These objections, as argued, are particular versions of general objections that have been traditionally raised against coherence theories of justification. More sophisticated coherence theories (as much in law as in any other realm) have been developed with a view to meeting or at least mitigating the force of these objections. There is a problem with coherence theories of law, however, that has not, received as much attention as it deserves. Arguably, this problem is not specific to legal coherentism, but it also affects coherence theories of justification in contexts other than the legal one. The problem is the following one. There are cases in which, because of serious defects in the processes of inquiry and deliberation, the theory that best satisfies the criteria of coherence is, nonetheless, intuitively unjustified. Two kinds of problem-cases may be distinguished:

(i) *Problems with the base of coherence*. These are cases in which the input to coherence-based reasoning, that is, the set of relevant normative elements and interpretative hypotheses over which the calculation of coherence proceeds, is the result of defective inquiry. In cases of this sort, a theory of the case that best satisfies the criteria of coherence does so only because the legal decision-maker has taken into account a less comprehensive body of beliefs about the law and political morality than the set that would have resulted if she had inquired properly about the case.

To start with, legal decision-makers may ignore relevant authorities. For example, they may selectively choose among the relevant precedents those earlier decisions that support their working interpretative hypothesis and purposefully disregard authorities which cast doubts upon this proposition. Furthermore, research in memory suggests that memory for information congruent with prior beliefs is better than memory for information irrelevant to prior beliefs; there is also evidence that people tend to drive their attention to encode information consistent with their expectations and are prone to seek information that supports rather than disconfirms their beliefs. This 'confirmation bias' may distort the generation of the base of coherence, which would be right from the beginning biased toward one of the decision alternatives being considered.[36]

Difficulties may also arise regarding the construction of the contrast set, i.e., the set of relevant alternatives. Lack of imagination, prejudice, excessive reliance on the parties' configuration of alternatives, or professional routines may lead legal decision-makers to ignore relevant alternatives. On the coherence-based account of legal inference, as explained above, legal inference in law works by exclusion. One theory is accepted as justified on the grounds that it is the best of a set of available alternatives on a test of coherence. But then unless the legal decision-maker has a reason to believe that she has ruled out the relevant alternatives to her claim about what the law requires, belief in the best interpretative hypothesis fails to be justified. To be sure, inferring to the best of a 'bad lot' cannot yield justified beliefs.[37] At the very least, a relevant set of alternative theories need to be considered for conclusions of inference to the best explanation to be justified.

(ii) *Problems with the coherence calculation*. There are cases in which a theory of the law satisfies the criteria of normative coherence but the reason why it does so traces back to certain defects in the way in which the legal decision-maker performs the coherence calculation. In these cases, the reasoning is defective for reasons that do not have to do with the input to such a process (i.e., the relevant normative elements) but rather with the quality of the process as such. Legal decision-makers may attempt to maximize coherence by inflating some alternatives while deflating others. In fact, there is substantial evidence that shows that this 'coherence bias' is at work in the evaluation of hypotheses in the legal context. In the process of legal decision-making, subjects restructure the diverse and conflicting considerations that provide equivocal support for different decision alternatives until they reach a representation in which the chosen alternative is supported by strong considerations and the rejected one is supported by weak considerations. Once the decision alternatives have been manipulated in such a way, the evaluation of those alternatives is already skewed toward one's preferred alternative (Simon 2004).

In all these cases, there does not seem to be sufficient reason to accept as justified the outcome of the process of a coherence-driven legal inference. The theory that best satisfies the criteria of coherence seems intuitively unjustified, as it satisfies the criteria better than alternative theories only because the legal decision-makers have reasoned about the case in a defective way. It is not only that legal decision-makers are to be blamed for inquiring and deliberating about the case in a defective way, but

[36] On the confirmation bias, see Mercier and Sperber (2011, 63–66).

[37] This is the so-called 'objection from the bad lot' or 'problem of underconsideration' raised by Van Fraassen against models of inference to the best explanation in science. See Bas Van Fraassen (1989, 142–150).

the justificatory status of the theories that result from such processes of inquiry and deliberation is undermined. To start with, the coherence that results from reasoning from a defective base or by distorting the deliberation factors does not seem to be epistemically valuable. There does not seem to be anything especially worthy about believing a theory about what the law requires by virtue of its coherence, when such coherence is but the product of one's cognitive failure. Besides, even if one might attribute some merit to having a coherent system of beliefs, even in cases in which such coherence results from objectionable epistemic behavior, there is a straightforward sense in which belief in such theories is unjustified: for, had the legal decision-maker been conscientious in forming his belief, he would not have accepted such a theory on the grounds that it coheres best. In other words, there seems to be a clear sense in which the legal decision-maker ought not to believe the way he does (Baehr 2009, 549–552).

This problem is a serious one for any coherence theory of legal justification. One may reach coherence—however such notion might be defined—by reasoning from a defective base or one may construct coherence over the course of decision-making in a biased way and the resulting system of beliefs seems unjustified, despite it enjoying a high degree of coherence. The coherence theory of justification needs to be modified in order to block ascriptions of justified belief in cases in which coherence-based inference is vitiated in some of these ways. The suggestion is that there is a need to impose a further condition on the process of coherence maximization for it to be justification-conferring, namely it has to be such that an 'epistemically responsible' agent could have reached such a conclusion in like circumstances.[38] Coherence irrespective of the process whereby it may be reached does not yield justification: a belief is justified only if it could be the outcome of epistemically responsible coherence-based reasoning.[39] In short, coherentism needs to be wedded to a 'responsibilist' account of justification. Thus, a full-fledged coherence theory of legal reasoning needs to provide a plausible account of the standards of epistemic responsibility that should govern legal decision-making.

Now, what is it is for a legal decision-maker to behave in an epistemically responsible way? Two main accounts of epistemic responsibility may be distinguished: a 'deontic' approach and an 'aretaic' approach. Under a deontic approach, epistemic responsibility is a matter of duty-fulfillment. One is epistemically responsible to the extent that one complies with one's epistemological duties, such as the duty

[38] For attempts to impose similar responsibility constraints on justification, see Baehr's account of evidentialism (2008, 484–485), BonJour's account of a priori justification (1998, 110–115), and Khalifa's defense of inference to the best explanation (2010).

[39] It critical to note that the responsibilist account of legal justification proposed here is 'counterfactual' rather than 'causal.' A causal version would say that a legal decision is justified if and only if it is the outcome of epistemically responsible coherence-based reasoning. In contrast, the counterfactual version states that a decision is justified if and only if it *could* be the outcome of epistemically responsible coherence-based reasoning. Thus, the counterfactual version allows for the possibility that a legal decision be justified, provided that an epistemically responsible legal decision-maker could have accepted it as justified, even if it results from an epistemically irresponsible process of coherence maximization.

to believe as the evidence dictates or the duty to seek out more evidence about propositions that are less than certain on one's evidence.[40] According to the aretaic conception of epistemic responsibility, one is epistemically responsible insofar as one properly exercises a number of intellectual virtues, such as diligence, courage to face criticism, perseverance in following a line of inquiry, or open-mindedness.[41] Epistemically responsible action amounts, in this approach, to intellectually virtuous behavior.

Initially, a deontic approach to the epistemic responsibility of legal decision-makers seems more adequate than an aretaic one, for the law aims at requiring standards of conduct that are minimally acceptable, rather than at setting up ideal models of conduct. We may distinguish, following Fuller, between two kinds of morality: the 'morality of aspiration' and the 'morality of duty.' While the former is, 'the morality of the Good life, of excellence, of the fullest realization of human powers,' the latter, 'lays down the basic rules without which an ordered society is impossible' (Fuller 1969, 5). It is the morality of duty, rather than the morality of aspiration, says Fuller, that provides us with 'workable standards of judgment in law,' for, 'there is no way open to us by which we can compel a man to live the life of reason. We can only seek to exclude from this life the grosser and more obvious manifestations of chance and irrationality' (Fuller 1969, 9). Thus, given law's aim of ordering social life and the limits of what can be achieved by legal means, it seems that we must turn to deontic concepts in order to provide an account of the standards of epistemic responsibility that should govern legal reasoning.

However, there are some reasons why, I would argue, an aretaic approach to legal decision-maker's epistemic responsibility is preferable.[42] First, virtue concepts have the advantage of greater richness than deontic concepts. Bernard Williams' well-known distinction between 'thin' and 'thick' concepts is to the point here (Williams 1985). Unlike deontic concepts, virtue concepts are 'thick' in Williams's sense and they convey not merely a negative or positive epistemic evaluation, but indicate the way in which the legal decision-maker acted properly or improperly. Second, a virtue approach does not (implausibly) reduce good epistemic practice to rule-following. Just as there does not seem to be any complete set of rules sufficient to give a determinate answer to the question of what one should do in a particular situation of moral choice, epistemic evaluation does not seem to be either strictly rule-governed. Third, virtue approaches to epistemology are in a better position than deontological ones to give an account of epistemic values such as wisdom, insight, or understanding, which are surely critical in the context of legal decision-making. Last, virtue epistemology allows us to put forward an ideal of the legal agent according to which legal decision-makers do not merely aspire to avoid prohibited epistemic

[40] The most influential contemporary statement of such a view is by Chisholm (1977). For an account of epistemological duties, see Feldman (1988, 2002).

[41] The most influential statements of 'virtue responsibilism' include Montmarquet (1993) and Code (1987). For a discussion of this trend in virtue epistemology, see Greco (2002).

[42] For a lucid discussion of some advantages of virtue theories in both epistemology and ethics, see Zagzebski (1996, 15–29).

conduct, but they aim to engage in epistemically valuable conduct. Thus, an aretaic approach to epistemology has some distinctive advantages over a deontic one.[43] In light of these reasons, virtue epistemology, I would argue, provides us with an adequate framework for analyzing the standards of epistemic responsibility of legal decision-makers.[44]

To conclude, coherence yields justification only against a background of epistemically responsible action. Virtue-based standards of epistemic responsibility, I have argued, ought to figure in a coherentist analysis of legal justification. A main finding in contemporary epistemology is the recognition of the relevance of features of the agent to attributions of justified belief. A plethora of virtue approaches to knowledge and justification have flourished in the last decades, which aim to give an account of the contribution that agents bring to processes of justification. The need for epistemic responsibility is even more acute in a coherentist approach to justification, given human's deep-rooted psychological tendency to make things add up. We seem to be 'biased' toward coherence, as Paul Ziff says, 'We humans are fanciers, connoisseurs, of coherence… coherence catches our eye, fixes our attention, focuses our mind' (Ziff 1984, 34). The drive toward coherence, which seems to be hard-wired in our cognitive make up, endows the coherence theory of justification with a high degree of psychological plausibility (more on this, later). However, this tendency toward coherence is also a source of problems, for one may aim to achieve coherence in objectionable ways, which detracts, and ultimately, deprives, the final outcome of any epistemic value. Hence, in order to put worries about coherence biases to rest, there is a need to include epistemic responsibility, and, more specifically, virtue as a core concept in a coherence theory of legal justification.

5 The Value of Coherence

Why does coherence justify in law? Why should one endorse a coherentist standard for the justification of normative propositions in law? The value of coherence as a standard of justification is often called into question by raising doubts about its truth-conduciveness. The strategy of the critic of coherence methods is as follows: It is first claimed that a defense of a theory of justification must show that accepting beliefs as justified according to the theory leads one to accept beliefs which are likely

[43] Alternatively, one could endorse an irenic approach to the standards of epistemic responsibility, which combines deontic and aretaic elements. I have defended such a view before (Amaya 2011). Or one could also make virtues derivative of duties. One way this could be done is by appealing to a duty to 'behave in ways that will maximize the person's number of true beliefs and minimize that person's number of false beliefs' (Feldman 2002, 372), where these behaviors would include the cultivation and exercise of intellectual virtues.

[44] The acceptance of a virtue framework does not, however, imply that deontic notions are irrelevant for determining whether a legal decision-maker has behaved in an epistemically responsible way. A virtue account of the epistemic responsibility of legal decision-makers is compatible with assigning to notions of duty important roles within the theory (Amaya 2015, 523–524).

to be true; next, it is argued that it has not been shown that coherence and truth are related in a proper way; hence, it is concluded, a defense of the coherentist standards of justification is doomed to failure.

Regardless of whether the prospects of meeting the truth objection are as poor as the critic of coherentism takes them to be, this strategy does not succeed in undermining the coherentist project because a thorough defense of the viability of a theory of justification does not—as the critic assumes—depend on showing the theory to be truth-conducive. While, to be sure, truth-conduciveness is a crucial standard for assessing the adequacy of a theory of justification, there are also other criteria that are relevant for evaluating a particular theory of justification. For one, even if truth is a momentous value when reasoning about both what to believe and what to do, we are also interested in achieving, in our reasonings, values other than truth, and, as long as this is so, the adequacy of a theory of justification will depend on how well it helps us realize the complex set of goals—truth included—that we aim at advancing in a particular domain. In what follows I shall suggest some lines of argument for supporting a coherence theory of justification for law.

5.1 The Argument from Antifoundationalism

The first reason for endorsing a coherentist standard for legal justification is a negative one, to wit, the enormous difficulties that the foundationalist view of justification encounters. As is well known, foundationalist theories of both moral and epistemic justification face severe objections. Similarly, the foundationalist view of justification that characterizes strong versions of legal positivism is in a serious predicament. The problems that foundationalism in its many varieties and in different domains faces make the coherentist alternative initially attractive, even if, of course, a full defense of coherentism requires that some positive reasons be given for its desirability.

5.2 Coherence and Emotion

Reasoning about both what to do and what to believe is a hot cognitive process as much in law as in any other domain. Thus, a complete theory of legal reasoning needs to give an account of the role that emotions play in reasoning about disputed questions of law. Some advocates of coherence theories have shown how emotions may be incorporated within their theories (Richardson 1994; DePaul 1993; Thagard 2006). Recent studies (Simon et al. 2015) have shown that coherence-based reasoning can be extended to encompass hot cognitions. Because relations of coherence (and incoherence) are not restricted to propositional elements, and because judgments of coherence (and incoherence) are sensitive to emotional responses, a coherence theory is better located than alternative theories to give an account of the role that emotions play in justification. This gives an important advantage to coherence theories of legal justification over alternative views of justification that are less easily amendable to making room for the 'hot' components of legal reasoning.

5.3 The Argument from Psychological Plausibility

As I mentioned before, our cognitive equipment seems to be geared toward coherence. There is substantial psychological evidence that shows the relevance of coherence in our reasoning processes. Empirical studies strongly suggest that we find explanatory thinking natural: Explanatory considerations are the engine that drives much inference in ordinary life (Lipton 2004, 108–113). Moravski (1990, 213) has persuasively argued that cognition can be viewed as an activity directed toward the goal of achieving understanding, and that, in an important sense, humans may be seen as *homo explanans*. Simon and collaborators (Simon 2004) have shown that complex decision-making tasks, such as judicial reasoning, are carried out by building up coherence among a number of decision factors. All the foregoing studies—among others—provide the coherence theory of justification with a solid empirical basis. The psychological plausibility of coherentism is an important argument in support of the coherence theory of legal reasoning. To start with, from the perspective of naturalism, it is a requirement of any theory of justification that it does not set up normative standards we are not even able to approximate. Facts about how we reason have a bearing—under this view—on questions about how we ought to reason. The coherence theory of legal reasoning seems to satisfactorily meet this naturalistic constraint. In addition, a theory of legal reasoning is expected to guide legal decision-makers in carrying out their task. The coherence theory, insofar as it builds upon legal decision-makers' natural reasoning processes, seems more apt to perform this guiding role than theories that impose forms of reasoning that are alien to them.

5.4 The Argument from the Dynamics of Justification

It is an advantage of the coherence theory that, unlike other models of justification, it has the resources to give an account of the dynamic aspects of justification (Haack 2000). Coherence accounts of justification allow us to model major conceptual changes, such as those involved in scientific revolutions, as well as piecemeal revisions in one's system of beliefs, like those formalized by belief revision theories (Hansson 2006; Olsson 1997, 1998; Thagard 1992). As opposed to the static view of justification that foundationalism assumes, coherence theories are congenial to a dynamic view of justification. In law, questions about how legal decision-makers assess the relative goodness of alternative interpretations about what the law requires in the particular case and integrate them into their accepted views about the law, are indeed important questions. These questions, given the dynamic dimensions of coherentist justification, may be profitably addressed within the framework of legal coherentism.

5.5 The Epistemic Value of Coherence

To be sure, even if, as I have said, a defense of a method of justification does not depend exclusively on having conclusive reasons for its truth-conduciveness, still, an argument to the effect that justified beliefs, according to the method, are also likely to be true is an important part of such a defense. How does the coherence theory of justification fare with regard to the goal of truth? As argued, although none of the arguments that have been designed to show that coherence and truth are related in the right way have succeeded in putting worries about the truth-conduciveness of coherence to rest, the case for the truth-conduciveness of coherence is not hopeless.

The problem of the truth-conduciveness of coherence is much less acute in the normative domain than in the factual one, given that non-realist accounts of the truth of normative propositions are generally viewed as more plausible than non-realist accounts of the truth of empirical propositions, and that coherentist theories of justification are less problematically combined with non-realist views of truth, than with views of truth as correspondence. Even if one endorsed a realist construal of the truth of propositions in law, the truth objection would not amount to a knockdown argument against legal coherentism, for—as argued before—there are several lines of argument whereby one may aim to connect up coherence with truth as correspondence. Thus, there seem to be reasonably good reasons for supporting the desirability of coherence methods from the perspective of advancing the goal of truth in law.

In addition, it is critical to note that there are several reasons why coherence is epistemically valuable regardless of its truth-conduciveness. As some proponents of probabilistic approaches to coherence have shown, coherence is conducive to reliability (Olsson and Schubert 2007, 2013; Schubert 2011, 2012) and confirmation (Dietrich and Moretti 2005; Moretti 2007) and it has also an important heuristic value (Angere 2007, 2008). Coherence has also been claimed to be intrinsically linked to the value of understanding as well (Cooper 1994; Österman 2001). Thus, there are a number of important epistemic reasons to seek coherence when reasoning in the legal context.

5.6 The Practical Value of Coherence

Legal institutions are designed to advance a number of different goals, and coherence—I shall argue—is a valuable tool for realizing some of these goals. As some advocates of coherentist approaches to practical reasoning have shown, coherence promotes successful coordination and effectiveness (Bratman 1987, 137; Richardson 1994, 152–158; Millgram and Thagard 1996, 67). In the legal context, some degree of coherence is also crucial for coordinating actions, and actions which cohere with each other tend to be more effective as well. Law is a collective enterprise, and legal decision-makers are more likely to succeed in their efforts at coordination if the decisions they make are part of a coherent plan of action. Similarly, law's project of regulating and transforming social life is more likely to work if it embodies a coherent plan than if it involves overlapping goals and conflicting courses of action. Thus, coherence has practical benefits as much in law as in any other practical domain. In addition, coherence aids the realization of values that are distinctive to the legal context. Coherence is instrumental to the value of legal certainty (Pino 1998; Moral 2003, 320). Among other ways in which coherence promotes legal certainty is by facilitating knowledge of the law, for a coherent body of norms is more easily remembered and understood. Coherence also promotes the efficacy of the legal system, for a coherent body of norms is also easier to be applied and followed. Besides, a certain degree of coherence in legal decision-making at both the legislative and the judicial level is also pivotal for securing the social stability the law aims to preserve (Alexy and Peczenik 1990, 145). Thus, there are valuable practical reasons for pursuing coherence in the course of legal decision-making (cf. McGarry 2013).

5.7 The Argument from the Social Function of Coherence

Several researchers in cognitive psychology have argued that coherence has an important social function. People who are not coherent are poorly perceived by others. In contrast, the coherence in one's statements makes it more likely that they will be accepted by others, i.e., coherence is positively correlated with consensuality. This significantly increases people's confidence, for when people think that their statements are likely to be shared by others, attitude certainty and strength increase. Thus, we have important incentives to 'look' coherent (Kurzban and Aktipis 2007). More specifically, there is evidence indicating that coherence evaluation, i.e., when we represent the coherence of different mental states and draw inferences from it, serves two interrelated functions. First, the evaluation of the coherence of other people's statements is used for 'epistemic vigilance,' i.e., to detect whether there is anything wrong with someone else's statements. And, second, given that communicated information is tested based on its coherence, we turn coherence evaluation mechanisms on ourselves in order to be perceived as being right (Mercier 2012).

Now, the social aims of coherence evaluation suggest another important set of reasons for pursuing coherence in legal reasoning: Coherence in legal decision-making increases the public acceptability of decisions[45]; it has a positive impact in

[45] Cf. Simon and Scurich (2011), showing that judicial decisions accompanied by monolithic reasoning—multiple one-sided reasoning—are judged as less acceptable than decisions accompanied by multiple two-sided reasoning. It is unclear, however, whether Simon and Scurich's results cast doubt over the persuasive value of coherence. As argued, coherence reasoning—and theorizing—rather than explaining conflict away, provides a way to proceed in the way of conflict. Thus, coherence reasoning need not be monolithic, but it is compatible with admitting to good reasons on both sides.

the citizens' confidence in the legal system; and it critically contributes to building consensus around issues which are, at times, fairly divisive, as happens when courts have to adjudicate in cases involving matters of fundamental principle. Thus, there are important social benefits associated with the coherence of legal decisions.

5.8 The Argument from Conflict Resolution

Law is a complex institution directed toward solving conflict through argumentative means (Atienza 2006, 59). Conflicts of values are pervasive in law. How may we rationally proceed when the different values that the law aims to protect come into conflict? Coherence is of a piece with a non-instrumentalist view of practical reasoning according to which we may rationally deliberate about ends, and not merely about what the best means to achieve some fixed ends are. Some of the coherence theories of practical inference provide illuminating accounts of how a coherentist approach to justification guides agents on how to deliberate about what the best course of action in light of a set of conflicting ends is. Richardson's (1994) coherentist version of specificationism provides us with a fruitful way of addressing normative conflict in law. When two norms or values come into conflict, legal decision-makers may rationally deal with this conflict by specifying them in a way that enhances coherence. Hurley's (1989) view that, when deliberating in the practical domain, judgments about what one should do, all-things-considered, are determined by the most coherent theory of the different values which apply is also extremely useful for understanding how a coherence theory would help deliberate about ends in the legal context. It follows from Hurley's account of deliberation that, when faced with conflict, legal decision-makers ought to take the decision, among competing courses of action, that is dictated by the theory that makes the best sense of the relationships among the conflicting reasons.

Thus, not only does coherence help us realize truth and other values in law, as I have argued above, but it also helps us reason about how to weigh and balance these values when they come into conflict in a legal case. That is, coherence is not merely instrumental to the ends that the law aims at promoting, but it crucially helps us deliberate about these very ends. In so doing, it expands the space of reason in law by providing a method for reasoning about ends in law, and not only for rationally assessing which, among competing courses of action, is best in promoting a given set of goals that are placed beyond the pale of deliberation. The capacity of coherence-based theories to provide guidance about how to deliberate about ends, and thus to realize law's function of resolving conflict by argumentative means, provides a strong reason in support of legal coherentism.

5.9 The Constitutive Value of Coherence

The foregoing considerations about the pervasiveness of conflict in law lead us to consider yet another reason for pursuing legal coherence, namely its constitutive value. Coherence plays a constitutive role in individual and political identity. A certain degree of coherence in individual and collective deliberation is necessary to be both a unified agent and part of a distinctive political community.[46] One may ask: In the face of conflict, why shouldn't one decide upon one of the conflicting values, as opposed to strive after coherence? The reason why this is so is that by deliberating about the values and goals in conflict when deciding a particular issue, legal agents are determining their own identity as members of a political community. Individual identity and group identity are not fixed—as Hurley (1989) brilliantly argued—but they are the result of self-interpretation. Legal decision-makers are not free to disregard a concern for coherence, because in so doing they would be refusing to determine their own identity as members of the political community to which they belong. The constitutive dimension of coherence in individual self-determination and group self-determination gives us a foundational reason for valuing coherence as a guiding standard in legal decision-making.

6 Conclusions

Coherence theories of justification are of a relatively new vintage—compared with the traditional, foundationalist, approach to justification. Nonetheless, coherentism has already been around for a while. What has the coherence approach to law and adjudication achieved thus far? Has it succeeded in articulating a solid alternative account of legal justification? What are its implications for the broader field of legal reasoning and legal theory? I shall conclude this chapter by briefly assessing where the coherence theory stands and suggesting some lines for further research.

In the last decades, coherence theories of justification have firmly moved from 'theory sketch' to 'actual theory.'[47] Coherentism—in law as well as in other disciplines—provides us, as of now, with a well-rounded and fairly detailed theory of justification. As such, how plausible are they? What credentials does coherentism have as a theory of justification and, more specifically, as a theory of legal justification? In the preceding sections, I have presented some of the main objections that may be raised against legal coherentism. Although these objections need to be taken seriously, they do not succeed in undermining legal coherentism. As argued,

[46] That coherence has a constitutive value in individual identity has been argued by Richardson (1994), Thagard and Millgram (1996), and, most extensively, by Hurley (1989). That coherence plays a constitutive role of political communities is argued in Dworkin (1986). See also Michelon (2011), arguing that the value of coherence may be grounded in a narrative conception of human life.

[47] The expression is Bender's (1989, 1).

coherence theories of legal justification have articulated several lines of response that overcome, or, at least, mitigate the force of these objections. In addition, there are a number of important arguments which jointly provide strong support for the claim that it is desirable to pursue (a certain) degree of coherence in the course of legal reasoning.[48] Thus, given the reasons that support the coherence theory in law and that the problems of legal coherentism are not intractable, a theory of legal justification that gives a prominent role to reasons from coherence may be claimed to provide a plausible account of the structure of justification in law.

Now, the plausibility of coherentism, I would argue, comes with 'impurity' as much in law as in any other domain. While it is plausible that coherence is a core ingredient of justification, the view that coherence is all there is to justification is highly questionable. More developed versions of coherentism across domains resort to non-coherentist elements in order to provide a detailed theory of justification that has the resources to counteract the traditional objections that have been directed against it. In this chapter, I have argued for the relevance of virtue-based standards of epistemic responsibility as a central element of legal coherentism. It is only, I have argued, by engaging in coherentist reasoning against a background of epistemically responsibility that coherence-driven legal inference may yield justification. Virtue coherentism allows us to exploit the drive toward coherence that is characteristic of our reasoning processes in ways that are not epistemically objectionable.

The discussion over coherentism in the last decades has had an important impact on the way in which key aspects of legal reasoning are theoited and, more broadly, on the way in which the field of legal theorizing should be conceived. The proposal of coherentist approaches has propitiated vigorous debates on important topics such as the proper scope of judicial discretion, the relevance of moral reasons in legal argument, the limits of rule-based approaches to legal reasoning, and the role of moral and legal principles within a theory of the sources of law. It has also put forward a conception of the structure of legal justification and knowledge alternative to the foundationalist—pyramidal—one that is traditionally assumed. In addition, the advancement of legal coherentism has also contributed to strengthen the connections between legal theory and A.I., build bridges between analytical philosophy of law and narrative theory, and integrate the philosophical and psychological aspects of legal decision-making. Thus, regardless of the merits of legal coherentism as a theory of law and adjudication, the advent of coherentist theories to the field has left thus far a profound and, I dear to say, long-lasting imprint on the way in which legal theory and legal reasoning are thought of and conducted.

There is much to be done on the coherentist research program in law. In the last years, there has been a revival of the traditional discussion about the connection between coherence and truth prompted by the use of probability theory and other formal tools. Only recently, have these results being applied specifically to law. The development of this area of research and its legal applications would shed further

[48] It is critical to note that a certain dose of 'incoherence' is also valuable. This does not detract from the plausibility of coherentism, as the benefits of incoherence may be recognized and accounted for from within a coherentist framework. See Haack (2004), Lariguet (2011), and Amaya (2017).

light on the knotted problem of how coherence and truth in law connect up. While there has been important work being done on the role of emotions in a coherence theory of legal reasoning, it would be desirable to have more detailed account of how coherence and emotion may be integrated in legal argument. Another interesting topic for further research is the connection between coherence studies and studies on defeasibility. More specifically, work on the concept of incoherence and its role in legal argument would significantly contribute to current research on defeasibility in law. Most work on legal coherentism assumes the individualistic perspective that characterizes traditional epistemology. It would be necessary to explore the mechanisms whereby coherence emerges in the course of collective decision-making and the relevant connections that may be established between coherence and consensus. Finally, applications of legal coherentism to specific areas of the law need to be further developed. Such applications are critical for devising a more nuanced account of the inner workings, value, and limits of arguments from coherence in law.

References

Alcoff, L. 1996. *New versions of the coherence theory*. Ithaca: Cornell University Press.

Alcoff, L. 2001. The case of coherence. In *The nature of truth*, ed. M. Lynch, 159–183. Cambridge: MIT Press.

Alexy, R. 1998. Coherence and argumentation or de genuine twin criterialess super criterion. In *On coherence theory in law*, ed. A. Aarnio, et al., 41–49. Lund: Juristförlager i Lund.

Alexy, R., and A. Peczenik. 1990. The concept of coherence and its significance for discursive rationality. *Ratio Juris* 3: 130–147.

Alonso, J.P. 2006. *Interpretación de las normas y derecho penal*. Buenos Aires: Editores del Puerto.

Amaya, A. 2015. *The tapestry of reason: An inquiry into the nature of coherence and its role in legal argument*. Oxford: Hart Publishing.

Amaya, A. 2007. Formal models of coherence and legal epistemology. *Artificial Intelligence and Law* 15: 429–447.

Amaya, A. 2011. Legal justification by optimal coherence. *Ratio Juris* 24: 304–329.

Amaya, A. 2017. *The tapestry of reason*. Oxford: Hart Publishing.

Angere, S. 2007. The defeasible nature of coherentist justification. *Synthese* 157: 321–335.

Angere, S. 2008. Coherence as heuristic. *Mind* 117: 1–26.

Araszkiewicz, M. 2010. Balancing of legal principles and constraint satisfaction. In *Legal knowledge and information systems*, ed. R.G.F. Winkels. Amsterdam: IOS.

Araszkiewicz, M. 2012. Coherence-based account of the doctrine of consistent interpretation. In *AI approaches to the complexity of legal systems*, ed. M. Palmirani, et al. Berlin: Springer.

Araszkiewicz, M., and J. Savelka. 2012. Two methods for representing judicial reasoning in the framework of coherence as constraint satisfaction. In *Legal knowledge and information systems*. JURIX: The twenty-fourth annual conference, ed. K. Atkinson. Amsterdam: IOS Press.

Araszkiewicz, M. 2013. Limits of constraint satisfaction theory of coherence as a theory of (legal) reasoning. In *Coherence: Insights from philosophy, jurisprudence, and artificial intelligence*, ed. M. Araszkiewicz, and J. Šavelka, 217–243. Dordrecht: Springer.

Atienza, M. 2006. *Derecho como argumentación*. Barcelona: Ariel.

Audi, R. 1988. Foundationalism, coherentism, and epistemological dogmatism. *Philosophical Perspectives* 2: 407–442.

Audi, R. 1993. *The structure of justification*. Cambridge: Cambridge University Press.

Baehr, J. 2008. Four varieties of character-based virtue epistemology. *The Southern Journal of Philosophy* 46: 469–502.

Baehr, J. 2009. Evidentialism, vice, and virtue. *Philosophy and Phenomenological Research* LXXVIII: 545–567.

Bench-Capon, T., and G. Sartor. 2001a. Theory based explanation of case law domains. In *Artificial intelligence, proceedings of the eighth international conference on artificial intelligence and law.* New York: ACM Press.

Bench-Capon, T., and G. Sartor. 2001b. A quantitative approach to theory coherence. In *Legal knowledge and information systems.* JURIX 2001: The fourteenth annual conference, ed. B. Verheij et al. Amsterdam: IOS.

Bench-Capon, T., and G. Sartor. 2003. A model of legal reasoning with cases incorporating theories and values. *Artificial Intelligence* 150: 97–143.

Bender, J. 1989. Coherence, justification, and knowledge: The current debate. In *The current state of the coherence theory: Critical essays on the epistemic theories of Keith Lehrer and Laurence BonJour, with replies,* ed. J. Bender, 1–14. Dordrecht, Kluwer.

BonJour, L. 1985. *The structure of empirical knowledge*. Cambridge: Cambridge University Press.

BonJour, L. 1998. *In defense of pure reason*. Cambridge: Cambridge University Press.

Bratman, M.E. 1987. *Intention, plans, and practical reasoning*. Cambridge: Harvard University Press.

Brink, D. 1989. *Moral realism and the foundations of ethics*. Cambridge: Cambridge University Press.

Chisholm, R.M. 1977. *Theory of knowledge*, 2nd ed. Englewood Cliffs: Prentice-Hall.

Code, L. 1987. *Epistemic responsibility*. Hanover: University Press of New England.

Coleman, J. 2001. *The practice of principle*. New York: Oxford University Press.

Coleman, J., and B. Leiter. 1993. Determinacy, objectivity and authority. *University of Pennsylvania Law Review* 142: 549–637.

Cooper, N. 1994. Understanding. *Proceedings of the Aristotelian Society* 68 Supp: 1–26.

Davidson, D. 2001. *Subjective, intersubjective, objective*. Oxford: Oxford University Press.

DePaul, M. 1993. *Balance and refinement: Beyond coherence methods of moral inquiry*. London: Routledge.

Dietrich, F., and L. Moretti. 2005. On coherent sets and the transmission of confirmation. *Philosophy of Science* 72: 403–424.

Dworkin, R. 1972. No right answer? In *Law, morality and society,* ed. P.M.S. Hacker, and J. Raz, 58–84. Oxford: Clarendon Press.

Dworkin, R. 1977. *Taking rights seriously*. Cambridge: Harvard University Press.

Dworkin, R. 1983a. A reply by Ronald Dworkin. In *Ronald Dworkin and contemporary jurisprudence*, ed. M. Cohen, 247–300. Totowa: Rowman and Littlefield.

Dworkin, R. 1983b. My reply to Stanley Fish (and Walter Benn Michaels): Please don't talk about objectivity any more. In *The politics of interpretation*, ed. W.J.T. Mitchell, 287–313. London: University of Chicago Press.

Dworkin, R. 1985. *A matter of principle*. Cambridge: Harvard University Press.

Dworkin, R. 1986. *Law's empire*. London: Fontana.

Dworkin, R. 1993. Natural Law revisited. In *Readings in the philosophy of law*, ed. J. Arthur, and W.H. Shaw. New Jersey: Prentice Hall.

Dworkin, R. 1996. Objectivity and truth: You'd better believe it. *Philosophy & Public Affairs* 25: 87–139.

Edmunson, W.A. 1996. The antinomy of coherence and determinancy. *Iowa Law Review* 82: 1–20.

Feldman, R. 1988. Epistemic obligations. *Philosophical Perspectives* 2: 235–256.

Feldman, R. 2002. Epistemological duties. In *The oxford handbook of epistemology*, ed. P. Moser, 362–385. Oxford: Oxford University Press.

Finnis, J. 1987. On reason and authority in law's empire. *Law and Philosophy* 6: 173–199.

Fuller, L.L. 1969. *The morality of law*, Rev edn. New Haven: Yale University Press.

Goldman, A.H. 1988. *Moral knowledge*. London: Routledge.

Goldman, A.H. 1989. Legal reasoning as a model for moral reasoning. *Law and Philosophy* 8: 131–149.

Goldman, A.H. 2002. *Practical rules: When we need them and when we don't*. Cambridge: Cambridge University Press.

Greco, J. 2002. Virtues in epistemology. In *The Oxford handbook of epistemology*, ed. P.K. Moser, 287–316. Oxford: Oxford University Press.

Günther, K. 1989. A normative conception of coherence for a discursive theory of legal justication. *Ratio Juris* 2: 155–166.

Günther, K. 1993. *The sense of appropriateness: Application discourses in morality and law*. Albany: State University of New York Press.

Haack, S. 2000. A founherentist theory of empirical justification. In *Epistemology: An antholog*, ed. E. Sosa, and J. Kim, 226–237. Malden: Blackwell.

Haack, S. 2004. Coherence, consistency, congruity, cohesiveness & c.: Remain calm! Don't go overboard! *New Literary History* 35: 167–173.

Habermas, J. 1996. *Between facts and norms*. Cambridge: MIT Press.

Hage, J.C. 2001. Formalizing legal coherence. In *Proceedings of the 8th international conference on artificial intelligence and law*. New York: ACM.

Hage, J.C. 2004. Law and coherence. *Ratio Juris* 17: 87–105.

Hage, J.C. 2013. Three kinds of coherentism. In *Coherence: Insights from philosophy, jurisprudence and artificial intelligence*, ed. M. Araszkiewicz, and J. Šavelka, 1–33. Dordrecht: Springer.

Hage, J.C., and A. Peczenik. 2000. Law, morals, and defeasibility. *Ratio Juris* 13: 305–325.

Hansson, S.O. 2006. Coherence in epistemology and belief revision. *Philosophical Studies* 128: 93–108.

Harman, G. 1980. Reasoning and explanatory coherence. *American Philosophical Quarterly* 17: 151–157.

Harman, G. 1986. *Change in view: Principles of reasoning*. Cambridge: MIT Press.

Hellman, C. 1995. The notion of coherence in discourse. In *Focus and coherence in discourse processing*, ed. G. Rickheit, and C. Habel, 190–202. Berlin: DeGruyter.

Hoffmaster, B. 1980. A holistic approach to judicial justification. *Erkenntnis* 15: 159–181.

Hurley, S.L. 1989. *Natural reasons: Personality and polity*. Oxford: Oxford University Press.

Hurley, S.L. 1990. Coherence, hypothetical cases and precedent. *Oxford Journal of Legal Studies* 10: 221–251.

Joseph, S., and H. Prakken. 2009. Coherence-driven argumentation to norm consensus. In *Proceedings of the 12th international conference on artificial intelligence and law.* ACM: New York.

Kennedy, D. 1997. *A critique of adjudication (Fin de siècle)*. Cambridge: Harvard University Press.

Khalifa, K. 2010. Default privilege and bad lots: Underconsideration and explanatory inference. *International Studies in the Philosophy of Science* 24: 91–105.

Kress, K. 1984. Legal reasoning and coherence theories: Dworkin's right thesis, retroactivity, and the linear order of decisions. *California Law Review* 72: 369–402.

Kress, K. 1994. Coherence and formalism. *Harvard Journal of Law and Public Policy* 16: 639–682.

Kress, K. 1996. Coherence. In *A companion to philosophy of law and legal theory,* ed. D. Patterson, 521–539. Cambridge: Blackwell.

Kurzban, R., and C.A. Aktipis. 2007. Modularity and the social mind: Are psychologists too selfish? *Personality and Social Psychology Review* 11: 131–149.

Lariguet, G. 2011. Todo lo que usted quería saber sobre la coherencia y no se atrevió a preguntarle a Amalia Amaya. *Discusiones* X: 87–139.

Laudan, L. 1977. *Progress and its problems: Towards a theory of scientific growth*. Berkeley and Los Angeles: University of California Press.

Lehrer, K. 2000. *Theory of knowledge*, 2nd ed. Boulder: Westview Press.

Lipton, P. 2004. *Inference to the best explanation,* 2nd edn. London and New York: Routledge. 1st edn (1991).

Lycan, W.G. 1988. *Judgment and justification*. New York: Cambridge University Press.

Lynch, D. (ed.). 2001. *The nature of truth*. Cambridge: MIT.

MacCormick, N. 1983. Dworkin as a Pre-Benthamite. In *Ronald Dworkin and contemporary jurisprudence,* ed. M. Cohen, 182–204. Totowa: Rowman and Allanheld.

MacCormick, N. 1984. Coherence in legal justification. In *Theory of legal science,* ed. A. Peczenik, L. Lindahl, and B. van Roermund. Dordrecht: Reidel. A revised version was published in *Theorie der Normen. Festgabe für Ota Weinberger zum*, ed. W. Krawietz et al. 1984. Berlin: Duncker and Humblot.

MacCormick, N. 1993. Argumentation and interpretation in law. *Ratio Juris* 6: 16–29.

MacCormick, N. 1994. *Legal reasoning and legal theory*. Oxford: Clarendon Press.

MacCormick, N. 2005. *Rhetoric and the rule of law: A theory of legal reasoning*. Oxford: Oxford University Press.

Mackie, J. 1983. The third theory of the law. In *Ronald Dworkin and contemporary jurisprudence*, ed. M. Cohen, 172–181. Totowa: Rowman and Allanheld.

Marmor, A. 1991. Coherence, holism, and interpretation: The epistemic foundations of Dworkin's legal theory. *Law and Philosophy* 10: 383–412.

McGarry, J. 2013. The possibility and value of coherence. *Liverpool Law Review* 34: 17–26.

Mercier, H. 2012. The social functions of explicit coherence evaluation. *Mind & Society* 11: 81–92.

Mercier, H., and D. Sperber. 2011. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences* 34: 57–74.

Michelon, C. 2011. Princípios e Coerência na Argumentação Jurídica. In *Direito e Interpretação: Racionalidade e Instituições*, ed. C. Barbieri, and R.P. Macedo Jr., 261–285. São Paolo: Saraiva.

Millgram, E., and P. Thagard. 1996. Deliberative coherence. *Synthese* 108: 63–88.

Montmarquet, J. 1993. *Epistemic virtue and doxastic responsibility*. Lanham: Rowman and Littlefield.

Moore, M. 1985. A natural law theory of interpretation. *California Law Review* 85: 459–475.

Moore, M. 2004. *Objectivity in ethics and law*. Dartmouth: Ashgate.

Moral, L. 2003. A modest notion of coherence in legal reasoning: A model for the European court of justice. *Ratio Juris* 16: 296–323.

Moretti, L. 2007. Ways in which coherence is confirmation conducive. *Synthese* 157: 309–319.

Olsson, E.J. 1997. A coherence interpretation of semi-revision. *Theoria* 63: 105–134.

Olsson, E.J. 1998. Making beliefs coherent. *Journal of Logic, Language and Information* 7: 143–163.

Olsson, E.J. 2002. What is the problem of coherence and truth? *The Journal of Philosophy* 99: 246–272.

Olsson, E.J. 2005. *Against coherence: Truth, probability, and justification*. Oxford: Oxford University Press.

Olsson, E.J., and S. Schubert. 2007. Reliability conducive measures of coherence. *Synthese* 157: 297–308.

Österman, B. 2001. Is there a general theory of understanding? *Acta Philosophica Fennica* 69: 43–54.

Pastore, B. 1991. Sistemi Normativi e Coerenza. In *Sistemi normativi statici e dinamici*, ed. L. Gianformaggio. Torino: Giappichelli.

Peczenik, A. 1990. Coherence, truth, and rightness in the law. In *Law, interpretation, and reality: Essays in epistemology, hermeneutics, and jurisprudence*, ed. P. Nerhot, 265–309. Dordrecht: Kluwer.

Peczenik, A. 1994. Law, morality, coherence and truth. *Ratio Juris* 7: 146–176.

Peczenik, A. 1998. A coherence theory of juristic knowledge. In *on coherence theory of law*, ed. A. Aarnio, et al. Lund: Juristförlager i Lund.

Peczenik, A. 1999. The passion for reason. In *The law in philosophical perspective*, ed. L.J. Wintgens, 173–223. Dordrecht: Kluwer.

Peczenik, A. 2000a. Certainty or coherence? In *The reasonable as rational? On Legal Argumentation And Justification. Festschrift for Aulis Aarnio,* ed. W. Krawietz et al. Berlin: Duncker and Humblot.

Peczenik, A. 2000b. Scientia Iuris—An unsolved philosophical problem. *Ethical Theory and Moral Practice* 2: 273–302.

Peczenik, A. 2004. Can philosophy help legal doctrine? *Ratio Juris* 17: 10–117.

Peczenik, A. 2009. *On law and reason*, 2nd ed. Dordrecht: Springer.

Peczenik, A., and J.C. Hage. 2004. Legal knowledge of what? *Ratio Iuris* 13: 326–345.

Pérez-Bermejo, J.M. 2006. *Coherencia y sistema jurídico*. Madrid-Barcelona: Marcial Pons.

Pérez-Bermejo, J.M. 2007. Alcune osservazioni sul valore della coerenza nei sistemi giuridici. *Diritto e Questioni Pubbliche* 7: 43–59.

Pino, G. 1998. Coerenza e verità nell'argomentazione giuridica. Alcune riflessioni. *Rivista Internazionale di Filosofia del Diritto* 1: 84–126.

Plantinga, A. 1993. *Warrant: The current debate*. Oxford: Oxford University Press.

Pollock, J. 1974. *Knowledge and justification*. Princeton: Princeton University Press.

Pollock, J. 1986. *Contemporary theories of knowledge*. Totowa: Rowman and Littlefield.

Psillos, S. 2002. Simply the best: A case for abduction. In *Computational logic*, ed. A.C. Kakas, and F. Sadri, 605–625. Berlin: Springer.

Putnam, H. 1981. *Reason, truth and history*. Cambridge: Harvard University Press.

Rabinowicz, W. 1998. Peczenik's passionate reason. In *On coherence theory of law*, ed. A. Aarnio, et al. Lund: Juristsförlaget i Lund.

Ratti, G.B. 2007. La coerentizzazione dei sistemi giuridici. *Diritto e Questioni Pubbliche* 7: 61–70.

Rawls, J. 1999. *A theory of justice*, rev ed. Cambridge: Harvard University Press.

Raz, J. 1985. Authority, law, and morality. *The Monist* 68: 295–324.

Raz, J. 1986. Dworkin: A new link in the chain. *California Law Review* 74: 1103–1119.

Raz, J. 1992. The relevance of coherence. *Boston University Law Review* 72: 273–321.

Rescher, N. 1985. Truth as ideal coherence. *Review of Metaphysics* 38: 795–806.

Richardson, H. 1994. *Practical reasoning about final ends*. Cambridge: Cambridge University Press.

Rodriguez-Blanco, V. 2001. A revision of the constitutive and epistemic coherence theories in law. *Ratio Iuris* 14: 212–232.

Sartorious, R. 1968. The justification of the judicial decision. *Ethics* 78: 171–187.

Sartorious, R. 1971. Social policy and judicial legislation. *American Philosophical Quarterly* 8: 151–160.

Šavelka, J. 2013. Coherence as constraint satisfaction: Judicial reasoning support mechanism. In *Coherence, insights from philosophy, jurisprudence, and artificial intelligence*, ed. M. Araszkiewicz, and J. Šavelka, 203–217. Dordrecht: Springer.

Schröter, M.W. 2006. European legal reasoning: A coherence-based approach. *ARSP* 92: 86–89.

Schuaer, F. 1986–87. The jurisprudence of reasons. *Michigan Law Review* 85: 847–868.

Schubert, S. 2011. Coherence reasoning and reliability: A defense of Shogenji's measure. *Synthese* 187: 305–319.

Schubert, S. 2012. Is coherence conducive to reliability? *Synthese* 187: 607–621.

Schubert, S., and E. Olsson. 2013. Coherence and reliability in judicial reasoning. In *Coherence, insights from philosophy, jurisprudence, and artificial intelligence*, ed. M. Araszkiewicz, and J. Šavelka, 33–59. Dordrecht: Springer.

Shogenji, T. 1999. Is coherence truth-conducive? *Analysis* 59: 338–345.

Simon, D. 2004. A third view of the black box: Cognitive coherence in legal decision-making. *The University of Chicago Law Review* 71: 511–586.

Simon, D., and N. Scurich. 2011. Lay judgments of judicial decision making. *Journal of Empirical Legal Studies* 8: 709–727.

Simon, D., L.B. Pham, Q.A. Le, and K.J. Holyoak. 2001. The emergence of coherence over the course of decision-making. *Journal of Experimental Psychology. Learning, Memory, and Cognition* 27: 1250–1260.

Simon, D., D. Stenstrom, and S.J. Read. 2015. The coherence effect: Blending cognition and emotion. *Journal of Personality and Social Psychology* 109: 369–394.

Sintonen, M., and M. Kikeri. 2004. Scientific discovery. In *Handbook of epistemology*, ed. I. Niiniluoto, M. Sintonen, and J. Wolenski, 205–253. Dordrecht: Kluwer.

Thagard, P. 1989. Explanatory coherence. *Behavioral and Brain Sciences* 12: 425–502.

Thagard, P. 1992. *Conceptual revolutions*. Princeton: Princeton University Press.

Thagard, P. 2000. *Coherence in thought and action*. Cambridge: MIT Press.

Thagard, P. 2006. *Hot thought: Mechanisms and applications of emotional cognition*. Cambridge: MIT.

Thagard, P. 2007. Coherence, truth, and the development of scientific knowledge. *Philosophy of Science* 74: 28–47.

Thagard, P. 2012. Coherence: The price is right. *The Southern Journal of Philosophy* 50: 42–49.

Van Fraassen, B.C. 1989. *Laws and symmetry*. Oxford: Clarendon Press.

Wacks, R. 1984. Judges and injustice. *South African Law Journal* 101: 266–285.

Waldron, J. 2008. Did Dworkin ever answer the crits? In *Exploring law's empire: The jurisprudence of Ronald Dworkin*, ed. S. Hershovitz. Oxford: Oxford University Press.

Walker, R.C.S. 1989. *The coherence theory of truth: Realism, anti-realism, idealism*. London: Routledge.

Weinrib, E.J. 1988. Legal formalism: On the immanent rationality of law. *Yale Law Journal* 97: 949–1016.

Weinrib, E.J. 1994. The jurisprudence of legal formalism. *Harvard Journal of Law and Public Policy* 16: 583–596.

Williams, M. 1980. Coherence, justification and truth. *Review of Metaphysics* 34: 243–272.

Williams, B. 1985. *Ethics and the limits of philosophy*. Cambridge: Harvard University Press.

Wintgens, L. 1993. Coherence of the law. *ARSP* 79: 483–519.

Wintgens, L. 2000. On coherence and consistency. In *The reasonable as rational? On legal argumentation and justification. Festschrift for Aulis Aarnio,* ed. W. Krawietz et al., 539–550. Berlin: Duncker and Humblot.

Young, J.O. 2001. A defense of the coherence theory of truth. *The Journal of Philosophical Research* 26: 89–101.

Zagzebski, L. 1996. *Virtues of the mind: An inquiry into the nature of virtue and the ethical foundations of knowledge*. Cambridge: Cambridge University Press.

Ziff, P. 1984. Coherence. *Linguistics and philosophy* 7: 31–42.

Precedent and Legal Analogy

Kevin D. Ashley

1 Introduction

This chapter provides guidance[1] in reasoning from precedent and by legal analogy as practiced in a common law context. The guidance is intended primarily for law students, although practitioners and legal theorists may benefit, as well. The legal profession, it has been observed, "spend[s] relatively little time refining general methods for discriminating between good patterns of reasoning and bad […] We seldom develop general accounts that explain to students how lawyers ought to reason and why." (Walker 2007, 1687f). In an effort to address that deficit, the ultimate goals of this work are to identify good patterns of argument associated with precedent and legal analogy and to distill them into argument schema that students, practitioners, and those who develop computational models of legal reasoning can implement.

As developed more fully (and as qualified) below, in reasoning from precedent, a rule derived from a source case, the precedent, applies by its terms to the facts of a new case, the target case, resulting in a presumption that a court should apply the rule to determine the outcome of the target case. In legal analogy, the rule of the source case does *not* apply by its terms to the target case, but the court finds good reasons to decide the target case in the same way as the source case, and generalizes the rule of the source case to extend to both. Court opinions, justifying whether or

[1] According to the *OED*, a "handbook" is "a compendious book or treatise for guidance in any art, occupation, or study." Given Holmes' characterization of the law as "prediction, the prediction of the incidence of the public force through the instrumentality of the courts," the OED's alternative definition of handbook as "a betting-book," that is, a book for bookmaking, may also be *apropos* (Holmes 1897).

K. D. Ashley (✉)
School of Law and Graduate Program in Intelligent Systems, University of Pittsburgh, Pittsburgh, PA, USA
e-mail: ashley@pitt.edu

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_23

not and how to apply a precedent or a legal analogy, evidence a variety of patterns of reasoning and argumentation.

Among those legal scholars who have spent time discriminating good from bad patterns of reasoning with precedent and legal analogy are the authors whose work is discussed in this chapter, especially Melvin Eisenberg, whose theory addresses the extent to which precedents are binding; Scott Brewer, who provides explicit patterns and interpretive schemes for exemplary legal argument, including reasoning from precedent and legal analogy;[2] Fred Schauer, who focuses on the constraints affecting the force of precedent implicit in fact characterization; and Robert Summers, who surveys courts' methods of interpreting precedent in a major common law jurisdiction. These and other examples in the jurisprudential literature provide some useful guidance for students and practitioners in making and responding to arguments from precedent and legal analogy and serve as a basis for distilling argument schemes.

Recent work in argumentation theory and Artificial Intelligence and Law (AI and Law) on argument schemes provides a complementary source of guidance. Researchers in these fields have employed argument schemes in order to "discriminat[e] between good patterns of reasoning and bad" in the legal domain (and others) (Walker 2007, 1687f; Walton 1996). An argument scheme corresponds to a typical framework in the domain for making an inference sanctioned by the argument; it corresponds to a kind of *prima facie* reason for believing the argument's conclusion (Prakken 2005, 5). The scheme often embodies a principle of practical reasoning that underlies the inference. The scheme also may include *critical questions* that affect the scheme's applicability or the believability of the sanctioned inference.[3]

Argument schemes can be thought of as general blueprints of arguments made in a specific context, for instance, a legal argument from precedent or by analogy. They provide a set of abstract criteria that are necessary for the argument to work. For example, in order to be persuasive, an argument by analogy needs a previously decided legal case (i.e., a source case), a measure of similarity to the current fact situation (or target case), and reasons why the similarities matter to a legal conclusion about the target case. When used to generate an argument, this skeletal generic scheme can be fleshed out with specific content to form a valid legal argument. Then, in a second step, this argument can be assessed qualitatively by asking certain critical questions from a predefined list of challenges that correspond to the specific scheme used to create the argument. For example, arguments by analogy have spe-

[2] Scott Brewer speaks frequently of patterns and interpretive schema of analogical and disanalogical argument (Brewer 1996, 962–966, n. 35, 1009f, 1016, 1021). Cass Sunstein speaks of analogies as a kind of patterned reasoning: "Their [i.e., analogies'] meaning lies in their use. They are not simply unanalyzed fact patterns; They are used to help people think through contested cases and to generate low-level principles. In this way they have a constitutive dimension, for the patterns we see are a product not simply of preexisting reality, but of our cognitive structures and our principles as well. The principles and patterns we develop and describe are in turn brought to bear on, and tested through confrontation with, other cases" (Sunstein 1993, 778–779, n. 129).

[3] "[A]rgument schemes can be formalised as prima facie reasons, [...] applications of schemes resulting in opposite conclusions can be regarded as rebuttals, while negative answers to critical questions about exceptional circumstances correspond to undercutters" (Prakken 2005, 5).

cific vulnerabilities with which they will typically be challenged, like insufficient or merely superficial similarity to the target case, the lack of an underlying reason why the similarities matter legally, or an assertion that the source case is no longer good law. Depending on the answers, critical questions can lead to counterarguments and surrebuttals.

This chapter asks whether approaches in argumentation theory can be applied to develop more detailed and, possibly, given the purpose of a Handbook, more useful accounts of reasoning with precedents and legal analogies. For instance, it can flesh out the critical questions and resulting counterarguments that arise in reasoning with precedents and legal analogies. The focus on argument schemes and patterns of justification in the works discussed below on applied legal philosophy, argumentation theory, and AI and Law addresses precedent and legal analogy from a somewhat broader vantage point than one often encounters in legal theory. Instead of focusing exclusively on the force of precedent from a theoretical viewpoint, the question is, given a theory (or theories) of precedent or legal analogy, what kinds of arguments should advocates make concerning the force of a precedent in a given context, what kinds of responsive arguments are reasonable, and how should courts evaluate the arguments and responses.

The chapter begins with brief accounts of Eisenberg's and Schauer's jurisprudential theories of the constraints that "following precedent" imposes on a court's discretion and of Brewer's and others' theories of legal analogy. In light of the theoretical accounts of precedent and analogy and based on Summers's account, the chapter then surveys the kinds of legal arguments one may use to convince a court to follow a precedent, or analogy, or not to do so. Then, the chapter turns to how to represent these patterns of reasoning in terms of argument schema. It focuses on particular argument schema for reasoning with precedent and for legal analogy that have been offered by AI and Law researchers. Since these researchers intend to enable computer programs to construct arguments from precedent and legal analogy, they specify detailed formalisms for representing the argument schema. With an eye toward accommodating a middle ground between jurisprudential and computational accounts, the discussion in this chapter eschews formalisms in favor of succinct textual descriptions of the argument schema as applied in some realistic legal examples.

Finally, the chapter briefly examines differences in such argument schemes as they are applied across the contexts of common law reasoning, statutory interpretation, and constitutional questions.

The goal is to bring jurisprudential theories and descriptive accounts of precedent and legal analogy, philosophical accounts of practical reasoning in terms of argument schemes and critical questions, and computational models of analogical legal reasoning together in a way that should be instructive for law students, but also has practical ramifications for the fields of AI and Law and of Jurisprudence.

2 Reasoning from Precedent

Reasoning from precedent, a characteristic common law mode of reasoning, is rooted in the principle of *stare decisis* and related to legal analogy. In 1988, Melvin Eisenberg defined both of these reasoning modes in *The Nature of the Common Law*.

2.1 Nature of Precedential Constraint

According to (Eisenberg 1988, 47), under the principle of *stare decisis*, a precedent's "holding" or "rule" (i.e., its *ratio decidendi* or ground of decision) binds courts in subsequent cases "if the precedent satisfies certain formal conditions, such as having been rendered by a court at a designated level in the relevant jurisdiction."

When a court reasons from precedent, however, it has to determine in a particular context how much its discretion is constrained, formally and substantively, by the requirement to "follow precedent." In Eisenberg's view, this raises two interdependent questions: "(1) How does a court faced with a precedent determine what rule the precedent stands for? (2) What is meant by the concept, expressed in the principle of *stare decisis*, that the rule of a precedent is binding?" (Eisenberg 1988, 51). In answering the first question, a court can begin with the rule announced in the precedent or construe a new rule out of the precedent's facts and result. Which of these a court chooses depends on the answer to the second question, what does "binding" mean? (Eisenberg 1988, 55).

Eisenberg breaks the bindingness of a precedent into three constraints:

(1) Formal constraint that "the rule established by the deciding court must be reconcilable with the result reached in the precedent,"
(2) Second formal constraint that "the deciding court must either follow the precedent or distinguish it," and
(3) Substantive constraint that "the announced rule of a precedent should be applied and extended to new cases if the rule substantially satisfies the standard of social congruence and a failure to apply or extend the rule to a new case would not be justified by applicable social propositions, given the social propositions that support the rule" (Eisenberg 1988, 161–164).

The substantive constraint is a key aspect of Eisenberg's theory of reasoning with precedent: Determining if the precedent is binding involves reasoning about *applicable social propositions*, "those moral norms, policies, and experiential propositions that it is proper for a court to employ," including "usages (experiential propositions about how the world works in a relevant subgroup)" (Eisenberg 1988, 43). "[S]ocial propositions always figure in determining the rules the courts establish and the way in which those rules are extended, restricted, and applied" (Eisenberg 1988, 3).

From this, one can see that, in Eisenberg's view, the binding quality of precedents is not ironclad. "[I]f the rule announced in a precedent substantially satisfies the standards of social congruence and systemic consistency, it should be consistently applied and extended even though another rule would be marginally better" (Eisenberg 1988, 75). The *standard of social congruence* is an ideal "that the body of rules that make up the law should correspond to the body of legal rules that one would arrive at by giving appropriate weight to all applicable social propositions and making the best choices where such propositions collide" (Eisenberg 1988, 44). In other words, if the court determines in light of applicable social propositions that there is no (or not sufficient) good reason to treat the case at hand differently, it should *follow* the precedent and apply its rule to the new case.

If, however, the balance of applicable social propositions favors *not* applying the rule of the precedent, the court may distinguish it, formulating an exception. "In *distinguishing*, a court normally begins with a rule, announced in a prior case, that is in terms applicable to the case at hand, and then determines that there is good reason to treat the case at hand differently. The court therefore reformulates the announced rule (or, what is the same thing, formulates an exception) in a way that requires the two cases to be treated differently" (Eisenberg 1988, 87).

In extreme situations, where applying the rule of the precedent is seriously out of step with applicable social propositions, the court may even radically reconstruct or overrule the rule of the precedent. "Whether a deciding court applies, extends, reformulates, radically reconstructs, or overrules an announced rule will always depend in part on whether the rule is socially congruent or incongruent" (Eisenberg 1988, 75).

According to Eisenberg, the decision of *MacPherson v. Buick Motor Co.*[4], discussed at length by Levi (1949, 20–25), is an example of radical reconstruction overruling precedent. The plaintiff had purchased from a dealer an automobile manufactured by Buick with a defective wooden wheel that suddenly collapsed, injuring the plaintiff. According to the then current rule, a negligent manufacturer of a defective product was ordinarily liable only to its immediate buyer (i.e., one with whom the manufacturer had privity of contract) unless the product was "inherently dangerous," an exception that had been applied narrowly if somewhat inconsistently. The policy of the rule, according to Eisenberg, was to protect an emerging industry against extensive liability at a time when, it was assumed, customers relied on the due care of the immediate vendor. "[W]ith the advent of nationwide distribution of brand-name merchandise [however], consumers had begun to rely more on manufacturers than on retailers. Because the rule came to be socially incongruent, it also came to lack consistency with the body of the law." In light of the history of inconsistent New York decisions on the liability of a negligent manufacturer to someone other than a buyer from manufacturer, Justice Cardozo affirmed the judgment for the plaintiff based on a new rule that "transformed the old rule by a radical reconstruction of precedent" (Eisenberg 1988, 60).[5]

[4] 217 N.Y. 382, 111 N.E. 1050 (1916).

[5] "If the nature of a thing is such that it is reasonably certain to place life and limb in peril when negligently made, it is then a thing of danger. Its nature gives warning of the consequences to be expected. If to the element of danger there is added knowledge that the thing will be used by persons

2.2 How Much Precedential Constraint Is Enough?

One consequence of Eisenberg's view is that there are no easy cases. "Social propositions are relevant in all common law cases, and [...] no case is easy in the sense that it can be decided solely on the basis of doctrinal propositions, without the employment of social propositions" (Eisenberg 1988, 75). "Since no two cases are identical, every new case raises an issue whether the rule announced in a precedent can consistently be distinguished," and that "turns on whether applicable social propositions justify different treatment of the two cases, given the social propositions that support the rule of the precedent" (Eisenberg 1988, 75).

Given the courts' obligation to always consider applicable social propositions, and the resulting possibility of distinguishing, radically reformulating, or overruling a precedent, one may ask if this model of reasoning with precedent is consistent with the underlying purpose of *stare decisis*. Subjecting a court's discretion to constraints, Eisenberg points out, serves a number of purposes including the principles of universality and evenhandedness: a court should not "decide a case on the basis of a rule unless it is ready to apply the rule to all similarly situated disputants," and "all other things being equal, once the court has adopted a rule to decide a case it should indeed apply that rule to similarly situated disputants;" that is, "like cases should be treated alike" (Eisenberg 1988, 48).

Fred Schauer has argued that Eisenberg's model of the bindingness of precedent is too weak (Schauer 1989, 470). It would be too easy for courts to avoid applying doctrinal rules embodied in precedent in favor of countervailing social propositions, rendering legal decision-making equivalent with just plain decision-making (hence, his question, "Is the common law law?"). In order to shore up reasoning with precedent, Schauer offers an additional constraint, a presumption in favor of applying the precedent's rule. That is, if a court does not follow a precedent, it should only happen when the weight of inconsistency with social propositions overwhelms a presumption that the doctrinal proposition controls (Schauer 1989, 470). Otherwise, the doctrinal proposition will not be given sufficient weight. Even though, as Eisenberg maintains, the institutionalized professional discourse in which judges engage "mediates their reception of social propositions," thus burnishing the claim to legitimacy of social propositions, Schauer argues that the presumption is needed to redress the imbalance in Eisenberg's account between doctrinal and social propositions.

2.3 Precedential Constraint and Characterization

For Schauer, the dependence of precedent on characterization imposes a constraint that Eisenberg's account tends to gloss over. "Reconcilability" in Eisenberg's first (formal) constraint is, Eisenberg maintains, relatively easy to satisfy since the court

other than the purchaser, and used without new tests, then, irrespective of contract, the manufacturer of this thing of danger is under a duty to make it carefully." 111 N.E. 1050, 1053 (N.Y. 1916).

can characterize the facts of relevant precedents in such a way that their results are consistent with “whatever rule the court proposes to adopt” (Eisenberg 1988, 61).

By contrast, Schauer’s model of precedent emphasizes the dependence of a precedent’s binding constraint on the leeway one has to recharacterize the case facts. There are two questions in deciding a precedent: (1) what characterization to choose in fashioning a legal rule for deciding the facts of the present case in furtherance of current normative goals (i.e., characterization) and (2) in light of the manner in which it can be anticipated new facts will be assimilated into those concepts in the future (Schauer 1987, 574, 577, 579, 580–582, 594). According to Schauer, a “characterization may be a simple word or phrase, as when we characterize a 1957 Chevrolet as a car, a vehicle, or an antique” or “a more complex description of what this event is an example of” (Schauer 1987, 577, n. 12). “[T]he articulated characterization acts like a specifically formulated rule” (Schauer 1987, 581). In this chapter, I will often refer to such articulated characterizations as intermediate legal concepts (also known as legal middle terms) which stand “as a mediating link between the requirements and the [normative] consequences” in legal inference (Lindahl 2004).

In focusing on the constraint on *future* decision-making that a precedent imposes, Schauer posits “rules of relevance” that reflect preexisting linguistic or social categories with which facts have been characterized in the past and constrain how facts may be characterized in the future.[6] A rule of relevance, a kind of principled or standardized “choice among alternative characterizations” (Schauer 1987, 578–579), can be authoritatively stated by the precedent decision maker or reflect preexisting linguistic or social categor[ies]” (Schauer 1987, 585) or “larger categories and rules of language” (Schauer 1987, 588) that impose a presumptive “argumentative burden[s] if [a court] wishes to depart from these categories” (Schauer 1987, 587).[7] Would-be precedent setters need wonder how future decision makers will categorize some aspect of the facts of the current case in terms of a legal concept that will apply more generally to other facts, an issue Schauer refers to as “*assimilation*, how we will group the facts and events of our world. The power of precedent depends upon some assimilation between the event at hand and some other event” (Schauer 1987, 579).

Schauer’s point is that such assimilation of facts into a legal rule’s concept is governed not only by normative concerns but also by linguistic expectations or rules

[6] Schauer says, “a careful study of precedent must confront the extent to which sticky and substantially nonnormative social or linguistic characterizations may impede the ability of a formulator of a principle to draw certain intrinsically sound distinctions or to employ certain intrinsically justifiable groupings” (Schauer 1987, 572, n. 4). He notes that the rules of relevance “are contingent upon both time and culture.” He provides the following example: “Holmes used his [the law of butter-]churn story to show that legal similarity is determined by broad and theory-soaked descriptions like “property” rather than by the nature of the objects involved. Would we draw that conclusion today? The existence of distinct principles for some classes of goods (consumer goods and securities, for example) shows that the rules of relevance in Holmes’ time are not necessarily those of today” (Schauer 1987, 578).

[7] “Whether the characterizing language is treated as holding or dictum, that language cannot absolutely prevent a subsequent interpreter from recharacterizing the first case. But that interpreter must at least confront an argumentative burden not present without an articulated characterization” (Schauer 1987, 580).

about what concepts mean and about what facts can reasonably be equated with what other facts.

> The grouping of trucks and motorcycles as vehicles is a rule of language, not solely within the control of the regulations of the park department. We must distinguish the rule of language from the rule about the kinds of things we want in the park. If we collapse this distinction, we ignore what is most important – the way in which larger categories and rules of language that generate certain groupings significantly constrain the substantive rules in particular regulatory contexts. (Schauer 1987, 587–588)

He gives an example of "a faculty meeting considering a request from a student for an excused absence from an examination in order to attend the funeral of his sister.[…] [T]he student's sister is simultaneously a woman, a sibling, a relative, a blood relative, and one with whom the student has a 'meaningful relationship.' How the relationship is characterized will determine whether later cases will be classed as similar." "Implicit in this objection [that this case establishes a precedent allowing absences for 'funerals of grandparents, aunts, uncles, cousins, nieces, nephews, close friends, and pets'] is a rule of relevance that treats death as a relevant similarity, 'caring' as a relevant similarity, and any distinction between siblings and other meaningful relationships as irrelevant" (Schauer 1987, 578).

Intuitively, this seems to be both true and important. Linguistic and social expectations attached to an advocate's choice of terms for categorizing the facts and formalizing a rule to govern the case condition the ways in which one may characterize new fact situations for purposes of applying the rule, and this conditioning effect seems not to be attributable merely to the meanings of the terms.

For purposes of constructing argument schema for reasoning with precedents, Eisenberg's and Schauer's accounts together suggest the need for critical questions that ask how appropriate it is to apply a precedent's rule given the possibly changed social propositions of a new case and given the underlying linguistic and social expectations obtaining when the precedent was decided.

3 Argument from Legal Analogy

Reasoning by analogy in Eisenberg's model, like reasoning from precedent, involves reasoning about rules in light of applicable social propositions. "In reasoning by analogy, a court normally begins with a rule, announced in a prior case, that is not in terms applicable to the case at hand, and then determines that there is no good reason to treat the case at hand differently. The court therefore reformulates the announced rule (or, what is the same thing, formulates a new rule) in a way that requires the two cases to be treated alike" (Eisenberg 1988, 87).

Reasoning by analogy, a court can extend a rule from a case that the rule covers by its terms to one it does not where "neither applicable social propositions nor a deep doctrinal distinction justifies different treatment of the new cases." The court can also use reasoning to consolidate rules in a consistent manner, "determining that one rule […] should be adopted in preference to a competing rule […] because neither applicable social propositions nor any deep doctrinal distinction justifies adopting [the latter rule] while adhering to some other previously announced rule" (Eisenberg 1988, 93).

As a result, in Eisenberg's view, reasoning from precedent and reasoning by analogy are substantively equivalent. Which mode of reasoning a court pursues depends on how general a rule was announced in the precedent (Eisenberg 1988, 94–96).

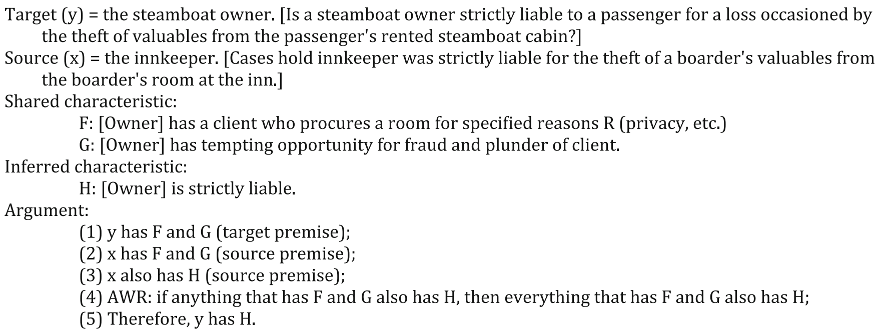
Interestingly, in defining reasoning by analogy, Eisenberg avoids referring to such terms as "relevant similarities" or "relevant differences." He presents his account in stark contrast with Edward Levi's description of common law as "reasoning by example" in which "the finding of similarity or difference is the key step in the legal process" (Levi 1949, 2–6). According to Levi, "The steps are these: similarity is seen between cases; next the rule of law inherent in the first case is announced; then the rule of law is made applicable to the second case. This is a method of reasoning necessary for the law."

By contrast, Eisenberg disparages the suggestion that case comparison can be effective in a normative context. For him, Levi's example of the privity of contract cases that Justice Cardozo reviewed in the *MacPherson* decision, cases involving liability for painter's scaffolds and coffee urns but not for balance wheels and steam boilers, illustrates the futility of reasoning by example. "Reason cannot be used to justify a normative conclusion on the basis of an example without first drawing a maxim or rule from the example (or, what is the same thing, without first concluding that the example 'stands for' a maxim or rule)." According to Eisenberg, reasoning by analogy "does not consist of either comparing the similarities and differences between cases or reasoning by example" (Eisenberg 1988, 87).[8]

Scott Brewer, however, another jurisprudential scholar, does treat legal analogy as exemplary reasoning. In his view, the rule underlying a legal analogy is the source of its rational force. Comparing cases or examples plays a key role in inferring the rule.

Brewer's account, thus, seems to occupy an intermediate position between those of Levi and Eisenberg. In reasoning by analogy, one presents an example (i.e., a source case) to support a conclusion that the current problem (the target case) has a particular legal property and an "analogy-warranting rule" or AWR that specifies "in what its exemplariness consists." The AWR should have a deductive logical structure and serve as a premise, which when applied to the target case (or the source case) deductively entails the conclusion that the target case (source case) has the desired characteristic (Brewer 1996, 971, 975). The AWR identifies the particular, shared features of the source and target that justify the conclusion that the target has the same property as the source. Under this rule, the source case is relevantly similar to the target case with respect to the shared characteristics of each and the inferred characteristic the rule warrants (Brewer 1996, 1015–1016).

[8] On the other hand, as discussed below, at least one of Eisenberg's examples of courts' reasoning by analogy seems to involve analogizing, that is, comparing relevant similarities between cases.



Target (y) = the steamboat owner. [Is a steamboat owner strictly liable to a passenger for a loss occasioned by the theft of valuables from the passenger's rented steamboat cabin?]
Source (x) = the innkeeper. [Cases hold innkeeper was strictly liable for the theft of a boarder's valuables from the boarder's room at the inn.]
Shared characteristic:
F: [Owner] has a client who procures a room for specified reasons R (privacy, etc.)
G: [Owner] has tempting opportunity for fraud and plunder of client.
Inferred characteristic:
H: [Owner] is strictly liable.
Argument:
(1) y has F and G (target premise);
(2) x has F and G (source premise);
(3) x also has H (source premise);
(4) AWR: if anything that has F and G also has H, then everything that has F and G also has H;
(5) Therefore, y has H.

Fig. 1 Brewer's example of legal analogy

In order to be compelling, a legal analogy also requires an analogy-warranting rationale (AWRa) that explains why, in the "eyes of the law," "the logical relation among the characteristics articulated by the analogy-warranting rule either does obtain or should obtain" (Brewer 1996, 965).

As an example, Brewer schematizes a judge's argument that a steamboat owner is strictly liable to a passenger for a loss due to the theft of valuables from the passenger's rented steamboat cabin (Brewer 1996, 1005).

The analogy-warranting rule (Fig. 1, item (4)) subsumes the source case of an innkeeper and the target case of the steamboat owner and states conditions under which an owner (and an innkeeper) is strictly liable. The opinion provides the AWRa: "The principle upon which innkeepers are charged … as insurers … [is that they] should be subjected to a high degree of responsibility in cases where an extraordinary confidence is necessarily reposed in them, and where great temptation to fraud and danger exists by reason of the peculiar relations of the parties" (Brewer 1996, 1004).

In response, an "argument by disanalogy" would involve rewriting the AWR to "impose additional conditions on the rules stated (or implied) in prior cases" (Brewer 1996, 1006, 1011). For instance, if in the cases where innkeepers were held strictly liable, they had not posted warning notices to customers to protect their valuables, and knowing that the steamship owner had posted such notices, a distinguisher might claim an exception E: "Owner failed to post a warning notice,"[9] and provide a disanalogy-warranting rationale (DWRa), perhaps to the effect that innkeepers who fail to post warning notices assume the risk that uninformed, unwary customers will leave valuables in the room for the convenience of thieves.

One may also distinguish a target problem from a competing line of source cases, which had the opposite result. For instance, the steamship owner might attempt to draw an analogy to a line of cases which held "that the owner of a railroad was not strictly liable to railroad passengers who had personal goods stolen from the open-berth sleeping cars on trains" (Brewer 1996, 1013). A distinguisher, arguing against the steamship owner, could maintain that the situation of the railroad is not

[9] This DWR is not in Brewer's example.

analogous to that of the steamship. Unlike a steamship berth, an open booth in a train does not afford the privacy of a room and thus does not present the railroad owner with the same tempting opportunity for fraud and plunder of the client. It follows that the line of railroad cases "does not satisfy the sufficient conditions for the inferred characteristic" H, strict liability.

4 Arguments from Hypotheticals, a Kind of Legal Analogy

Arguments from hypotheticals are an important subset of arguing from legal analogy. "A hypothetical is an imagined situation that involves a hypothesis." In arguing from a hypothetical, an "arguer designs and poses the hypothetical in order to help demonstrate and test [the] consequences [of the arguer's] proposed test or standard for deciding an issue in the case before a court" (i.e., of the arguer's hypothesis.) (Ashley 2009, 323).

Eisenberg regards reasoning from hypotheticals as formally comparable to reasoning by analogy, "since it turns on the question whether two cases can be distinguished" but substantively different, since it "depends on an interplay between applicable social propositions and *conceivable* doctrinal propositions and cases. [emphasis added]" (Eisenberg 1988, 102). Brewer treats arguments from hypotheticals as simply a kind of analogical argument in which the items compared are hypothetical cases rather than authoritative precedents (Brewer 1996, 964–965).

In Eisenberg's model, a court may argue from hypotheticals in order to decide which legal rule to apply to a situation or whether not to apply a rule to a situation. A hypothetical case that is (a) factually different from the problem but that (b) seems easier to decide, and crucially, that is (c) indistinguishable "under applicable social propositions" can help the court decide which rule to apply, namely the rule from the hypothetical or, rather, an extension thereof. A hypothetical case that is (c) indistinguishable from the problem under applicable social propositions to which (d) it "seems clearly improper" to apply a proposed rule helps a court justify not applying the rule to the problem because the rule cannot be generalized to cover the hypothetical (Eisenberg 1988, 99–101; Ashley 2009, 324).

Eisenberg's example, of a court's posing a hypothetical to help decide which rule to apply, is based on a case, *Vincent v. Lake Erie Transportation Co.* 109 Minn. 456, 560 (1910), whose facts are similar to those in *Ploof v. Putnam*, 81 Vt. 471, 71 A. 188 (1908). In *Ploof*, the plaintiff alleged that while sailing, a sudden and violent tempest threatened destruction of his sloop and family. To save them, he moored the sloop to the defendant's dock, but the defendant's servant unmoored the sloop, whereupon "it was driven upon the shore by the tempest," destroying the sloop and its contents and injuring him and his passengers. The plaintiff vessel owner won in *Ploof*. In the *Vincent* case, however, the defendant vessel owner, Lake Erie Transportation, had tied its vessel to Vincent's dock to unload cargo. When a violent storm arose, the defendant kept the vessel at the dock for safety's sake, but the wave-rocked vessel damaged the dock. Vincent, the dock owner, sued the vessel owner for damages and won.

In concluding that the defendant vessel owner was liable, the court in Vincent drew an analogy to *Ploof*. "If, in [the *Ploof*] case, the vessel had been permitted to remain, and the dock had suffered an injury, we believe the shipowner would have been held liable for the injury done" (109 Minn. at 460). In support of this conclusion, the court posed the following hypothetical: "Let us imagine in this case that for the better mooring of the vessel those in charge of her had appropriated a valuable cable lying upon the dock. No matter how justifiable such appropriation might have been, it would not be claimed that, because of the overwhelming necessity of the situation, the owner of the cable could not recover its value" (109 Minn. at 460). Since the hypothetical case was (a) factually different from the problem but (b) seemed easier to decide, and was (c) indistinguishable under applicable social propositions, the court, in extending the rule from the hypothetical to the problem, decided that the defendant vessel owner was liable for the damage to plaintiff's dock.

Similarly, in Brewer's model "heuristically well-chosen" hypothetical source cases can help the reasoner search for an analogy-warranting rule to propose and confirm that that AWR "effects an acceptable sorting of a range of particular items, actual or hypothetical, thought relevant by the legal reasoner" (Brewer 1996, 1021–1022).

In elaborating the *Vincent* court's reasoning patterns in the above example, Eisenberg explicates inference steps that the court left implicit. His interpretation of the court's reasoning seems entirely plausible. It also illustrates how reasoning with hypotheticals serves as a kind of shorthand for reasoning about a proposed rule's or decision's consistency with underlying values, but without requiring much explicit discussion of values in the abstract. Instead, the reasoning focuses on comparing the rule's or decision's effects on values in specific factual scenarios. In a related way, case analogies flesh out the content of abstract principles without necessarily refining the language in which the principle is expressed as a proposition. SIROCCO (McLaren 2003) is an example of an AI and Law program that modeled how decided cases incrementally elaborate the meanings of abstract principles, information it used to improve its retrieval effectiveness in objectively measurable ways.

5 Roles of Differences, Similarities, and Rules in Precedent and Legal Analogy

Two aspects of the jurisprudential accounts of precedent and legal analogy deserve further discussion: the roles of relevant similarities and differences and of rules.

From a practitioner's viewpoint, it is commonly thought that both reasoning from precedent and legal analogy involve comparing a problem scenario and cases in terms of relevant similarities and differences. At least, that is how law students are taught.[10]

[10] According to a widely used instructional text on legal reasoning, "An *analogy* shows that two situations are so parallel that the reasoning that justified the decision in one should do the same

With respect to distinguishing, jurisprudential accounts jibe with practitioners' intuitions: A common way to respond to arguments from precedent and legal analogy is to point out relevant differences, in Brewer's sense of the term, between the current problem and a cited case. As Brewer notes, Eisenberg recognizes distinguishing as a type of argument by analogy: "At its core, reasoning by analogy is the mirror image of the process of distinguishing" (Eisenberg 1988, 87; Brewer 1996, 1011, n. 254).

Pointing out relevant *similarities* between the current problem and cited case, however, seems to be somewhat more controversial, at least from Eisenberg's viewpoint. As noted, unlike Schauer's or Brewer's models, neither Eisenberg's model of reasoning with precedent nor his model of legal analogy explicitly refers to pointing out relevant similarities, although they seem to do so implicitly. In Eisenberg's model of legal analogy, the need to determine whether applicable social propositions or a deep doctrinal distinction justifies different treatment of the new cases seems likely to involve comparing the application of social propositions in the source and target cases (i.e., whether they are similar or different). That is, the need to determine whether applicable social propositions warrant *not* applying the rule of the precedent seems likely to lead to a consideration of whether applicable social propositions that warranted applying the rule *in* the precedent apply as well to the current case, that is, whether the precedent and current case are analogous (i.e., similar) in this respect and the types of facts that make them so.

At least one of Eisenberg's examples of legal analogy, the *Ploof* case, described above, seems to belie his attempt to circumvent the practical reality of judicial case comparisons involving assessing relevant similarities. It may be true, as Eisenberg maintains, that the court in *Ploof* reasoned about the inconsistency of having one rule denying landowners damages for unauthorized entry to save a life and another rule allowing landowners to eject the intruder,[11] but the court does not say so explicitly. Instead, the court in *Ploof* seems to have based its decision for the vessel owner on a factually analogous case. After citing some principles,[12] the court recites the facts of a

in the other. [...] *Distinguishing* is the opposite of analogy: a demonstration that two situations are so fundamentally dissimilar that the same result should not occur in both. Analogizing and distinguishing [...] help find and state the rule for which a precedent stands, together with something about how that rule is to be applied. Distinguishing does so by showing what the rule is not and how it is not to be applied. There are three steps in analogizing or distinguishing. First, make sure that the issue in the precedent is the same one you are trying to resolve. Second, identify the precedent's determinative facts [...] facts that the precedential court treated as crucial and on which it really relied. Finally, compare the precedent's determinative facts to the facts you are trying to resolve" (Neumann 2009, 154).

[11] "The law denies a landowner damages for unauthorized entry against an intruder under necessity because the purpose of saving life or property is more important than the purpose of giving inviolate status to property, and it would be morally improper for the landowner to deny entry. Those reasons apply equally well when the issue is whether the landowner can use self-help to eject the intruder. Accordingly, damages and self-help cannot be distinguished for this purpose. It would therefore be inconsistent to adopt a rule that the landowner can use self-help while adhering to the rule that he cannot recover damages" (Eisenberg 1988, 94).

[12] "This doctrine of necessity applies with special force to the preservation of human life [...] One may sacrifice the personal property of another to save his life or the lives of his fellows."

similar case in language that highlights the relevant similarities. In *Mouse's Case*, 12 Co. 63, the court rejected plaintiff's suit against a ferryman for "taking and carrying away the plaintiff's casket and its contents" in the midst of a "great tempest[…], and a strong wind, so that the barge and all the passengers were in danger of being lost if certain ponderous things were not cast out, and the defendant thereupon cast out the plaintiff's casket" (71 A. 189).

Eisenberg's critique of Levi's model of reasoning by example and, in particular, of its focus on similarity may stem from the same problem that Cass Sunstein and Scott Brewer recognize, the need for a set of criteria by which analogical reasoning can assess relevant similarities and differences. "The method of analogy is based on the question: Is case A relevantly similar to case B, or not?[…] To answer such questions, one needs a theory of relevant similarities and differences. By itself, analogical reasoning supplies no such theory. It is thus dependent on an apparatus that it is unable to produce" (Sunstein 1993, 774; Brewer 1996, 932f).

These authors find the relevance criteria in the analogy-warranting rule or low-level principle informing the analogy. For Brewer, analogy-warranting rules subsume the source and target cases and lend justificatory weight to the analogies for the reasons elaborated in the analogy-warranting rationale (Brewer 1996, 1020). For Sunstein, the low-level principle informs and is informed by the analogy (Sunstein 1993, 778; Ashley 2002, 172–175). Brewer's analogy-warranting rules and rationales and Sunstein's principles produced by analogical reasoning are plausible attempts to provide content to the concept of relevant similarity, more plausible, I believe, than Eisenberg's attempt to circumvent the concept entirely.[13]

As I have argued elsewhere (Ashley 2002, 173), the focus of Brewer's theory of relevance on analogy-warranting rules and of Sunstein's on principles are probably closely related. Sunstein's principles produced by analogical reasoning "operate at a low or intermediate level of abstraction" (Sunstein 1993, 747), and Brewer's account of analogical reasoning draws on a similar process of analogical reasoning in ethics, in which "relatively precise norms or principles" play the role of AWR (Brewer 1996, 979). Brewer uses the term "subsumption" (as in the analogy-warranting rule subsumes the relevant features of the source and target cases). Sunstein focuses on principles operating at a "low or intermediate level of abstraction," that is, closer to the level of facts. As noted, since Eisenberg's applicable social propositions include normative principles, one is tempted to conclude that, properly considered, "relevant similarities" play a de facto role in his accounts of reasoning from precedent and by legal analogy, as well.

[13] Eisenberg asserts that "cases are not determined in the common law simply by comparing similarities and differences," but his conception of comparison seems to be limited to comparing the numbers of similarities and differences. For example, he argues, "Here there are nine similarities between the cases and only one difference, but obviously the difference is decisive, and it would be decisive if ninety more similarities were added" (Eisenberg 1988, 84). That, no doubt, is true, but as various AI and Law models of case-based legal reasoning have demonstrated, one can do more in comparing cases than count similarities and differences. One can compare sets of similarities and differences shared among cases and a problem (Ashley 1990), and that set comparison can be informed by underlying reasons (Aleven 2003) and values (Bench-Capon and Atkinson 2009).

The jurisprudential accounts of "relevant similarities" and "relevant differences" in precedent and legal analogy bring together factual descriptions, legal rules and their component concepts, and underlying normative values and principles, in a complex and dynamic interaction that reflects legal practice:

> The judicial, scholarly and juristic practice of arguing for or against the applicability of precedent regularly takes the form of close analysis of material facts. Judges carefully examine the material facts to ascertain degrees of relevant similarity. Scholars use similarity or lack thereof as support for argument. Advocates stress factual similarity or dissimilarity as a major structural component of the position they take in their presentation to the court[…]. Argumentation over the applicability of precedent also involves isolating rules and principles. In so doing, courts, scholars and advocates will argue that a rule or principle explicit or implicit in the holding does or does not apply to the case at hand. Further, courts sometimes determine the applicability of a precedent by considering whether the next case falls within the substantive reasoning underlying the precedent. (Summers 1997, 387)

The exact nature of this interaction is still not very clear (Marshall 1997, 512–513). Manifestly, it involves a process of selecting facts and choosing in what terms and how abstractly to describe them. This process is driven by the role of the proponent in the argument (i.e., the advocates or sometimes the judge), the factual and legal contexts of the problem scenario, the way an argument has already proceeded, the available precedents and legal rules, the legal concepts employed in those rules and their putative meanings, and the values and policies underlying those rules.[14]

The interaction also varies with time. As Eisenberg's views on the role of applicable social propositions and Schauer's views on characterization and assimilation, discussed above in Sect. 2, imply, assimilation of facts into a legal rule's concept is temporally dynamic. Given that the differences and similarities deemed relevant depend on applicable social propositions and on linguistic rules of relevance, both of which change over time, what counts as relevantly different or similar will also change over time.

As Brewer points out, another temporal consideration about legal rules associated with precedents and legal analogies (i.e., analogy-warranting rules in Brewer's terminology) is their defeasibility. "A defeasible argument is one in which the addition of premises can weaken the force of the conclusion." A judge knows "that later judges may well come along and rewrite the AWR." Therefore, "in a context of doubt, the legal reasoner uses the resources of analogy both to build and to maintain confidence in her judgment about how that doubt is to be resolved" (Brewer 1996, 1017, 1020). "[T]he reasoner keeps her eye on the shared characteristics of source and target and thus does not simply dispense with the example, because she is confident that source and target are alike in the respects specified by the AWR, that those respects are relevant to being 'defeased' or not, that the source case managed to defeat defeasibility, and that therefore one ought to adjudge defeasibility as being likewise defeated in the target case as well" (Brewer 1996, 1020). Thus, as analogies are drawn they invite more detailed comparisons of current facts with precedents to see whether facts that

[14] An AI and Law model that included an algorithm for determining how and how abstractly to characterize facts in an argument comparing cases was developed in CATO (Aleven 2003).

purportedly matter now were not also present then, even if their significance was not, or might not have been, recognized.[15]

This interaction is closely related to the task of determining the *ratio decidendi*, or rule of decision of a case, which raises similar issues of selection, characterization, and abstraction of case facts.

> Juristic discussions, involving both judges and scholars, do thus exhibit a certain confusion, a difficulty in capturing or conceptualizing in clear terms some central elements of practice. The *ratio* is perhaps to be considered an essentially contested concept, because it is not purely descriptive but also evaluative or normative in force. The difficulty, perhaps impossibility, of achieving consensus in definition should not be thought to mirror a like confusion in practice; for on the whole experienced lawyers and judges are able, with a relatively high degree of common understanding, to operate effectively in practice with a system of precedent in its application case-by-case, as distinct from abstract description (Marshall 1997, 512–513).

While jurisprudential scholars may meet this reassurance skeptically, it does seem to make a wise point that the rule of the case is not purely descriptive but also evaluative and often contested in a case. As Marshall notes, common law jurisprudence exhibits an "absence of standardization in the opinions that come under consideration" (Marshall 1997, 513).

There are, however, identifiable schemes for contesting proposed formulations of precedents' decision rules. For example, practitioners may argue that a proposed rule of a precedent is too broad or too narrow by posing a carefully constructed hypothetical, as described below. The aim of a more empirically focused descriptive account such as (MacCormick et al. 1997) is to identify the types of such arguments and illustrate them with examples, as discussed in the next section. The aim of much work in AI and Law is to identify argumentation schemes associated with those types and examples and to implement them computationally, or to enable computers to identify them in decision texts for purposes of automated reasoning and improving legal information retrieval.

Synthesizing the above, reasoning from precedent and legal analogy, distinguishing, and relevant similarity involve characterizing facts in terms of intermediate legal concepts to be used in legal rules, assimilation of facts into those concepts given "sticky" social and linguistic categories, and decisions about applying or distinguishing those rules in light of normative principles that reveal themselves in comparing the problem and cases in light of applicable social propositions. Even if these processes elude precise definition in jurisprudential theories, one may still hope to identify examples of legal practitioners' arguments dealing with these issues and to extract the corresponding argument schemes. "Experienced lawyers [. . .] are perfectly able to handle all these situations [. . .] and [. . .] variants [. . .], and to implement a coherent system of precedent" (Marshall 1997, 513).

[15] For instance, suppose that in one of the innkeeper cases referred to in Sect. 3, an innkeeper had been held strictly liable even though he had posted notices warning customers to protect their valuables. Such a source case had managed to defeat defeasibility in Brewer's sense and could be useful in a future argument that the posting of notices is irrelevant.

6 Arguments in Practice for Following Precedent or Legal Analogy

The above jurisprudential accounts of reasoning from precedent or legal analogy provide examples of courts' reasoning as recorded in legal opinions for which the authors construct somewhat idealized explanations in terms of their theoretical models.

In *Interpreting Precedents: A Comparative Study*, Neil MacCormick and Robert Summers undertook a systematically more descriptive approach to surveying the ways in which courts employed precedents and analogies in legal reasoning. They also cite examples of courts' reasoning as recorded in legal opinions and explain them in terms of models, but the focus of their models is succinctly to describe the fairly diverse modes of reasoning with precedent across eleven civil and common law jurisdictions including the USA in Summers (1997).

Summers first describes the general role of precedent in relation to other sources of law in the following terms.

> [P]recedent in the New York courts is the primary source of decisive authority in common law subject areas such as contract, tort and property. […] [A] relevant statutory text […] prevails over any conflicting precedent. (So, too the constitution prevails over statute and case law.) Also a precedent interpreting a statute becomes a binding interpretation for future cases. When there is no relevant precedent from New York or elsewhere, and no controlling statute, a scholarly treatise may have the most influence on the court. Also, in such cases, the court gives more than usual weight to purely substantive considerations of policy and principle (Summers 1997, 365).

In the course of elaborating upon the uses of precedent, Summer's illustrates kinds of argument that, based on past practice, one can reasonably make to a court in support of following a precedent, departing from a precedent, being guided by a legal analogy (or not), and arguing from hypothetical cases. This section presents a series of tables listing, categorizing, and summarizing these kinds of arguments. It then illustrates some of them in the context of a small set of related case examples. (Section 8 focuses briefly on special considerations for arguments from precedent or by analogy involving statutory interpretation.)

In presenting these lists, my assumption is that a legal forum would regard any argument type Summers includes, at least, as a reasonable type of argument move for an advocate to advance, even if a court may not agree that it is a convincing argument in a particular case. Thus, the lists and summaries below gloss over the finer qualifications on argumentation that Summers reports, such as, that "courts *sometimes* determine" [emphasis added] or that "some courts are less 'substantive rationale-minded'" than others. The lists also omit the more technical assumptions that need to be satisfied in order for a type of argument to be appropriate. These include, for instance, the formal requirements that the court in the target case be lower in the jurisdiction's judicial hierarchy and subject to the authority of the precedent court.[16] Other assumptions capture more substantive conditions that should be true

[16] "Lower courts, for example, are expected to respect the decisions of higher courts. But the hierarchical ordering of decision makers implicates considerations different from those involved when

Table 1 Types of arguments for following precedent

1. Relevant similarities	The source precedent's material facts are relevantly similar to the target case (Summers 1997, 387)
2. Source-rule-applies	"A rule or principle explicit or implicit in the holding of" the source precedent applies to the target case" (Summers 1997, 387)
3. Source-rationale-applies	The target case falls within the substantive rationale in the source precedent (See Summers 1997, 387)

for the argument to have force. Some of these are addressed below in the section concerning argument schemes for arguing from precedent or by analogy.

6.1 Arguments for Following/Departing from Precedent

The three main types of arguments in favor of following a precedent, set out in Table 1, focus on mapping something from the source precedent to the target case: relevant factual similarities, an applicable rule, or an applicable rationale. The argument types can be used in combination ideally where the relevant similarities, source rule, and rationale are all in alignment.

For each of the main types of arguments in favor of following a precedent, there are complementary argument types against doing so.[17] Collectively, Tables 2 through 6 below provide a comprehensive listing of the kinds of critical questions that apply to arguments with precedents (or to arguments by analogy or with hypotheticals) and, depending on the answers, that can lead to counterarguments. The types of counterarguments that focus on distinguishing a precedent's facts are in Table 2, those that target the precedent's rule are grouped in Table 3, and those that target the precedent's rationale are in Table 4. An argument that a source is distinguishable (see Table 2, type 2) can be used to attack the applicability of a precedent's rule or rationale based on a showing that the precedent is factually distinguishable, but it has not been repeated in Tables 3 or 4. Finally, Table 5 summarizes types of arguments for departing from precedent because the source was mistaken, and Table 6 lists arguments involving additional factors, some of which can negatively affect the force of a precedent.

It should also be noted that some argument types in the tables overlap and could probably have been consolidated. The various formulations in the MacCormick and

a decision maker is constrained by its previous actions as opposed to the orders of its superiors in the hierarchy" (Schauer 1987, 576).

[17] As noted, I omit responses that argue that the court lacks authority to distinguish or overrule a precedent.

Table 2 Types of arguments for departing from precedent: regarding facts

1. Not-relevantly similar	The source precedent's material facts are *not* relevantly similar to the target case (Summers 1997, 387, 390)
2. Source-distinguishable	"[A]lthough there is an alleged rule or principle that covers the instant and earlier cases, the facts are sufficiently different (specify some, of many, differences) to conclude that the rule or ratio does not apply to the present case" (Marshall 1997, 516)[a]

[a] As discussed in Sect. 5, the principle and its underlying values provide guidance as to the type of similarities and differences that are legally significant

Table 3 Types of arguments for departing from precedent: regarding rule

1. Source-rule-inapplicable	"A rule or principle explicit or implicit in the holding of" the source precedent does *not* apply to the target case" (Summers 1997, 387, 390)
2. Source-rule-too-broad	"It may be held that a principle apparently laid down as the reason for a particular decision was too widely stated or in some other way inappropriate" (Marshall 1997, 516)
3. Source-rule-not-binding	"[C]onclude that what has been alleged to be a binding ratio falls into the category of *obiter dictum*" (Summers 1997, 516)

Summers text have been preserved, however, since they may highlight variations on the themes of the major types of responding arguments.

Table 5 summarizes types of arguments for departing from precedents because the precedents were ill-conceived from the beginning or proved unworkable from the start.

Summers identifies a number of additional factors that can affect the force of a precedent and that one would expect to observe in arguments urging a court to follow a precedent or not (See Table 6). These factors concern reliance on long-established precedent, whether the precedent has become a leading case, trends in other states or jurisdictions, related changes in other areas of law, academic critiques, and whether there is an alternative process for changing the rule of the precedent.

6.2 *Arguments from Legal Analogies or Hypotheticals*

Two formulations of arguments from legal analogy are presented that map a rule, a principle, or an approach from a factually different although fundamentally sim-

Table 4 Types of arguments for departing from precedent: regarding rationale

1. Changed-social-circumstances	"Changes in the political, economic or social background […] require adjust[ing] the law to conform to the new circumstance" (Summers 1997, 374)[a]
2. Changed-social-values	"when the […] social or moral […] substantive values upon which the precedent was based are no longer tolerable" (Summers 1997, 396)
3. Source-rationale inapplicable	The target case does *not* fall within the substantive rationale in the source precedent (See Summers 1997, pp. 387, 390).[b] "Courts will often examine closely the facts or rationale of a prior decision to ensure that a rule, which seems superficially relevant to resolving a dispute, is not applied to the facts of a case that would not be well resolved under the prior decision" (Summers 1997, 391)
4. Source-obsolete	Technological innovations or improvements have made the precedent obsolete (Summers 1997, 396)

[a] "[A]s the probability of unsound results increases, and in the face of mounting evidence that the rule no longer fits, courts do overturn precedent and create a new rule" (Summers 1997, 375) providing examples of the New York Court of Appeals overturning decades-old decisions due to changes in society, conceptions of justice, and evidentiary practicalities. This seems consistent with Eisenberg's focus on applicable social propositions

[b] "A lawyer confronted by an unfavourable precedent will frequently attempt to distinguish the precedent, arguing either that the material facts were different or that the substantive rationale for the ruling does not apply to the facts of the case under consideration. The principle enunciated by Llewellyn in The Bramble Bush, 'the rule follows where its reason leads; where the reason stops, there stops the rule' (Llewellyn 1951, 15–18), is applied with vigour by many American lawyers and judges." (Summers 1997, 390)

ilar case. MacCormick and Summers refer generally to this type of argument as "illustrative or analogical precedent."

As summarized in Table 8, Summers elaborates a rich description of ways to employ hypotheticals in legal argument to place a target problem in relation to clear or borderline examples, to critique proposed results as leading to absurdities or incoherence, and to reason about proposed legal rules (MacCormick et al. 1997, 528–529).

6.3 Examples

Quite a few of the above argument types can be recognized in three related examples provided by Summers (1997, 375, 388). The examples are consolidated and elabo-

Table 5 Types of arguments for departing from precedent: regarding source mistaken

1. Source-ill-conceived	While it is not enough "merely to demonstrate that, [if the court] had decided the precedent originally, it would have decided differently," (MacCormick and Summers 1997, 525),[a] if an advocate "overtly challenges the arguments which support the precedent, the court may be compelled to evaluate the soundness of the earlier arguments in determining whether to apply or distinguish a precedent." (Summers 1997, 375). "[W]here subsequent experience with a precedent shows that it was substantively quite erroneous or ill-conceived from the beginning," (MacCormick and Summers 1997, 396, 526), or that it was "taken per *incuriam*, that is, in ignorance of applicable precedent or statute" (Marshall 1997, 516)
2. Source-never-settled	If the rule of the precedent led to confusion or became unworkable and thus never became part of settled law (Summers 1997, 376)

[a] Compare with Schauer's statement, "But if we are truly arguing from precedent, then the fact that something was decided before gives it present value despite our current belief that the previous decision was erroneous" (Schauer 1987, 575)

Table 6 Arguments re additional factors affecting precedential force

1. Reliance	Some precedents may be observed because they have been the law for so long, "citizens have relied on them" (Summers 1997, 375). Precedential force depends on the area of law involved. Is the field "assumed to give rise to a high degree of reliance on settled precedent" (e.g., property and contractual rights)? (Summers 1997, 376)
2. Leading case	When a precedent becomes a leading case, it has even more normative force (Summers 1997, 389)
3. Trends	Is a precedent contrary to a trend in other states? (Summers 1997, 376)
4. Related change	"Legal change in related areas of the law may lead courts to examine the relevance of such change to the legal issues before them" (Summers 1997, 376)
5. Academic critique	"The formal bindingness or other degree of normative force of a precedent may be affected by academic writings [which] provide support for adhering to or not adhering to precedent" (Summers 1997, 377)
6. Alternative	Is there an alternative process of amendment? (Summers 1997, 376)

rated here with references to the argument types noted with [Table No., Argument Type, Item No. Item].

In *Woods v. Lancet*, 303 N.Y. 349 (1951), the New York Court of Appeals (New York's highest court) overruled *Drobner v. Peters*, 232 N.Y. 220 (1921) according to which a complaint "alleging prenatal injuries, tortiously inflicted on a nine-month fetus, viable at the time and actually born later" did *not* state a valid cause of action. The *Woods* case involved a similar complaint on behalf of an infant plaintiff, alleging that, "while the infant was in his mother's womb during the ninth month of her pregnancy, he sustained, through the negligence of defendant, such serious injuries that he came into this world permanently maimed and disabled." Even though there was "no material distinction between" the two cases, 303 N.Y. 351, the Court declared that "*Drobner v. Peters* must be examined against a background of history and of the legal thought of its time and of the thirty years that have passed since it was handed down." 303 N.Y. 352. [Table 4, Types of Arguments for Departing from Precedent: Re Rationale, 1. Changed-Social-Circumstances, 2. Changed-Social-Values, 3. Source-Rationale Inapplicable, 4. Source-Obsolete].

The Court noted that, "since 1921, numerous and impressive affirmative precedents have been developed," citing decisions from five other states and Canada holding in favor of the availability of such a claim. [Table 6, Arguments Re Additional Factors Affecting Precedential Force, 3. Trends].

The court opined that, "surely, as an original proposition, we would, today, be hard put to it to find a sound reason for the old rule. Following *Drobner v. Peters (supra)* would call for an affirmance but the chief basis for that holding (lack of precedent) no longer exists." 303 N.Y. 353-4. [Table 4, Types of Arguments for Departing from Precedent: Re Rationale, 3. Source-Rationale Inapplicable].

The Court notes the reliance interest underlying following precedent: "Of course, rules of law on which men rely in their business dealings should not be changed in the middle of the game, but what has that to do with bringing to justice a tort-feasor who surely has no moral or other right to rely on a decision of the New York Court of Appeals? [Table 6, Arguments Re Additional Factors Affecting Precedential Force, 1. Reliance (Not Followed)]. Negligence law is common law, and the common law has been molded and changed and brought up-to-date in many another case." 303 N.Y. 354.

The Court also noted that "some kinds of changes in the common law could not safely be made without the kind of factual investigation which the Legislature and not the courts, is equipped for." "Legislative action there could, of course, be, but we abdicate our own function, in a field peculiarly nonstatutory, when we refuse to reconsider an old and unsatisfactory court-made rule." 303 N.Y. 355. [Table 6, Arguments Re Additional Factors Affecting Precedential Force, 6. Alternative (Not Followed)]. "This child, when injured, was in fact, alive and capable of being delivered and of remaining alive, separate from its mother. We agree with the dissenting Justice below that 'To deny the infant relief in this case is not only a harsh result, but its effect is to do reverence to an outmoded, timeworn fiction not founded on fact and within common knowledge untrue and unjustified.'" 303 N.Y. 357. [Table 4, Types of Arguments for Departing from Precedent: Re Rationale, 1. Changed-Social-Circumstances, 2. Changed-Social-Values, 4. Source-Obsolete].

In a subsequent case, *Albala v. City of New York* 54 N.Y.2d 269, the Court affirmed a determination that "a tort committed against the mother of a child not yet conceived" did not give rise "to a cause of action in favor of the child if that tort caused injury to the child during gestation." It was contended on behalf of the infant plaintiff "that as a result of the alleged malpractice of defendants in negligently perforating [his mother's] uterus, seven years prior to this lawsuit, plaintiff was born with a damaged brain." 54 N.Y.2d 271. The Court distinguished the *Woods* case: "The instant case differs significantly from *Woods v. Lancet* (303 NY 349) where we upheld a cause of action on behalf of a child for prenatal injuries incurred in utero as a result of a tort committed against the child's mother during her pregnancy. In that case at the time the tort is committed there are two identifiable beings within the zone of danger each of whom is owed a duty independent of the other and each of whom may be directly injured." 54 N.Y.2d 272. [Table 2: Types of Arguments for Departing from Precedent: Re Facts, 2. Source-Distinguishable].

The Court also distinguished the case of *Park v. Chessin* 46 N.Y.2d 401 (1978) where the Court had "refused to recognize a cause of action asserted on behalf of [a] child born with a condition on the basis of the allegations that had the negligence not occurred the afflicted child would never have been conceived, or if conceived the pregnancy terminated." The mother's first child had died of the same condition, but an obstetrician negligently assured the woman before conceiving the plaintiff that the condition was not hereditary and that the probability of another child's being born with the condition were "practically nil". 54 N.Y.2d 272. The Court distinguished the Park case because, "assuming the allegations in the complaint to be true, had the alleged negligence not occurred and Ruth Albala's uterus not been perforated, plaintiff would have in all likelihood been born normal. Here, the defendants' alleged negligence made the difference between life in an impaired state and life in an unimpaired state, whereas in *Park* the alleged negligence made the difference between life in an impaired state and nonexistence." 54 N.Y.2d 272-3. [Table 2: Types of Arguments for Departing from Precedent: Re Facts, 2. Source-Distinguishable].

Despite the distinction, the Court found a useful analogy in the *Park* case, where, it said, " we isolated the central concern [...] for judicial recognition of the birth of a defective child as an injury to the child. We noted the staggering implications of any proposition which would honor claims assuming the breach of an identifiable duty for less than a perfect birth and by what standard and the difficulty in establishing a standard or definition of perfection." 54 N.Y.2d 273. [Table 7: Analogical Argument Types, 1. Sufficient Similarity that Rule Should Apply, 2. Source Illustrative if Not on Point].

The Court went on to pose a hypothetical: "Were we to establish liability in this case, could we logically preclude liability in a case where a negligent motorist collides with another vehicle containing a female passenger who sustains a punctured uterus as a result of the accident and subsequently gives birth to a deformed child? Unlimited hypotheses accompanied by staggering implications are manifest. [Table 8: Types of

Table 7 Analogical argument types

1. Sufficient similarity that rule should apply	"[A]lthough the facts of this case allegedly differ from those in earlier cases, they are sufficiently similar (mention some, of many, similarities) to conclude that the rule or ratio should apply to the present case" (Marshall 1997, 516)
2. Source illustrative if not on point	"Precedents which are not strictly in point may be used as illustrative examples or analogies to clarify a court's conclusion." In cases of "first impression," a court may examine a source case involving a similar issue and map either the considerations or a principle to the target case. "As a result such precedents may also generate an overall harmonization argument favouring one outcome rather than another." "Thus, [the analogous source case] although distinguishable and distinguished, clarifie[s] the 'central concern' upon which the decision in [the target problem can be] based" (Summers 1997, 387–388)

Arguments from Hypotheticals 1. Hypo Demonstrating Distinctions/Continuum, 2. Reductio Ad Absurdum Hypo, 3. Hypo Demonstrating (In)coherent Application, 4. Hypo Clarifying Rule/Rationale, 6. Hypo to Reject Applying Rule]. The perimeters of liability although a proper legislative concern, in cases such as these, cannot be judicially established in a reasonable and practical manner." 54 N.Y.2d 273-4.

The Court responded to three cases in other jurisdictions that recognized the validity of a cause of action for preconception tort. It found two "unpersuasive" as "decided on policy grounds based largely on a misplaced reliance upon precedent in prenatal injury cases which, as we noted above, are inapposite" [as was the *Woods* case]. 54 N.Y.2d 274. The third, it distinguished, as having been decided "[u]nder a products liability theory, [where] the liability of the manufacturer is extended to the entire class of persons thereby affected regardless of privity, foreseeability or due care […]. Accordingly, the necessity of establishing manageable bounds for liability is conspicuously absent." 54 N.Y.2d 274. [Table 4, Types of Arguments for Departing from Precedent: Re Rationale, 3. Source-Rationale Inapplicable].

6.4 Comparing Jurisprudential and Descriptive Accounts

It is worth pausing briefly to compare the types of arguments Summers reports from actual cases with the jurisprudential models discussed in the beginning of the chapter.

In the main, the argument types in practice are consistent with the jurisprudential models. The Summers account confirms Eisenberg's and Brewer's focus on the

Table 8 Types of arguments from hypotheticals

1. Hypo demonstrating distinctions/continuum	Construct hypothetical cases to articulate "distinctions between paradigm cases and borderline cases and the creation of conceptual bridges between cases along a continuum" and relate to case to be decided
2. Reductio ad absurdum hypo	Construct analogous hypothetical case for use in "reductio ad absurdum argument [] demonstrating the unsoundness of proposed applications of code sections, statutes or doctrinal formulations"
3. Hypo demonstrating (in)coherent application	Construct hypothetical cases to elaborate "coherent patterns of applications of authoritative language and [to] demonstrat[e] how proposed or possible applications [e.g., to target case] would not be coherent"
4. Hypo clarifying rule/rationale	Construct a "clear case[s] to which a code section, statue or doctrine must apply if it is to have any rational application"; construct a "paradigm case [] so as to display a policy rationale in its clearest application" [e.g., a rule/rationale relevant to the case to be decided.]
5. Hypo to extend rule	Construct "a hypothetical case to which it would be justified to extend the application of the rule in a precedent.[…] then argue [] that the facts of the hypothetical cannot be rationally distinguished from another case the court […] is about to decide[…] and conclude [] that it is therefore justified to so extend the rule.[…]"
6. Hypo to reject applying rule	"The hypothetical case functions to help justify rejecting the application of a rule in a precedent to the case just decided (or about to be decided) because the rule can be shown to be unsuitable when applied to the hypothetical, yet the hypothetical cannot be rationally distinguished from the case involved"

central importance of the rule of the precedent, but also on the precedent's rationale and on Eisenberg's focus on the effect on the rationale of changed circumstances since the precedent was decided. The Summers account confirms the central role of distinguishing a precedent as a response to an argument that a court should follow a precedent, its rule or its rationale. As noted above, although Eisenberg disparages the role of relevant similarities in legal analogy, the Summers account confirms its practical importance.

The Summers account also identifies some considerations that, as a practical matter, can strengthen or weaken an argument from precedent (Table 6): the extent of reliance on a precedent, its status as a leading case, trends in subsequent cases, other jurisdictions, or related areas of law, academic critiques, or the existence of alternatives to a court's overruling the precedent.

One may include these with the considerations identified in the jurisprudential accounts that add to or subtract from the strength of an argument from precedent or by analogy. Brewer's theory can "provide the legal reasoner with heuristic guidance in assessing the strengths and weaknesses of analogical and disanalogical reasoning. [It focuses] attention on the fact that an argument by analogy is more likely to succeed when there are compelling reasons to believe that the presence in an item of some characteristics supports the inference that some other characteristics are also present." These components of an argument's "rational force" come from the "AWR and the AWRa that either has been explicitly supplied or could plausibly be supplied" (Brewer 1996, 930).

On Eisenberg's account, an argument to depart from a precedent is strengthened to the extent that following precedent would counter currently prevailing social propositions. Schauer agrees subject to his view of precedent as raising presumptive argument burdens rather than absolute prohibitions or constraints. His

> account asserts that a doctrinal proposition is presumptively controlling even when it is inconsistent with some, or all, social propositions. Only when the weight of the inconsistency is overwhelming is the presumption overcome[…]. At times this strength may be embodied in a particularly compelling factual situation, such as that in Riggs v. Palmer. [FN39] At times it may be embodied in a particularly compelling policy argument, as in MacPherson v. Buick Motor Co. [FN40] And at times it may be manifested in a particularly outmoded doctrine or in a particularly compelling moral argument, both of which were present in Brown v. Board of Education. [FN41]. (Schauer 1989, 470)

7 Argument Schema for Legal Analogy

This section describes more detailed argument schemes for some of the kinds of arguments addressed in the jurisprudential and descriptive models presented above. Specifically, these schemes are for drawing case-based analogies and for using hypothetical cases to argue that a proposed rule for deciding a case is too broad or too narrow. Matthias Grabmair, a former Ph.D. candidate at the University of Pittsburgh Graduate Program in Intelligent Systems, and I formalized these schemes as part of his dissertation project. He transformed these formalized schemes into a "computational model" of legal argument, a computer program which, given inputted scenarios, outputs appropriate arguments based on the formal models.

A particular focus of this research is to design a computational model of legal argument that incorporates the values underlying legal rules in a legally reasonable way. Like Eisenberg's model, it treats the decision to apply a rule from a case as a value judgment. In our model, a legal rule not only states a set of conditions from which a legal consequence shall follow, but also is associated with certain underlying values. Likewise, a case decision represents not only a judge's determination whether the rule's conditions are fulfilled, but also the judge's value judgment whether the positive effects on underlying values of the rule's conclusion outweigh the negative effects on those values (Grabmair and Ashley 2011).

This research extends jurisprudential efforts like those of Eisenberg and Brewer to make explicit the patterns of legal argumentation by analogy and by posing hypotheticals and to account for more of the patterns identified in descriptive models like that of Summers. While some legal theorists, notably Robert Alexy, have focused on providing formulas for such value judgments that take into account the abstract weights of values and their degree of interference with each other in a specific case (Alexy 2003), Grabmair focuses on a computational process for making the *arguments* how the conflicting values should be resolved (Grabmair and Ashley 2011).

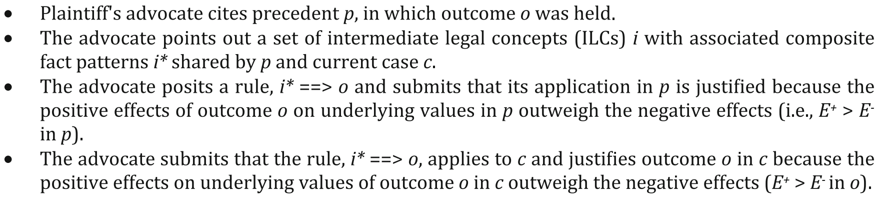
A computer program implementing our model could predict that a consequence of applying the legal rule in a fact situation would promote some values at the expense of others and explain its reasoning explicitly in an argument. Although it may not be apparent from the textual summaries presented below, the actual formalizations are detailed enough to construct algorithms enabling a computer program to make and respond to arguments by analogy and with hypothetical cases.

Apart from computer implementations and more relevant for this Handbook, making these patterns of argument explicit in textual form could be useful for instructing law students about them even without an intelligent tutoring system. Not only can students identify particular argument patterns, but they can use them as templates for reconstructing arguments they encounter in legal opinions, or for creating new arguments. They can also critique the adequacy of the formalization and critically assess the degree of persuasiveness to accord them in particular contexts.

7.1 Argument Schemes for Case Analogies and Hypotheticals

As suggested above, the argument schemes Grabmair has defined all assume a particular model of reasoning about actions in terms of their expected effects on values. Applying a legal norm to a fact situation promotes certain applicable values at the expense of other values. In the formalism, E^+ represents the net positive effects on applicable values of a decision. E^- represents the decision's net negative effects on applicable values. In the formalism, a judge's value judgment that the positive effects on underlying values of a proposed rule's conclusion outweigh the negative effects is expressed as $E^+ > E^-$ (Grabmair and Ashley 2011).

As we have seen, an advocate may attempt to convince a judge how to decide a case at bar by analogizing the case at bar to a favorable precedent, pointing out shared facts giving rise to reasons to decide the problem in the same way as the precedent. Assume that the case at bar c involves a dispute about outcome o in situation s. We assume that the argument proceeds in a 3-ply format (Ashley 1990): A proponent cites an analogous precedent, the opponent responds by distinguishing and/or citing counterexamples, and the proponent offers a surrebuttal downplaying distinctions or



- Plaintiff's advocate cites precedent *p*, in which outcome *o* was held.
- The advocate points out a set of intermediate legal concepts (ILCs) *i* with associated composite fact patterns i* shared by *p* and current case *c*.
- The advocate posits a rule, i* ==> *o* and submits that its application in *p* is justified because the positive effects of outcome *o* on underlying values in *p* outweigh the negative effects (i.e., $E^+ > E^-$ in *p*).
- The advocate submits that the rule, i* ==> *o*, applies to *c* and justifies outcome *o* in *c* because the positive effects on underlying values of outcome *o* in *c* outweigh the negative effects ($E^+ > E^-$ in *o*).

Fig. 2 Argument schemes for case-based analogies: argument from sufficient similarity

distinguishing the counterexamples. Section 7.2 provides examples illustrating the application of some of the following argument schemes in a real case.

In the opening move, the plaintiff's advocate cites precedent *p* whose outcome was *o*. A scheme for an argument by analogy is shown in Fig. 2 (Grabmair and Ashley 2011). According to the scheme, the advocate posits a rule whose conditions describe in terms of intermediate legal concepts the basis of the analogy that justifies deciding the problem case like the precedent given the expectation of a similar net positive effect on underlying values.

The opponent may respond by citing its own precedents and distinguishing the proponent's, pointing out unshared circumstances that give rise to reasons for deciding the problem and precedents differently. Figure 3 shows argument schemes for these various rebuttals (Grabmair and Ashley 2011).

The responsive arguments express disagreements with the proposed rule or the asserted analogy, its cause or effects. The distinctions involve various unshared features or undesirable consequences that interfere with the analogical inference based on the precedent that the positive effects on underlying values of outcome *o* in *c* outweigh the negative effects or lead to an alternative rule that does not apply in *c*. In addition, the response may cite a counterexample to support an analogical inference that the negative effects on values of outcome *o* in *c* outweigh the positive effects.

The proponent may then downplay the opponent's distinctions or point out compensating features following the argument schemes for surrebuttals in Fig. 4, distinguish the opponent's counterexamples, or present an alternative argument (Grabmair and Ashley 2011).

As discussed above, in Brewer's model of analogical reasoning, illustrated in Fig. 1, analogy-warranting rationales or principles inform the relevance of similarities and differences. The argument schemes in Figs. 2, 3, and 4 suggest one way to make that model computational, by taking into account effects on underlying values. These schemes (including those for hypothetical reasoning below) model many of the types of arguments by analogy identified in Summers (1997), including those for following precedent (Table 1), departing from precedent based on facts (Table 2), on the rule (Table 3, source-rule-inapplicable and source-rule-too-broad), on the rationale (Table 4, source-rationale-inapplicable), and analogical argument (Table 7, sufficient similarity that rule should apply).

In order to extend the computational model to include hypothetical reasoning, Grabmair has also formalized a set of argument schemes for critiquing a proposed

Distinction Due to Missing Feature in Current Case

- Defendant's advocate argues that an ILC m in p justifies the inference that the positive effects on underlying values of outcome o in p outweigh the negative effects.
- The advocate argues that p and c are not analogous because m is not present in c.

Argument from Undesirable Consequence

- The defendant's advocate posits an alternative rule, $(i^* \cup m)$ ==> o that applies in p.
- The advocate argues that an undesirable consequence q arises if m were omitted from the rule (or a desirable consequence if m were included).
- The advocate may support the distinction if the opinion in p explicitly states that m was relevant for the decision.

Distinction from Non-occurring Undesirable Consequence

- The defendant's advocate hypothesizes that the purpose of the rule in the precedent p was to prevent an undesirable consequence u from occurring.
- The advocate argues that u is not entailed in the case at bar c, and hence the precedent's rule should not apply because the purpose it serves does not apply.

Distinction from Missing Feature in Precedent Case

- The defendant's advocate point out that an ILC m is given in the current case c but not present in the precedent p.
- The advocate may argue that the presence of m in c conflicts with i^*.
- The advocate argues that since m is in c, the negative effects on values outweigh the positive effects.

Counterexample

- The defendant's advocate may state her own precedent $p0$, a case that (ideally) shares i^* with current case c but whose outcome is not o.
- The advocate argues that precedent $p0$ justifies the inference in c that the negative effects on values of outcome o outweigh the positive effects.

Fig. 3 Argument schemes for case-based analogies: rebuttal arguments

Downplaying Significance of Distinction

- The plaintiff's advocate points out an undesirable consequence if the distinctive ILC m were to be required for the analogy to p (or a desirable consequence if m were not required.)

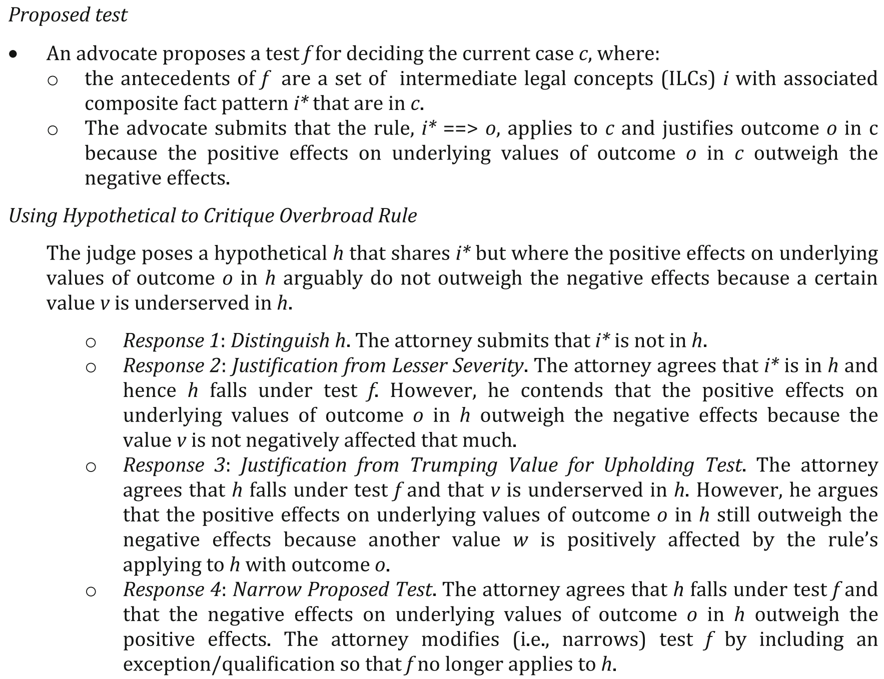
Feature Substitution

- Plaintiff's advocate can point out an ILC n, which is present in current case c but absent in p, which compensates for the absence of m in c.
- The advocate submits that given n, the fact pattern associated with i^* justifies the inference that the positive effects on values outweigh the negative effects.

Fig. 4 Argument schemes for case-based analogies: surrebuttal arguments

test as too broad or too narrow. Like the argument schemes for analogical argument, these argument schemes are based on the assumption that applying a legal rule to a fact situation promotes certain applicable values at the expense of other values and that this tradeoff is the basis for critiquing the rule.

A creative feature of hypothetical reasoning is imagining a scenario that satisfies the conditions of a proposed test but where the positive effects on underlying values



Proposed test

- An advocate proposes a test *f* for deciding the current case *c*, where:
  - the antecedents of *f* are a set of intermediate legal concepts (ILCs) *i* with associated composite fact pattern *i** that are in *c*.
  - The advocate submits that the rule, *i** ==> *o*, applies to *c* and justifies outcome *o* in c because the positive effects on underlying values of outcome *o* in *c* outweigh the negative effects.

Using Hypothetical to Critique Overbroad Rule

The judge poses a hypothetical *h* that shares *i** but where the positive effects on underlying values of outcome *o* in *h* arguably do not outweigh the negative effects because a certain value *v* is underserved in *h*.

- ○ *Response 1*: *Distinguish h*. The attorney submits that *i** is not in *h*.
- ○ *Response 2*: *Justification from Lesser Severity*. The attorney agrees that *i** is in *h* and hence *h* falls under test *f*. However, he contends that the positive effects on underlying values of outcome *o* in *h* outweigh the negative effects because the value *v* is not negatively affected that much.
- ○ *Response 3*: *Justification from Trumping Value for Upholding Test*. The attorney agrees that *h* falls under test *f* and that *v* is underserved in *h*. However, he argues that the positive effects on underlying values of outcome *o* in *h* still outweigh the negative effects because another value *w* is positively affected by the rule's applying to *h* with outcome *o*.
- ○ *Response 4*: *Narrow Proposed Test*. The attorney agrees that *h* falls under test *f* and that the negative effects on underlying values of outcome *o* in *h* outweigh the positive effects. The attorney modifies (i.e., narrows) test *f* by including an exception/qualification so that *f* no longer applies to *h*.

Fig. 5 Argument schemes for reasoning with hypothetical cases: rule too broad

of following the rule arguably do *not* outweigh the negative effects because, for instance, a certain value is not adequately served in the hypothetical case. When faced with such a critique, an advocate may distinguish the hypothetical and argue that the test does not apply, concede that the test applies but contest the assertion that the positive effects of following the test do not outweigh the negative effects or assert some compensating trumping value that flows from applying the test, or concede that the test does apply but should not apply to the hypothetical and narrow the proposed test so that it no longer applies. These moves are elaborated in Fig. 5.

Alternatively, the hypothetical may be constructed in such a way that the test does not apply but should because the positive effects on underlying values of following the test would outweigh the negative effects. Here, an advocate may respond: by arguing by analogy that the test does apply to the hypothetical, by conceding that the test does not apply and arguing that the test *should* not apply because the negative effects on underlying values of applying it outweigh the positive effects due to a desirable consequence of the non-coverage, or by conceding that the test does not apply and that it should apply, and broadening the test so that it covers the hypothetical. These moves are elaborated in Fig. 6.

The argument schemes in Figs. 5 and 6 deal with Summers' types of arguments from hypotheticals listed in Table 8, items 5 and 6, above. They expand on a process model of hypothetical reasoning in (Ashley 2009) and exemplify one way to model

Using Hypothetical to Critique Too Narrow Rule

The judge poses a hypothetical *h* that does not share *i** but where test *f* intuitively should cover *h* because the positive effects on underlying values of outcome *o* in *h* outweigh the negative effects.

- *Response 1*: *Analogize h*. The attorney submits that *i** is in *h* and hence that test *f* applies to *h*.
- *Response 2*: *Upholding Narrow Test Using Undesirable Consequence*. The attorney agrees that *i** is not in *h* and hence test *f* does not cover *h*. The attorney submits, however, that the negative effects on underlying values of outcome *o* in *h* outweigh the positive effects due to an undesirable consequence.
- *Response 3*: *Pointing Out Desirable Consequence*. The attorney agrees that test *f* does not cover *h*. The attorney upholds the test by pointing out a desirable consequence of the non-coverage, submitting that the positive effects on underlying values of not holding outcome *o* in *h* outweigh the negative effects.
- *Response 4*: *Broaden Proposed Test.* The attorney agrees that *f* applies to *h* and that coverage is desirable because the positive effects on underlying values of outcome *o* in *h* outweigh the negative effects. The attorney hence modifies (i.e., broadens) the test to cover *h*.

Fig. 6 Argument schemes for reasoning with hypothetical cases: rule too narrow

the kind of reasoning that Eisenberg deems important about the consequences for underlying values of deciding the hypothetical according to the proposed rule. See Sect. 5.

7.2 Examples

As an example, these argument schemes can be used to reconstruct the arguments in a real case, *Cady v. Dombrowski* 413 U.S. 433 (1973) (Grabmair and Ashley 2011). In this example, we present a series of arguments that it would have been reasonable for the parties to have made, provide citations to similar actual arguments in the briefs or oral arguments associated with this and related cases, and explain the arguments in terms of the above argument schemes.

The facts of the *Cady* case are as follows. Mr. Dombrowski was on his way to his brother's farm in Wisconsin when he crashed his rental car into a bridge abutment. When the police picked him up, Dombrowski appeared to be intoxicated and identified himself as a Chicago police officer. The damaged car was taken into police custody and deposited in a private garage a few miles away from the police station. The police searched the car to remove Mr. Dombrowski's service revolver and came across evidence of a murder.

The issue that brought the case to the US Supreme Court was whether the evidence had been rightfully obtained or whether the police should have obtained a warrant before the search. The conflicting values or principles underlying the issue included Mr. Dombrowski's right of privacy and freedom from government searches under the Fourth Amendment of the US Constitution versus the State's need for effective law enforcement and maintenance of public safety.

In seeking to establish that a warrant was not required, the State's advocate could analogize the situation to a Supreme Court precedent, the *Carroll* case: "Your honor, this court must hold that the search by the police did not require a warrant because it was not 'unreasonable' since it falls under the recognized automobile exception of the Fourth Amendment. As this court held in *Carroll v. United States*, a vehicle (*vh*) can be searched without obtaining a warrant (*wa*) beforehand if obtaining a warrant would create a risk (*rsk*) because of the delay in time."[18]

The State's advocate could justify the analogy thusly: "In this case we find that there is a risk because the police had knowledge of a firearm (*fa*) in the vehicle (*vh*). Since the car (*c*) was parked in a lot far from the police station, a delay in searching would have created the risk of the car being opened and the firearm being stolen (*st*). Hence, the *Carroll* holding applies and a warrant is not required."[19]

In terms of the argument scheme for case-based analogy in Fig. 2, the State's advocate proposes a similarity standard justifying the analogy to *Carroll* in terms of intermediate legal concepts, namely the inference that the combination of searching a vehicle and a risk of loss of evidence due to a delay in time justifies the police in not obtaining a warrant ($vh \cup rsk \Rightarrow \neg wa$), all of whose elements are present in the precedent. That is, arguably, the *Cady* case shares the relevant similarities with the precedent case: Car c is subsumed by vh (there is a vehicle). A firearm that could be stolen ($fa \cup st$) is subsumed by rsk. Hence, there is a risk of the possibility of the firearm being stolen from the car (i.e., c is subsumed by rsk). Implicitly, that advocate proposes that positive effects on relevant values of not requiring a warrant outweigh the negative effects (i.e., in terms of the formalism: $E^{+}(\neg wa;\, s \cup vh \cup rsk)$ $> E^{-}(\neg wa;\, s \cup vh\ rsk)$).

In response, Mr. Dombrowski's advocate could argue: "As distinguished from this case, *Carroll* involved the police acting with probable cause (*pc*) as they had a reasonable suspicion of an offense having been committed and the car containing evidence of this offense. Further, said car was functioning and intact (mobility *mb*). This legitimized the warrantless search as the evidence could instantly be driven beyond reach (*ebr*). Requiring a search warrant in that case would have seriously impaired effective law enforcement."[20] The advocate points out the desirability of

[18] See Brief for Petitioner, Cady v. Dombrowski, 1973 WL 171687 (U.S.), Appellate Brief, No. 72-586. p. 21. In Carroll et al. v. United States, 267 U.S. 132 (1925) pp. 153–154, 158–162, the majority holds that probable cause is needed. We use the risk analogy for purposes of the example and address probable cause below. Although the State's advocate did not expressly rely on *Carroll* in his brief, he noted with disapproval that the Court below found that the *Carroll* "pigeonhole" did not apply, and argued that a search could nevertheless be reasonable when carried out in pursuance of the police officer's responsibilities for protecting public health and safety.

[19] Brief for Petitioner, Cady v. Dombrowski, 1973 WL 171687 (U.S.) (Appellate Brief) No. 72–586. pp. 27-28; Cady v. Dombrowski, 413 US 433, 442–443, 447 (1973).

[20] Brief for Respondent, Cady v. Dombrowski, 1973 WL 171688 (U.S.), Appellate Brief, No. 72–586. p. 24. See Carroll et al. v. United States, 267 U.S. 132 (1925) pp. 153–154, 158–162 (suspicion of contraband smuggling).

his or her version of the *Carroll* ruling by explaining that its main purpose was to prevent loss of evidence: "The actual test in *Carroll* requires probable cause and evidence of functioning vehicles. Here, however, the vehicle was rendered immobile by an accident. It cannot be driven away instantly. Also, at no time before the search did the police have knowledge or suspicion of an offense having been committed."[21]

In terms of Fig. 3's argument schemes for rebutting case-based analogies, Mr. Dombrowski's advocate contests the analogy to *Carroll* using a distinction due to a missing feature, disagreeing with the underlying value judgment. The advocate points out the ILCs *probable cause* and *mobile vehicle*, which were present in the precedent (since the car was intact and the police acted on a suspicion of contraband smuggling) but absent from the current case (where the vehicle was immobile and the police did not yet know of the crime.). The advocate argues that the rule in *Carroll* was in fact $vh \cup mb \cup rsk \cup pc \Rightarrow \neg wa$ with a corresponding value judgment that the positive effects on relevant values of applying such a rule outweighed the negative effects.

Dombrowski's advocate (or a Justice) might also argue that the State's proposed rule is too broad, employing hypothetical reasoning to push on the meaning of an intermediate legal concept: The rule from *Carroll* should not be expanded beyond scenarios of functioning cars and probable cause, as the State's advocate suggests, to include any abstract "risk." Imagine a risk (*rsk*) emanating from a motor home (*mh*) at a campsite with a broken engine so that it is not mobile ($\neg mb$). This is very similar to a home, where there is no risk resulting from mobility. Not requiring a warrant to search a dwelling compartment (*sdc*) would frustrate the owner's constitutionally protected expectation of privacy.[22]

In other words, as per the argument scheme for reasoning with hypothetical cases that the rule is too broad, Fig. 5, the advocate or Justice challenges the proposed rule ($vh \cup rsk \Rightarrow \neg wa$) by posing a hypothetical ($h = rsk \cup mh \cup \neg mb$) and points out the undesirable consequence *sdc* arising when the proposed rule is applied. Applying the proposed rule negatively affects relevant values, namely privacy.

The state's advocate could respond to the hypothetical based on the options in Fig. 5's argument schemes for reasoning with hypothetical cases. For example, the advocate could argue that rule would not apply by distinguishing the hypothetical (Fig. 5, Response 1), urging that a disabled motor home on a campsite is not really a vehicle rule. Or he or she could respond that a person living in a motor home has less of an expectation of privacy than a person living in a house, thus upholding the value judgment that the promotion of public safety outweighs the demotion of privacy in an argument from lesser severity (Fig. 5, Response 2).

[21] Brief for Resp., Cady v. Dombrowski, 1973WL 171688 (U.S.), Appellate Brief, No. 72–586. p. 24.

[22] Hypothetical inspired by California v. Charles B. Carney, 1984 US TRANS LEXIS 209, No. 83-859 (US Sup. Ct.), Argument of Mr. Homann for Respondent p. 27.

7.3 Toward Computer Implementation

For his dissertation project, Grabmair implemented a computer program that makes arguments like those in some of the above examples (Grabmair 2016). He is developing a methodology for representing the fact situations of cases and problems as progressions of events with which the program can predict effects of legal decisions on underlying values. Developing computer programs that can interpret rules in light of underlying values and effects of consequences on values is currently a focus of research in AI and Law.[23]

As a way of empirically evaluating the formalism, the arguments generated by the computer implementing the formalism could be compared in a blind test with arguments prepared by law students addressing the same problems. Once perfected, the computational model could drive an intelligent tutoring system to give law students practice in making and responding to such arguments.

The computerized argument schemes and knowledge representation techniques may someday serve as the basis of programs that can learn to extract information about arguments from legal opinion texts and other documents. IBM's *Jeopardy!*-winning Watson program extracts questions and answers from texts. Equipped with argument schemes and information, such a program might even learn to explain its answers. Conceivably, the computational model could also improve legal information retrieval by enabling users to specify, and database systems to find, case opinions that employ examples of such argument schemes. A legal information system might learn to select case texts on the basis of relevance criteria that take into account the kind of argument scheme the user is trying to instantiate (See Ashley 2017).

8 Special Argument Schemes for Statutes or Constitutions

A final question is whether, and the extent to which, the above argument schemes for reasoning with precedent, analogical reasoning, and hypothetical reasoning apply to statutory interpretation. The jurisprudential authors discussed above agree that they do to a substantial extent.

Eisenberg briefly addresses a similar question concerning his model of common law reasoning:

[23] Grabmair's work contributes to prior work in AI and Law on theory construction (Bench-Capon and Sartor 2003) and value-based frameworks for computationally modeling legal argumentation (Bench-Capon 2003; Greenwood et al. 2003; Atkinson et al. 2011). The prior work also uses contextual value promotion/demotion by actions, introduces a decision maker's preferences between actions, and provides corresponding argument schemes. As far as we can tell, however, the prior work relies on an abstract, context-independent hierarchy of values to resolve conflicting arguments. Our approach to comparing cases in terms of values takes into account the specific facts of the situations and the degree of promotion/demotion in those situations. Our approach to contextual balancing expressed by value judgments avoids the need to associate thresholds with values employed in Bench-Capon and Prakken (2010).

> Statutory and constitutional law[…] differ from the common law in that they are rooted in canonical texts, which the courts cannot properly reformulate […]. It is […] not uncommon[, however,] for a court to prefer a given reading of a statutory or constitutional text because that reading is given by a precedent, or because it makes the text more congruent with some relevant moral norm or policy, or more consistent with the body of the law, than would its alternatives. (Eisenberg 1988, 196–197, n. 35).[24]

Schauer notes that "the degree of judicial lawmaking surrounding nominally statutory areas of law has increased more than commensurately, and as a result, common law method flourishes as it has never before" (Schauer 1989, 456).

Brewer confirms that, "Although this type of exemplary, analogical reasoning is quite familiar in common law, it is obviously also thoroughly familiar—in Anglo-American legal systems, anyway in statutory and constitutional cases" (Brewer 1996, p. 936). "[E]ven in legal decisions whose principal source of authority is statutory, regulatory, or (in America) constitutional, the method of decision is the same exemplary process that courts use in decisions whose principal source of authority is 'common law' in the first sense" (Brewer 1996, 936, n. 30).

The authors of the descriptive accounts agree. MacCormick and Summers identify eleven argument types in four broad categories employed in statutory interpretation, including as a systemic argument: "Argument from precedent: interpret statute in conformity with interpretations given it by other courts" (MacCormick and Summers 1991, 512–514).

Summers formulates some considerations for the more specialized arguments from precedents and legal analogies in statutory interpretation. Reliance interests, the availability of a legislative alternative to critiquing a decision or changing the law, and analogies to other statutes play more of a role, and obsolescence, even of a statutory provision, can justify a departure.

While Grabmair's more detailed argument schemes do not deal specifically with statutory interpretive arguments by analogy or with hypotheticals, they could be specialized to do so. A number of the types of arguments from hypotheticals in Table 8 in Sect. 6.2 deal with using hypotheticals to demonstrate the unsoundness of proposed applications of statutes or demonstrate clear cases of a statute's rationale. Table 9 elaborates considerations that could support such arguments as well as those that could serve as critical questions, the answers to which could lead to effective counterarguments. As one addresses specific contexts and cases of statutory interpretation, one is also likely to need to ask more detailed, domain-specific questions about the meaning of terms in, and the construction of, statutory texts (MacCormick and Summers 1991).

[24] Eisenberg continues, "Although courts must faithfully employ constitutional and statutory texts in cases to which they are applicable, it might be said that constitutional and statutory law is not comprised solely of those texts, but consists of the rules that would be generated at the present moment by the application to those texts of the governing principles of interpretation (including the standard of doctrinal stability, the standard of systemic consistency, and a standard of congruence with relevant social propositions)" (Eisenberg 1988, 196–197, n. 35).

Table 9 Argument types for statutory interpretation from precedents and legal analogies

Reliance	Precedents involving statutory interpretation, especially if part of a long line of precedent, have a high degree of force. “Stability and reliability are thought to be crucial in cases construing statutes” (Summers 1997, 377)
Legislative alternative	“It is generally assumed that the legislature is in a better position to rectify a misconstrued statute than is a court.” “For example, if a statute is misinterpreted by the courts, the legislature may easily step in and pass a new law, and do so prospectively (rather than retroactively, as commonly with courts)” (Summers 1997, 377) “American courts often adhere to a precedent construing a statue even while strongly criticizing the result, observing that, if a reading of the statue leads to an unjust result, it is the business of the legislature to modify the rule, a power of amendment it unquestionably has. In regard to constitutions, however, legislatures have no power of amendment” (Summers 1997, 372)
Disapproval	“Thus, that a precedent is a plainly incorrect interpretation of a code or statute in the first place, or that a precedent has been subsequently reversed by statute, or that there is other clear evidence that the legislature disapproves of the precedent, or that the prospective departing court has itself already in prior cases been veering away from the precedent are all considered justifications for departure” (MacCormick and Summers 1997, 525)
Analogy from other statutes	“One form of this argument is that a word should be interpreted in a given way because this will treat similar cases similarly under related statutory provisions” (MacCormick and Summers 1991, 414)
Obsolescence	“affirmatively justifying grounds for departing from precedent[:][…] the precedent is somehow obsolete, given changes in social or other conditions[…] a precedent in the form of an interpretation of a suitably open-ended statute may become obsolete too, and might justify a departure from the precedent even though the language of the statute is not itself amended.” “new moral or social enlightenment since the date the precedent was decided may similarly justify a departure from a precedent, even one under a general statutory or code provision” (MacCormick and Summers 1997, 526)

9 Conclusions

This chapter has provided guidance in reasoning from precedent and by legal analogy as practiced in a common law context. The guidance has been based on jurisprudential theories and descriptive accounts of precedent and legal analogy, philosophical accounts of practical reasoning in terms of argument schemes and critical questions, and computational models of analogical legal reasoning. The requirement to "follow precedent" is not an absolute constraint on a court's discretion, and courts regularly entertain a variety of types of arguments that they should or should not follow a precedent's rule. Some of these argument types invite courts to take into account the effects of applying a rule on relevant values in the current and hypothetical circumstances. Formal and descriptive accounts of these kinds of legal arguments have been considered, including argument schemes detailed enough to enable computers to construct and respond to such arguments. Reasoning with precedents, analogous, and hypothetical cases is employed in common law statutory and constitutional cases, as well, suggesting a future task of adapting the argument schemes to accommodate the additional constraints imposed by authoritative legal texts.

References

Aleven, V. 2003. Using background knowledge in case-based legal reasoning. *Artificial Intelligence* 150: 183–238.

Alexy, R. 2003. On balancing and subsumption. A structural comparison. *Ratio Juris* 16: 433–449.

Ashley, K. 1990. *Modeling legal argument: Reasoning with cases and hypotheticals*. Cambridge, MA: The MIT Press.

Ashley, K. 2002. An AI model of case-based legal argument from a jurisprudential viewpoint. *Artificial Intelligence and Law* 10: 163–218.

Ashley, K. 2009. Teaching a process model of legal argument with hypotheticals. *Artificial Intelligence and Law* 17: 321–370.

Ashley, K. 2017. *Artificial intelligence and legal analytics: New tools for law practice in the digital age*. Cambridge University Press.

Atkinson, K., T. Bench-Capon, D. Cartwright, and A. Wyner. 2011. Semantic models for policy deliberation. In *ACM proceedings of the thirteenth international conference on artificial intelligence and law (ICAIL 2011)*, 81–90.

Bench-Capon, T. 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation* 13: 429–448.

Bench-Capon, T., and K. Atkinson. 2009. Action-state semantics for practical reasoning. In *Proceedings of the 2009 AAAI fall symposium on the uses of computational argumentation*. AAAI Technical Report SS-09-06, 8–13. Palo Alto, CA: AAAI Press.

Bench-Capon, T., and H. Prakken. 2010. Using argument schemes for hypothetical reasoning in law. *Artificial Intelligence and Law* 18: 153–174.

Bench-Capon, T., and G. Sartor. 2003. A model of legal reasoning with cases incorporating theories and values. *Artificial Intelligence* 150: 97–143.

Brewer, S. 1996. Exemplary reasoning: Semantics, pragmatics, and the rational force of legal argument by analogy. *Harvard Law Review* 109: 923–1028.

Eisenberg, M.A. 1988. *The nature of the common law*. Cambridge, MA: Harvard University Press.

Grabmair, M. 2016. *Modeling purposive legal argumentation and case outcome prediction using argument schemes in the value judgment formalism*. Ph.D. thesis, University of Pittsburgh, Pittsburgh, PA.

Grabmair, M., and K. Ashley. 2011. Facilitating case comparison using value judgments and intermediate legal concepts. In *Proceedings thirteenth international conference on artificial intelligence and law (ICAIL 2011)*, 161–170. New York: ACM Press.

Greenwood, K., T. Bench-Capon, and P. McBurney. 2003. Towards a computational account of persuasion in law. In *ACM proceedings of the 9th international conference on artificial intelligence and law (ICAIL 2003)*, 22–31.

Levi, E. 1949. *An introduction to legal reasoning*. Chicago, IL: University of Chicago Press.

Lindahl, L. 2004. Deduction and justification in the law. The role of legal terms and concepts. *Ratio Juris* 17: 182–202.

Llewellyn, K. 1951. *The bramble bush*. Oxford: Oxford University Press.

Holmes Jr., O.W. 1897. The path of the law. *Harvard Law Review* 10: 457.

MacCormick, D.N., and R. Summers (eds.). 1991. *Interpreting statutes: A comparative study*. London: Routledge.

MacCormick, D.N., R. Summers, and A.L. Goodhart (eds.). 1997. *Interpreting precedents: A comparative study*. London: Routledge.

Marshall, G. 1997. What is binding in a precedent. In *Interpreting precedents: A comparative study*, ed. D.N. MacCormick, R. Summers, and A.L. Goodhart, Ch. 16, 503–517. London: Routledge.

McLaren, B. 2003. Extensionally defining principles and cases in ethics: An AI model. *Artificial Intelligence* 150: 145–181.

Neumann Jr., R.K. 2009. *Legal reasoning and legal writing: Structure, strategy, and style*, 6th ed. New York, NY: Aspen Publishers.

Prakken, H. 2005. AI & Law, logic and argument schemes. *Argumentation* 19: 303–320.

Schauer, F. 1987. Precedent. *Stanford Law Review* 39: 571–605.

Schauer, F. 1989. Is the common law law? *California Law Review* 77: 455.

Summers, R. 1997. Precedent in the United States (New York State). In *Interpreting precedents: A comparative study*, ed. D.N. MacCormick, R. Summers, and A.L. Goodhart, 355–406. London: Routledge.

Sunstein, C. 1993. On analogical reasoning. *Harvard Law Review* 106: 741–791.

Walker, V. 2007. Discovering the logic of legal reasoning. *Hofstra Law Review* 35: 1687–1707.

Walton, D.N. 1996. *Argumentation schemes for presumptive reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Economic Logic and Legal Logic

Lewis A. Kornhauser

1 Introduction

In the mid-70s, Richard Posner (in Posner 1973) claimed that common law legal rules both *are* and *ought to be* efficient (in a sense to be made clear below in Sect. 2). These claims posit a deep connection between economic logic and the substance of the law. The positive claim asserts that common law legal rules are *in fact* efficient; the normative claim asserts that common law adjudicatory institutions *should* adopt rules that are efficient. The claims do not, however, assert that, in the course of reasoning about the law, judges, lawyers, or other legal officials deploy or ought to deploy economic reasoning. Empirical confirmation that common law rules were in fact efficient would not entail that common judges deploy an economic logic. The efficiency of legal rules might result from processes other than the reasoning of judges.

Similarly, a requirement that courts announce efficient rules does not entail that judges should adopt an economic logic. Given the structure of adjudication, judges might better achieve efficiency by aiming at something else. Conversely, a rejection of these claims does not imply that judges or other legal actors ought to renounce economic reasoning. One might reason economically but not pursue efficiency. The relation between economic logic and legal logic is thus an intricate one.

This essay seeks to disentangle some of these intricacies. This task presents several difficulties. First, as stated above, economic logic might have at least two distinct senses. "Economic logic" might refer to the process of economic reasoning. In this sense, the claim of the efficiency of the common law reduces to a claim about the nature of legal reasoning.

Alternatively, "economic logic" might refer to the way in which we understand the social processes that result in law. In this sense, Posner's claim about the common

L. A. Kornhauser (✉)
School of Law, New York University, New York, NY, USA
e-mail: lewis.kornhauser@nyu.edu

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_24

law reduces to a claim about the causal mechanisms underlying the institutional structures of common law adjudication rather than a claim about the reasoning of judges or a claim about the causal mechanisms underlying the institutional structures of legislation and administrative law making.

Second, and related to the first, Posner's claim interpreted as the process of economic reasoning adopts a very restrictive sense of the nature of economic reasoning. After all, efficiency, in the economic sense, is a goal or a criterion against which to assess institutional performance rather than a methodology. Economists assess the performance of institutions against the standard of efficiency. Presumably, economic reasoning refers to the process or methods of assessment not the criterion against which the institutions are measured. Further, economic methodology understood as a method of assessment or analysis is not monolithic. The logic of welfare economics may differ from the method of microeconomics which in turn may differ from the methods of macro-economics or econometrics.

Third, legal reasoning may vary across and within legal systems. The logic of a judicial decision rendered by the French Cour de Cassation may differ from the logic of a judicial decision rendered by the US Supreme Court.[1] Even within a legal system, the nature of "legal reasoning" might depend on the substantive law. One might think, for example, that the decisions in competition law should deploy economic reasoning but not the decisions in domestic relations law. On this account, then, we might endorse Posner's claims with respect to competition and contract law but not with respect to domestic relations law or criminal law. Fourth, and related to the third, the logic of justification, in both law and economics, differs from the logic of discovery. The argument in an opinion does not necessarily reflect the cognitive process that led the judge (or the court) to the announced disposition or legal rule. In what follows, I shall generally restrict attention to the logic of justification.

Fifth, the logic of legal decision extends beyond the scope of Posner's claim. Posner made a claim about the logic of judicial decision; but other actors also deploy legal reasoning and make legal decisions. Legislators, administrators, constitutional designers, and citizens, both in their role as private agents and their political role as voters make legal decisions. The logic of the law and its relation to economic logic may depend on the role the decision maker occupies or the nature of the institution in which she sits. One might argue that a legislator should deliberate about which legal rule to enact differently than she would deliberate if she were a judge (or an administrator) applying an extant legal rule. Specifically, one might contend that legislators, but not judges, should incorporate economic reasoning. Alternatively, one might, for example, contend that the institution of legislation is shaped by an economic logic. Public choice theorists and positive political theorists make claims of this type though, in their models, the economic logic consists of causal mechanisms that explain *inefficient* rather than efficient outcomes.

[1] Certainly the logic of justification differs across the two courts as a comparison of the opinions of the two courts demonstrates. The logic through which the courts reach their decisions may be more similar; see Lasser (2005). Nevertheless, given the significant differences in the institutional processes for reaching judgment, the logics of discovery are apt to differ as well.

Sixth, and perhaps most importantly, we must distinguish reasoning normatively about what our goals and values are from reasoning about the means for achieving those goals.

Economic reasoning, though it addresses some normative questions, is primarily instrumental. It reasons about the means we should use to achieve given ends and it reasons about the causes of various social phenomena. Economics typically takes the aims of individual agents and, sometimes, collective entities, as given. Legal reasoning, by contrast, often involves the elaboration or interpolation of ends. Purposive interpretation of constitutions and statutes provide the clearest examples of this.

This essay strives to disentangle these interconnections between economic logic and the logic of the law. At the outset, Sect. 2 offers a brief account of economic logic as economic reasoning; this account will be deployed in Sect. 4 through 7 which investigate the relation between legal and economic reasoning. Section 3 discusses economic logic understood as economic processes that may causally explain legal phenomena. Section 8 concludes.

2 What Is Economic Logic?

2.1 Introduction

Economics consists of many subdisciplines that encompass a variety of distinct logics and methods. I restrict attention here to two of these subdisciplines: micro-economic theory understood as an explanatory theory and welfare economics, particularly social choice theory, that investigates normative issues. I thus leave to one side both macro-economics and empirical methods. One should note, however, that legal institutions, particularly judicial institutions, are not well-designed to conduct or to evaluate empirical studies. This lack has significant consequences for the successful deployment of economic logic. Empirical studies serve two functions: they help select among different causal accounts of the relation between legal rules and institutions and the behavior that these rules and institutions seek to guide.[2] Second, empirical studies are necessary to determine or at least to estimate background conditions.

Economics is both an explanatory and a policy science. Though welfare economics has played a central role in the legal academic debates over the economic analysis of law,[3] the development of explanatory theory has played the dominant role in determining the structure of economic logic, even the logic of welfare economics.

[2] Relatedly, empirical studies may determine the relative importance of different causal mechanisms. Many theoretical models are complementary rather than exclusionary; the analysts posit a specific path of causal influence without necessarily rejecting the existence of another pathway. A particular legal rule, for example, might have both (multiple) incentive effects and transmit information. These effects may or may not push in the same direction.

[3] See, for example, Posner (1979, 1980), Dworkin (1980), and Kornhauser (1980).

The variety and complexity of economic phenomena have provoked an array of explanatory methods. I shall focus here primarily on (a subset of) the explanatory methods of micro-economic theory but one should recognize that macro-economic theories perhaps do and should play a role in some legal reasoning. Macro-economics obviously bears on legislation concerning monetary and fiscal policy. It is hard to imagine the creation of a central bank without reference to macro-economic theory or the regulation of financial institutions without some concern for the effects of the failures of financial institutions on the macro-economy.

I focus, however, on the logic of micro-economic theory for two reasons. First, most work in economic analysis of law rests on micro-, rather than macro-, economic principles. Indeed, the seminal articles—Coase (1960) and Calabresi (1961)—of the current research program in economic analysis of law[4] rest on micro-economic principles. Second, both most legal rules and micro-economic theory focus on the behavior of individual actors. Primary rules guide the behavior of individual agents; secondary legal rules typically guide the behavior of public officials. If the legal policymaker chooses legal rules in part on the basis of its consequences, then she must predict how individual agents will behave in light of these legal rules. Micro-economic theory provides a systematic framework within which to make such predictions. I discuss the logic of micro-economic explanation in Sect. 2.2 below.

The evaluative criteria that economists generally adopt have provoked significantly more criticism and outrage than the logic of micro-economic explanation. Accordingly, Sect. 2.3 outlines these common welfare criteria.

2.2 The Logic of (Some) Economic Explanation

Since Adam Smith, indeed since Bernard de Mandeville, economic logic has sought to understand how actions of individual agents might lead to aggregate outcomes that no one intended or anticipated. Both Smith and de Mandeville argued informally that individuals acting in their own interest, in fact generate public benefits. Both the style and substance of reasoning, however, has changed dramatically over time.

Since World War II, the substantive arguments have pursued parallel tracks. First, economists have sought to characterize the precise conditions under which the "invisible hand" indeed produced publicly desirable outcomes. This track culminated in the identification of the conditions under which a competitive equilibrium exists and in the first and second welfare theorems which hold that, under appropriate conditions, every competitive equilibrium is efficient and that every efficient allocation can be supported by a competitive equilibrium process.[5]

[4] The work of Coase and Calabresi provoked the second law and economics movement. An earlier attempt to apply economic reasoning to legal problems occurred at the beginning of the twentieth century and is exemplified by the work of Robert Hale and John Commons. For a discussion, see Kornhauser (2008).

[5] For an elegant exposition of these results, see Debreu (1959).

Simultaneously, the verbal tradition of Smith has given way to mathematical methods.

Virtually every article in leading journals of economics contains a formal model.[6] Legal reasoning typically eschews the formal mathematics that lies at the heart of most exercises in economic reasoning. Several key features that might more plausibly characterize economic reasoning underlie these mathematical models. Economic models abstract away from many details of the phenomenon under study. The complexity of social phenomena contrasts starkly with the simplicity of economic models.

In addition, theorists reason deductively from the assumptions of the model to conclusions. Most such models assume that all agents are rational in a sense to be made more precise below. Actual decision makers, of course, are rarely fully rational; and in some instances could hardly be described as rational at all.

The basic idea underlying economic rationality is most easily understood in the context of decision making under certainty. In this context, economic rationality requires that each agent have a complete ranking of outcomes.[7] A rational agent then seeks to achieve the best outcome that she can. In environments in which the agent acts alone against nature, this problem is relatively straightforward. In environments with strategic interaction, however, the problem becomes significantly more complex and the concept of rationality reflects the additional complexity.[8]

This bare description isolates an important feature of economic reasoning. Economic rationality is *instrumentally* rationally; economic reasoning typically takes the agent's goals as given. Legal actors might thus deploy economic reasoning in pursuit of *any* goal. We might thus observe judges, for example, deploying economic reasoning to further non-economic ends.

The use of formal models characterizes the *style* of economic reasoning but not its substance. Substantively, economic analysis focuses on the incentives and information available to individual agents who act rationally. Rationality implies that an agent makes decisions on the basis of her beliefs about the costs and benefits that she derives from her actions. She adjusts her beliefs, rationally in the light of the evidence available to her[9] and her choice is an optimal one. Within micro-economic theory, we may usefully distinguish behavioral from rational choice approaches.

[6] The text focuses on theoretical models. Most empirical papers also have an econometric model that allows one to interpret the data. These models too will be formal but they may present different issues than those discussed in the text.

[7] Under risk and uncertainty, the agent must have rankings over actions. Economic theory, however, has identified the conditions under which such rankings can be represented by a utility function that ranks outcomes and, in the case of uncertainty, a set of beliefs about states of the world that conforms to the probability calculus. In either case, an agent may calculate the "expected utility" of an action, i.e., the weighted sum of the valuations of the outcomes in each state of the world. The agent ranks act A more highly than act B if and only if the expected utility of act A exceeds the expected utility of act B. When the simple model of decision making under certainty is made more complex, there are some complications.

[8] The diversity of solution concepts in both cooperative and non-cooperative game theory illustrates this complexity. Each solution concept, after all, reflects a distinct conception of rationality.

[9] That is, she adjusts her beliefs according to Bayes' rule.

These approaches share some features. First, each predominantly deploys formal models to isolate specific causal mechanisms. I shall largely ignore this feature of micro-economic logic as legal reasoning is rarely formal. Second, each usually considers the agent's ends as given; both behavioral and rational choice micro-theories are instrumental. Third, each explains social phenomena in terms of individual decisions. They differ in their analysis of individual decision making. Neo-classical or rational choice micro-theorists assume that individuals are rational in the sense that they have well-defined preferences over the relevant domain and that they maximize their preferences given the constraints under which they act.[10] Behavioral economists recognize that agents may systematically deviate from these rationality assumptions; they investigate the consequences of such deviations.

2.3 The Logic of Economic Evaluation: Welfare Economics

2.3.1 Normative Economics Generally

Welfare economics addresses the evaluation of economic institutions, processes, and outcomes. Welfare economists have predominantly adopted a specific set of evaluative criteria rather than a distinctive set of evaluative methods. In this subsection, I confine myself to two remarks—one substantive and the other methodological—about welfare economists.

Substantively, economists typically adopt *welfarist* criteria to evaluate economic phenomena.[11] I discuss these in Sect 2.3.2.

Methodologically, much welfare economics proceeds axiomatically in two directions. One set of projects identifies a particular normative criterion such as utilitarianism and then offers an axiomatic characterization of that criterion. One might understand Harsanyi's theorem in this fashion.[12] Similarly, D'Aspremont and Gevers (1977) provide axiomatic characterizations of utilitarianism and leximin, a

[10] In analyzing decisions under certainty, preferences are over consequences or outcomes; in analyzing decisions under risk, preferences are over lotteries; and in analyzing decisions under uncertainty, preferences are over actions. One of the central projects of the latter two theories concerns the proof of representation theorems which state the conditions under which the preferences over lotteries or actions can be represented by preferences over consequences or actions and beliefs about the likelihood of each outcomes so that the agent prefers lottery L (or action A) to lottery L$'$ (or action A$'$) if and only if the expected value of lottery L (or action A) using the relevant preferences over outcomes and probabilities of outcomes exceeds the expected value of Lottery L$'$ (or action A$'$). Throughout, unless otherwise stated, I shall assume that the relevant conditions are met and distinguish between the agent's beliefs and her "desires" or preferences. Notice that assumption of rationality requires rational belief formation.

[11] Not every economist adopts welfarism. Some investigate libertarian, contractualist, or rawlsian principles. But normative economic studies are called "welfare economics" for a reason.

[12] See Harsanyi (1955). Broome (1991) offers a philosophical defense and interpretation of this theorem; Broome (1990) proves a similar theorem in a somewhat different framework. The theorem states that if each individual has "coherent" preferences and social preferences are also coherent

quasi-Rawlsian criterion according to which policies should maximize the well-being of the least well-off.[13]

Conversely, economists begin with a set of axioms and ask which set of normative criteria simultaneously satisfy each of the axioms. Arrow (1963) famously introduced this method with a demonstration that no criterion satisfied a small, arguably plausible set of axioms. Similarly, John Nash characterized a baseline bargaining solution using an axiomatic method. Subsequent work has investigated the consistency of the Pareto criterion with ideas of minimal liberty Sen (1970) and with various notions of responsibility (Fleurbaey 2005). This line of research bears on the *aim* of legal reasoning as it addresses the desirability of the Pareto efficiency as a normative goal for legal systems, but it does not address the process of legal decision making. These methods, though, very powerful and illuminating, play only a limited role in legal argument.[14]

2.3.2 Welfarism and Efficiency

Welfarism is a form of consequentialism that assesses the social value of an outcome, an action, a policy or an institution solely on the basis of the well-being of individuals. Assessment of outcomes, however, is fundamental; the consequentialist evaluation of actions, policies, and institutions rest on the evaluation of the outcomes that these actions, policies, or institutions may cause.

Forms of welfarism may differ in two respects: (1) on the nature of individual well-being—i.e., on which consequences matter, and (2) in the functional form that integrates the well-being of each individual into social value. Economists typically adopt a subjective account of well-being that identifies the agent's well-being with a ranking of alternatives. For evidentiary reasons, economists often identify the agent's well-being with her motivational preferences, also represented by a ranking. An agent presumably has better information about her (subjective) well-being than external observers; and the identification with her motivational preferences permits the analyst to infer the agent's well-being from her behavior. This inferential basis underlies various implementations of cost-benefit analysis.

Economists typically represent an agent's well-being by a ranking that itself can be represented by a "utility function." Usually, economists identify this "well-being ranking" with the agent's motivational preferences; this identification underlies the first and second welfare theorems (see, e.g., Debreu 1959) which state that, under suitable conditions, (1) every competitive equilibrium is a Pareto optimum; and (2) every Pareto optimal allocation can be supported by a competitive equilibrium.

and satisfy the Pareto criterion (defined below), then social preferences can be represented as the sum of the representations of the individual preferences.

[13] The criterion is "quasi-Rawlsian" because Rawls himself focused on the space of primary goods not the space of well-being.

[14] Of course, legal analysts and decisionmakers might rely on come of the conclusions of these axiomatic investigations.

These theorems make sense only on the assumption that the preferences that underlie individual choices correspond to the individual's well-being.

Posner claimed that the rules of the common law both should be and in fact are "efficient." As noted above, evaluation of this claim requires that we resolve at least two important ambiguities. First, we must determine, in the positive case, whether the efficiency of the common law arises from structural features of the system of adjudication or from the intentionality of individual judges. We must make a similar determination for the normative claim: should judges aim at efficiency or should we design adjudicatory institutions that yield efficiency (even if, in the best design, judges do not aim directly at it)?

Second, we must determine what "efficiency" means. At an individual level, the central question of a normative theory of adjudication asks: what should judges want? The central question of a positive theory of adjudication has a parallel structure: what *do* judges want? At the systemic level, the central question of a normative theory of adjudication asks: what should adjudication achieve? Similarly, at the systemic level, the central question of a positive theory of adjudication asks: what does adjudication achieve. This section addresses this second, normative question of the aim of the judge or the adjudicatory system.

In elaborating his claims, Posner interpreted the call for "efficiency" as a call to maximize social "wealth." Posner (1979, 1980) argued that wealth differed from utility and had independent moral value; these arguments were heavily criticized—see for example Dworkin (1980), Kornhauser (1980), Bebchuk (1980), among others —and largely unsuccessful. Here I sketch two accounts of Posner's idea of efficiency. The first takes Posner's characterization of the judicial task as *wealth* maximization; it interprets *wealth* in terms of cost-benefit analysis.

The second takes the term "efficiency" seriously and interprets the Posnerian injunction in terms of the neo-classical conception of Pareto efficiency.

The best interpretation of Posner's claim that a court should maximize wealth interprets "wealth" in terms of individual valuations of the relevant policies or legal rules. These valuations may be understood as a form of welfarism in which well-being is interpreted as preference satisfaction. Indeed, Posner's conception of wealth seems to correspond closely to the conception of value embedded in cost-benefit analysis. Cost-benefit analysis, too, is controversial (see, e.g., Nussbaum 2000; Richardson 2000; Anderson 1995; Satz 2010). I discuss it more fully in Sect. 5, in the context of agency rule-making and legislation rather than adjudication. Here it is sufficient to provide a bare characterization of the formal theory underlying cost-benefit analysis.[15]

Economic theorists assume that each agent has a well-defined preference over consequences. A preference is a complete linear ordering of the objects in the domain. The agent, that is, can rank any two objects A and B in the domain; either she prefers

[15] The practice of cost-benefit analysis relies on a formal theory that derives social preferences from individual preferences, an implementation theory that seeks to implement the theoretical apparatus by identifying measures of individual and social preference, and applications of these two theories to particular problems. This section sketches the formal theory and a few objections to this theory. Section 5 discusses the theory of implementation and some objections to that theory.

A to *B*, or she prefers *B* to *A*, or she is indifferent between them. In addition, her preference is transitive: for any three objects *A, B,* and *C* in the domain, if the agent prefers *A* to *B* and she prefers *B* to *C*, then she prefers *A* to *C*. Under an additional, technical assumption, this preference can be represented by a (n utility) function $u(.)$ that assigns a number to each element of the domain of preference such that *A* is preferred to *B* if and only $u(A) > u(B)$. The information embedded in the preference, and hence the utility function, is *ordinal* only; it contains no intensity information.[16]

Cost-benefit analysis begins with a description of the domain of preference.[17] Each agent has (ordinal) preferences[18] over (*policy, wealth*) pairs; the two elements describe the complete state of the world under each policy with the wealth element indicating the agent's wealth under that policy. One may represent these same preferences by a utility function that assigns an amount of money *m* to each policy that represents the agent's willingness to expend for that policy.[19] Cost-benefit analysis sums the willingness to expend of each agent to arrive at the "social value" or "wealth" in Posner's terms.

The best interpretation of Posner's claim understands "efficiency" in the traditional economic sense of *"Pareto efficiency."* To understand this term, we must begin with the more basic concept of the Pareto criterion. The Pareto criterion is an *aggregate, welfarist* criterion that compares "states of the world" on the basis of the individual agents' rankings of these states of the world.[20] The Pareto criterion ranks a state *R*

[16] Additional assumptions are required to represent a preference over uncertain prospects by a set of beliefs satisfying the probability calculus and a cardinal preference over outcomes such that the agent prefers prospect L to prospect L' if and only if the expected utility of L exceeds the expected utility of L'.

[17] Kornhauser (Chap. 5, part II, this volume, on "Choosing Ends and Choosing Means: Teleological Reasoning in Law") elaborates more fully on the nature of the domain of preference and how it may differ from the domain of choice. Kornhauser (ibid.) elaborates more fully on the nature of the domain of preference and how it may differ from the domain of choice.

[18] Preferences are ordinal when only the agent's ranking of two alternatives matter; there is no measure of intensity of preference.

[19] More precisely, we first pick a baseline policy-wealth pair (*P0, w0*) from which we may assess the agent's preferences. Consider some other policy-wealth pair (*P1, w1*). Let *m1* be the amount of money one would have to give the agent (or the amount the agent would be willing to pay) to replace *P0* with *P1*; i.e., the agent is indifferent between the two policy-wealth pairs (*P0, w0−m1*). We may proceed in this fashion to assign to each policy *Pj* a monetary amount *mj*. The agent then prefers policy *Pj* to policy *Pk* if and only if $mj > mk$. Notice that the monetary amounts assigned to each policy will differ if we change the baseline policy—the baseline policy always has $m = 0$ assigned to it—but that the order of policies remains unchanged.

[20] I noted above that each agent has a ranking of elements in her domain of preference. Policy assessment on the basis of these rankings makes several assumptions. First, it assumes that each agent has an identical domain of preference; i.e., each has preferences over the same objects. Of course, agents may rank these objects differently. Indeed, they may differ as to which aspects of the elements in the domain are important. Second, policy assessment assumes that each agent's ranking of these elements reflects the agent's relative well-being in each state of the world. Third, at least for purposes of implementations of cost-benefit analysis, assessment assumes that each agent's ranking corresponds to the agent's all-things—considered motivations; from any given choice set, the agent selects the option she ranks most highly.

over a state *S* if *every* agent ranks *R* over *S*.[21] In this instance, one says that "*R* is Pareto superior (or Pareto preferred) to *S*". A state *R* is Pareto efficient (or Pareto optimal) if and only if there is no other state *S* that is Pareto superior to *R*. Notice that the Pareto criterion does not rank all pairs of states. For some pairs of policies *P* and *Q*, some agents will prefer *P* to *Q* while others will prefer *Q* to *P*; in these cases, *P* is Pareto non-comparable to *Q*.

The Pareto criterion has a strong intuitive appeal. Some authors, Kaplow and Shavell (2002) most vigorously, have relied on that appeal to defend welfarism, the position that the evaluation of social states should depend solely on the well-being of individuals in a social state. Several concerns about this argument are worth noting. First, formally, the Pareto criterion simply aggregates a set of "individual" rankings into a (partial) "social" ranking. The substantive content of this formal aggregation depends on one's interpretation of the underlying individual rankings. These rankings might be anything: values that bear on the decision of a case, voter rankings of candidates, orders of finish in athletic events.

Economists typically identify the underlying ranking with an agent's well-being[22] but one might interpret the ranking as an all-things-considered evaluative judgment or in terms of the agent's "good" rather than her well-being (see e.g. Broome 1991). There is thus no necessary connection between the formal structure of the Pareto criterion and welfarism.

Second, the intuitive appeal of the Pareto criterion derives from its structure as an "unanimity requirement;" the criterion endorses a policy only if each person ranks it more highly than the status quo. Formulated in these terms, the Pareto criterion has a strong intuitive appeal as a normative criterion in part because it trades on two distinct normative understandings of "unanimity." On the first account, "unanimity" relies solely on what is often called the "principle of personal good" (see, e.g., Broome 1991 or Temkin 1993) which states that if P is socially better than (or ranked more highly) than Q, then P must be better for *someone*. The Pareto criterion understood as a unanimity criterion says more strongly that if P is better for *everyone* than Q then P must be socially better than Q. Taking the contrapositive we have if P is not socially better than Q then P must not be better than Q for someone, a reformulation of the principle of personal good. The principle of personal good has strong appeal, but it does exclude the possibility of communal good.[23]

[21] Or somewhat more strongly if every agent thinks *R* is at least as preferred as *S* and at least one agent strictly prefers *R* to *S*.

[22] In fact the identification is more complex. Economists typically identify the agent's ranking either with her choices (as in revealed preference theory) or with her motivational "preferences" generally understood as her self-interested preferences. They then identify the agent's well-being with this behavioral ranking. In the standard economic context of agents choosing consumption bundles, this elision of motivational and evaluative preferences is reasonably plausible but it becomes more suspect in non-market contexts.

[23] Equality often serves as the standard, if disputed, example of a communal good (see Temkin 1993; Broome 1991). Notice that the principle of personal good is not a requirement of welfarism. A criterion that cares about the average well-being of individuals in society and about the degree of dispersion of well-being in that society will be welfarist as it evaluates social states solely on

On the second interpretation, "unanimity" suggests consent; if each agent ranks P over Q, then each agent consents or would consent to a move from Q to P. One easily, but erroneously, slides from the claim: "agent *A* ranks policy *P* over the status quo policy *Q*" to the claim "agent A consents to the change of policy Q to policy P." Then, when *each* agent ranks *P* over *Q*, surely society should rank *P* over *Q*. Consent, or even hypothetical consent, to the move to *P* from *Q*, however, does not follow from the fact that the agent ranks *P* over *Q*. After all, the agent might prefer a third policy *R* to *Q*, with *R* Pareto non-comparable to *P,* as well. Why should we assume that the agent consents to a move to *P* rather than to a move to *R*? To be more concrete, suppose there are only two individuals in society and that *P* allocates all gains to individual 1 while *R* allocates all gains over *Q* to individual 2.

The interpretation of the unanimity feature of the Pareto criterion as consent suggests an interpretation of the underlying individual rankings as the *motivational* ranking of each agent. The ranking represents how the agent would choose among alternatives. But, again, choice, even hypothetical choice, is not equivalent to consent.

Third, we must carefully specify the environment against which we measure efficiency of a particular arrangement. The discussion thus far has implicitly assumed a world of certainty.

The actions of all agents yield a certain outcome that is known to all agents. Agents rarely, if ever, act in such an environment. More typically, the consequences of agents' decisions will depend on which state of the world is realized. A wheat farmer's income depends not only on the amount of wheat she plants and the effort she takes in cultivating it but also on the weather. Similarly, legislatures enact policies in a world of uncertainty. The consequences of a climate change policy may depend on the rate of technological progress in different clean energy sources.

In a world with uncertainty, we may distinguish two conceptions of efficiency that vary with the timing of the assessment of consequences. *Ex ante* efficiency evaluates actions or policies before the state of the world is realized. Suppose, for example, that there are two possible states of the world s and s′. The policymaker must choose between two policies P and Q that have different consequences in each of the two states.[24] Assume further that the policymaker evaluates policies on the basis of the preferences of the citizenry.

We may now define an *ex ante* Pareto criterion. Policy P is *ex ante* Pareto superior to policy Q if each citizen *ex ante* ranks policy P at least as highly as policy Q and at least one citizen ranks it strictly higher. A policy P will be *ex ante* Pareto efficient if there is no policy R that is *ex ante* Pareto superior to policy P.

the basis of the well-being of individuals. But society may rank one state in which everyone has equal well-being over another in which some have very high well-being and others relatively low well-being even though each member of society prefers the unequal society to the equal one.

[24] Further analysis requires an account of how the policymaker ranks these alternatives. If Policy P leads to better outcomes than policy Q in both states s and s′, then, clearly, the policymaker should rank P above Q. But if P leads to a better outcome in state s than Q but a worse outcome in s′ than Q, different criteria for ranking the two states exist. Kornhauser (Chap. 5, part II, this volume, on "Choosing Ends and Choosing Means: Teleological Reasoning in Law") for a brief discussion of some of these criteria. The text adopts the standard criterion: expected utility maximization.

Ex ante Pareto efficiency contrast with *ex post* Pareto efficiency. An *ex post* Pareto criterion evaluates policies on the basis of the realized state of the world. *Ex post* efficiency, that is, corresponds to efficiency defined under certainty for each possible state of the world. Clearly, an action that is *ex ante* efficient may not be *ex post* efficiet, for some—possibly all—realizations of the state of the world. Suppose, for example, that there are two, equally likely states of the world s and s′ and three actions. Action A yields the payoffs (4, 1) in states s and s′ respectively: action B yields the payoffs (1, 4); and action C yields the payoffs (3, 3). Then action A is *ex post* efficient when state s is realized while action B is *ex post* efficient when state s′ is realized. But state C is *ex ante* efficient as it has an expected value (indeed a certainty value) of 3 while actions A and B have expected value of only 2.5.[25]

Finally, many policies are more complex than simple actions; they establish *institutions*. An institution must operate in many different environments. We may demand that the institution work well in many or all of these environments. This demand may give rise to additional problems related to those that arise when uncertainty is introduced.

To understand these additional complexities, we must think more carefully how an institutional environment may vary. Crudely, one might say that the environment might vary in terms of the motivations (and evaluative preferences) of the population affected by the institution, in terms of the motivations of the population *inside* the institution, or in terms of external constraints under which the institution operates.[26] External constraints includes the extant of resource abundance available.

This structure introduces a third type of efficiency called *interim efficiency* that arises from the introduction of a second uncertain element.[27] Initially we considered uncertainty over the state of the world understood as, say, the nature of the weather. The environment now considered also includes uncertainty about the motivations or types of individuals who will inhabit the (level 1) uncertain world.

The structure of institutional design provides a clear illustration of this additional complexity. Consider the situation of the policymaker as institutional designer. She acts *ex ante*, before any uncertainty is resolved. She knows neither the distribution of motivations in the population in which the institution will operate nor what state of the world will be realized. After the institution is created, the motivations of both its personnel and the citizenry will be determined. We must now specify the nature of the uncertainty that these agents face.

[25] The expected utility calculations in the text assume that the agent is risk neutral. If the agent is risk-averse the *ex ante* appeal of action C is even stronger.

[26] This characterization of the environment is crude because an institution may in part influence the environment in which it is situated. Its selection (or hiring and retention) policies, for example, will determine in part the motivations of its personnel. Similarly, the institution may create incentives that structure the motivations of the population that the institution serves.

[27] In fact, interim efficiency arises in the evaluation of institutions. It results from the nature of the uncertainty about agent types. Typically, however, there may be no *ex ante* stage in which no agent knows her own type and thus cannot act *ex ante*; rather, the agents act in an environment of asymmetric information in which each knows her own type but only the distribution of types from which her opponent's type is drawn.

Clearly, neither group knows the underlying state of the world so they act *ex ante* relative to that. Similarly, each agent presumably knows her own motivations or her own type but does she knows the types of all other agents in the population? If each agent knows the type of herself and every other agent, then we are in the situation of uncertainty analyzed above. More typically, however, each agent knows her own type but not the types of other agents. When *S* decides to sell her apartment, she knows the lowest price that she is willing to accept but she doe not know the highest price that any specific buyer *B* is willing to pay. Consequently, though gains from trade may exist, *S* and *B* may fail to reach and agreement.

One might define naively define a concept of efficiency under asymmetric information directly in terms of the realized types of each individual. One would do so in the obvious way: a decision d dominates a decision d′ if each agent ranks d at least as highly as d′ and at least one agent ranks d more highly. The decision d would then be efficient if there were no decision d" that dominated it.

The Pareto criterion has very strong implications. It is inconsistent with the existence of any communal good and with the value of autonomy understood variously as a concept of self-government, as a concept of liberty, and as concept of responsibility.[28]

In what follows, however, I shall interpret "efficiency" generally as "Pareto efficiency" and I shall not contest its importance.

3 Economic Logic as a Social Process Rather Than a Reasoning Process

Economists have offered two different procedural accounts of how common law rules might be efficient. One account relies on the general social mechanism of *exchange*. The second examines non-judicial features of the process of adjudication. The claims for these procedures are quite limited; actual social mechanisms are unlikely to satisfy the conditions that ensure that the processes lead to efficient outcomes. In conclusion, I address and dismiss a third set of processes of aggregation of judicial judgments.

3.1 The Power of Private Ordering

In the early 1960s, Coase (1960) offered a simple but highly influential argument about the *irrelevance* of (certain) legal rules.[29] Using some simple numerical

[28] The arguments behind this claim are complex and set out at length in Kornhauser (2017, para 6.2).

[29] Coase (1959) had made a similar argument in the context of the allocation of the electro-magnetic spectrum. His argument generated much controversy, particularly after Stigler (1966) formulated "the Coase theorem."

examples, Coase argued that, when transaction costs are zero, private exchange leads, for any assignment of the liability rule, to a Pareto efficient outcome. Thus, on Coase's account, common law liability rules are efficient because they are irrelevant to the achievement of (Pareto) efficiency.

For Coase, bargaining or exchange is the social process that insures the achievement of efficiency. Coase contends that, when transaction costs are zero, parties will exhaust all potential gains from trade. If agent A values an entitlement more than agent B, the holder, A will pay for the transfer of that entitlement.

A proper understanding of Coase's argument takes some care. Notice first that only common laws rules that *assign* entitlements are irrelevant. Rules that govern the *transfer* of entitlements matter a great deal.[30] Courts must enforce whatever contracts the parties enter. They must also prevent non-consensual transfers of entitlements. The efficiency of rules governing property and tort thus depend on the efficiency of the rules governing contract and the security of property interests, even in a world of zero transaction costs.

Second, though the assignment of the entitlement is irrelevant to the achievement of efficiency, it may matter greatly to the parties. After all the assignment of the entitlement determines the relative wealth (or bargaining power) of the parties. Even in a Coasean world, then, entitlements have distributional significance. Legal reasoning, in this world, would then be divorced from central concepts of economic reasoning; judges and legislators assigning entitlements would focus exclusively on non-efficiency concerns—distribution and perhaps deontological matters—because, in a world of zero transaction costs, efficiency would take care of itself.

Third, Coase (1960) offers no elaboration of the concept of "transaction costs." We may usefully distinguish between two classes of transaction costs: *resource* costs and *strategic* costs. In a world of zero transaction costs, agents would face no resource costs; they would not have to expend time negotiating or money to communicate with their counterparties or to draft, monitor and enforce contracts. Resource costs, in the real world, of course are not zero; exchange requires time and money. An agent driving down the road cannot costlessly negotiate with the pedestrian about to dash across the street from between parked cars. Similarly, the monitoring and enforcement of entitlements and contracts to transfer entitlements are costly. Consequently, the allocation and structure of the entitlement will have efficiency as well as distributional implications. A role for judicial insight into efficiency thus exists.

Strategic costs may be either pure or informational. Informational costs are easier to understand. If information costs were zero, each party would have the same knowledge and understanding as the others. Typically, however, a buyer knows her own valuation of an entitlement but not the seller's valuation while the seller knows his own valuation of the entitlement but not the buyer's. This asymmetry of information

[30] I have shifted from Coase's formulation in terms of *liability rules* to one in terms of *entitlements*. Calabresi and Melamud (1972) introduced a distinction between an *entitlement* and its *protection*. Common law rules typically protect an entitlement with either a *liability* rule that gives the court the power to set the price at which the entitlement transfers and a *property* rule under which the holder of the entitlement sets the transfer price. Whether the shift in terminology matters may depend on one's interpretation of the term "transaction costs."

implies that inevitably some gains from trade will not be realized.[31] The party that values the good most highly may thus not end up with the good; final holding of the entitlement will not be efficient. Ideally, the law would, in these situations, allocate the entitlement, to the party who values the entitlement more highly. Again, there would be a role for a judiciary that reasoned economically.

Pure strategic costs may arise even when all parties have complete and hence symmetric information. An agent's strategic situation derives from the structures that govern her interaction: the preferences of the other agents, the choices available to her and to the other agents, and the underlying technology. Indeed, the allocation of the entitlement determines in part the agents' strategic situation. Strategic costs are thus not independent of the allocation of the entitlement. Not surprisingly, then, the allocation of the entitlement may affect the efficiency of the outcome.[32]

Strategic costs have a different character than resource costs. Generally, resource costs do not depend on to whom the entitlement is assigned. The assignment of the entitlement, however, will often determine the nature and extent of the strategic costs that the agents face.

Coase (1960) of course did not argue that common law legal rules, or even rules governing entitlements, were efficient. To the contrary, he recognized the pervasive presence of substantial transaction costs that made the allocation of the entitlement (and the choice of its protection) central to the achievement of Pareto efficiency. The efficiency of common law legal rules, if they are in fact efficient, must thus be explained by some other process.

3.2 Blind Justice: The Social Processes of Legal Development

Many critics of economic analysis of law argued that common law judges rarely explicitly invoked economic reasoning or economic facts in their opinions. That the common law rules were in fact efficient was thus highly unlikely. Paul Rubin (1977); George Priest (1977) then offered formal "evolutionary" models of the legal process that explain how "efficiency-blind" judges might nonetheless end up announcing efficient common law rules.

In both models, the judicially announced rule tends toward efficiency because, by a *differential litigation* assumption, more efficient rules are *less* likely to be litigated than less efficient rules.[33] This assumption, however, is not sufficient to insure that the legal process tends toward efficiency. Rubin adopts the additional, and implausi-

[31] See Myerson and Satterthwaite (1983) for a proof of this claim.

[32] Benoît and Kornhauser (2002) provides an example in which, under one allocation of an entitlement, the players achieve an efficient outcome while under a competing entitlement, the outcome is inefficient.

[33] In fact, nothing in these arguments rests on the rules being *efficient*. The models simply require that the rules be ranked with lower ranked rules litigated more often than more highly ranked rules. This formulation highlights the crucial step in the argument: why are lower-ranked rules litigated more often than more highly ranked rules?

ble assumption that efficient rules are never litigated. Generally, however, multiple efficient rules exist and each has different distributional consequences. Thus, even when an efficient rule prevails, there will be some potential litigants who would be better off under an alternative efficient rule (or even better off under a different, inefficient legal rule). These distributionally dissatisfied individuals would have reason to litigate even an efficient rule. Thus, an efficient rule, once announced, might not persist, as Rubin assumed, forever.

More generally, in many contexts, we have reason to believe that more efficient rules may not be litigated more often than inefficient rules. In an article that preceded the evolutionary models discussed here, Galanter (1974) argued that, in may areas of law, stakes were asymmetric with one party having significantly greater incentives to develop the law favorably to themselves. In product liability cases, for example, we might expect that manufacturers, as repeat players, have greater incentives to develop the law favorably to themselves. Plaintiffs, after all, are likely to be harmed only once.[34] Of course, this asymmetry in stakes does not imply that the common law rules will not be efficient. Manufacturers would like the most favorable rule possible and that rule might itself be efficient.

To insure that differential litigation resulted in the selection of an efficient rule, Priest assumed that judges selected one of only two possible rules. In this setting, the more efficient rule prevails more often than the inefficient rule. With more than two rules, however, this efficient result may not hold (see Kornhauser 1996 for details).

Subsequent research has developed in several directions. Cooter and Kornhauser (1980) set the evolutionary argument in a more natural Markovian setting and suggest that, in this more general setting, the tendency to efficiency does not necessarily follow.[35] Goodman (1978) identifies a different selection mechanism: the relative investments in litigation which affect the likelihood that a given rule will be adopted.

Another line of research investigates the evolution of the law when judges are, at least to some extent, seeking efficiency. Hadfield (1992) argues that judges who consciously pursue efficiency will not achieve it because the set of cases litigated and then appealed is a biased sample of the universe of cases. Judges thus do not discover the efficient rule. Cooter, Kornhauser and Lane (1979) present a model in which a common law court adjusts a standard of care using an economic logic; for suitable technologies of care, this rule of adjustment converges to the efficient standard. Fernandez and Ponzetto (2008), following Gennaioli and Shleifer (2007) consider a model in which some judges pursue efficiency but other judges are biased either for plaintiff or defendant. Distinguishing a prior case is more costly the more radical the deviation. In this model, the law tends toward the efficient rule.

[34] In this instance, a plaintiff's bar exists that does have repeated interactions with manufacturers and thus an interest in developing doctrine favorable to plaintiffs. Plaintiff's attorneys do not capture the entire benefit for a favorable doctrine; they will thus have less incentive than plaintiffs themselves to develop the law favorably.

[35] Kornhauser (1996) shows that the prior literature had conflated two conjectures; a strong and a weak one. Cooter and Kornhauser (1980) provide a counterexample to the strong conjecture but the weak conjecture remains open.

The interest of this literature stems from its analysis of the role of litigant selection in driving the development of the law and less in the specific claims about the "efficiency" of the resultant rules. The importance of litigant selection to the development of the law is difficult to underestimate and to evaluate.

3.3 "Efficiency" on Collegial Courts

The two prior subsections considered non-judicial processes that might generate efficient common law rules. In this subsection, I consider a third process that relies on the competence of judges but leaves the reasoning process unexplicated.

Consider a collegial court composed of n judges, none of whom actively pursues efficiency. It is imaginable but highly implausible that the court would nonetheless announce efficient rules; the process of aggregating the independent views of the n judges would somehow yield an efficient rule. The structure of such an aggregation process, however, is unknown and difficult to construct.

So, imagine instead that each judge on the collegial court seeks to promote efficiency understood as the correct resolution of the case. The judges disagree over which disposition (or which rule) is correct. Suppose that each judge is more likely than not to get the correct answer. Then, under a suitable voting procedure, the court will announce the correct disposition.[36]

3.4 Legislative and Other Processes

Economists have also used economic logic to analyze the behavior of public officials generally, and legislators in particular. These inquiries assume that public officials act as private individuals do; they maximize their utility subject to constraints. In the legislative contexts, this assumption often leads the analyst to assume that each legislator seeks to ensure her re-election or seeks to maximize a preference that combines policy preferences with a preference for re-election.

Two classes of models exist that correspond, roughly, to the two sides of the "market" for legislation. Public choice theory essentially studies the demand side of

[36] The correct voting procedure requires that the judges understand that they are "deciding together," and hence each must act differently than she would were she deciding alone. For this distinction see Kornhauser (2013). Under this decision procedure, the Condorcet jury theorem applies. If the judges are choosing among more than two legal rules, the argument is a bit more complex with plurality rule substituting for majority rule (see Dietrich and List 2004). Many models of collegial courts assume that the court chooses a legal rule from a one-dimensional space of legal rules according to a condorcet-consistent procedure. More specifically, the court uses majority rule with the appropriate amendment process the median voter theorem applies. The court will announce the rule of the median judge which is a Pareto efficient rule but then, in a one-dimensional policy space, every rule the cut-point of which lies in the interior of the interval spanned by the ideal points of the judges, is Pareto efficient.

the market for legislation; crudely, the analyst holds political institutions constant, and asks how the content of legislation varies with the distribution of preferences within the relevant population (see Buchanan and Tullock 1962). Positive political theory, by contrast, focuses on the supply side of the market; again crudely, the analysts hold the distribution of preferences within the relevant population constant and asks how the content of legislation varies with the electoral and legislative institutions.

The logic of these processes does not yield efficiency. But the processes do have important consequences for the structure of legislation and our understanding of law.

Specifically, the models suggest that, typically, legislation will be incoherent in the sense that it does not represent a coherent plan; rather it will reflect the different interests of competing factions (under public choice theory) or reflect the interests of skilled, agenda-controlling legislators (positive political theory). From a legal theoretic perspective, these insights suggest that we require a more complex understanding of what instrumentalism about law means.[37]

4 Situated Legal Reasoning

When lawyers and jurisprudes discuss legal reasoning, they typically mean reasoning that is done by or directed to judges. Laws and reasoning about and with the law, however, is ubiquitous. Constitutional designers create complex legal structures; presumably they must in part use legal reasoning to do so. Similarly, citizens, in their role as individual agents, must, if they are to act either prudently or morally, know what the law requires of them. In their role, as citizens who participate in the polity, these same individuals must, or at least should, consider what legal rules the polity should have. Even if, in the first role, they act as the paradigmatic Holmesian "bad man," they may still use legal reasoning to predict how the courts will treat them.

More significantly, public officials must create, administer, and apply the law. To do so, they must reason legally, or at least consult a lawyer. Though theorists typically treat the logic of legal reasoning as uniform, it is not clear that the skill of the legislative (or contract) drafter is identical to the skill of the litigator or judge. The two sets of actors are situated very differently and how each reasons is surely determined in part by the problems posed by their respective positions. Rawls (1955) made this point when he distinguished rule application from rule selection or justification.

For legislators, administrators, and adjudicators, a large portion of legal logic concerns the elaboration and the interpolation of ends. The legislator asks what legal rules should we have? This question has two aspects. First, the legislature must decide what ends to pursue. Second, they must determine how to pursue those ends. This second task requires the use of instrumental logic and economic logic provides a useful tool for the shaping of legal means to achieve legislative ends. The first task, however, entails reasoning about our ends; economic logic here provides little

[37] For a longer discussion of instrumentalism, see Kornhauser (2000, 2010).

guidance though, as noted above, it argues substantively that our ends should be welfarist.

For judges, who must apply the law, the interpolation of ends plays the central role as rules are necessarily incomplete so that the law appliers—administrators and judges must often interpret legislative and constitutional texts to determine what ends the legal regime pursues. The second, instrumental task of finding appropriate means, plays a more limited role in adjudication as adjudication is inherently backward-looking while rule elaboration is primarily forward-looking.

Judges and litigators deploy legal reasoning to resolve disputes. The nature of common law reasoning as well as reasoning in civil law countries is much contested.[38] The debate, however, tends to focus narrowly on the courts as law makers rather than law appliers. This focus elides the differences between the logic of legislation and the logic of adjudication.[39] We often say that each must *apply* the law to the facts before them. I shall call this reasoning judicial reasoning. What application of the law entails precisely is unclear and relatively unexplored. At a minimum, however, the judge must pay close attention to the facts; judicial reasoning is thus backward looking.

Administrative rule-making constitutes the most appropriate forum in which to study the second, instrumental stage of rule-making. Delegation to administrative agencies is often justified on the ground that a given realm—e.g., health and safety regulation or environmental regulation—requires a significant amount of technical expertise. The statutory framework thus establishes a set of performance goals and constraints that the administrative must elaborate into a set of operative norms that govern primary conduct.

5 The Logic of Legislation

Legislators perform a wide variety of tasks; they hold hearings, they deliberate, they serve constituents, and they legislate. Legislation itself takes many forms. In the USA, some legislation simply expresses “facts, principles, opinions and purposes” of one or both houses of Congress (Sullivan 2007, 7). Legislative enactments may take many other forms as well. A legislature may enact a *program* or a *project*, or it may make an *appropriation*.

A *project* authorizes the production of a commodity or the provision of a service. In 1956, for example, the US Congress enacted the Federal-Aid Highway Act which authorized the creation of the interstate highway system in the USA, essentially an immense public works project managed and directed by an office in the federal

[38] On common law reasoning, see e.g., Llewellyn (1951), Levi (1948), Eisenberg (1988), Schauer (2009).

[39] In addition it is useful to differentiate the logic of justification and the logic of discovery as well as normative and positive accounts of each. The typical legal theory of adjudication is a normative one that focuses on justification not discovery.

government. The Highway Act, however, did not give any private citizen a right or a duty. No one had a right to an interstate highway near their home or city.

Other legislation establishes *programs*; a program creates individual rights and duties. Some programs are very simple; others extremely complex. The Internal Revenue Code, for example, creates obligations to pay taxes while the Social Security Act grants rights to retirement income of various amounts to individuals with appropriate employment histories.[40] Notice that both the Social Security Act and the Internal Revenue code simultaneously created projects: the administration of the programs established by the statutes.

Third, the legislature may appropriate funds to support either projects or programs it has previously authorized.[41]

Reasoning about projects, programs, and appropriations may differ. The choice of projects seems largely a question of interest or preference aggregation. Program creation, by contrast, may mingle both expressions of interest and judgments of value. Programs, at the very least, must satisfy certain fairness and equality criteria.

Many legal agents make rules, most prominently, legislatures and administrative agencies.

6 The Logic of Administrative Agencies

Legislatures often delegate the formulation of primary rules of conduct to administrative agencies. The legislation establishes a framework for such rule-making, typically announcing one or more performance goals and a set of constraints. The agency then promulgates operational norms that meet the performance goals and satisfy the constraints. The delegation occurs, typically, when appropriate regulation involves many technical issues that render the formulation of effective regulation complex and difficult. The agency develops and relies on technical expertise not available to the legislature.

In rule-making, an agency must complete two tasks. It must determine the consequences that any proposed rule will have. In environmental regulation, for example, in setting a standard for auto emissions, the agency must predict the effect of such a standard on air quality and on the cost of automobiles. Neither prediction is straightforward. For each, micro-economic reasoning may be beneficial.

The agency must then assess the predicted consequences of the proposed rule. This assessment also involves multiple steps. The legislation determines the criteria of assessment but, often, the agency must elaborate the statutory criteria when applying them to the concrete problems it confronts. As in the environmental context, the

[40] Both the IRC and SSA simultaneously authorized projects in the form of the relevant government bureaucracies necessary to enforce such laws. Not every statutory program also creates a bureaucracy to enforce it. The Ku Klux Klan Act of 1871, for example, created several individual rights of nondiscrimination by public and private individuals. No enforcement machinery was created.

[41] The list in the text is not exhaustive. Much modern legislature creates administrative agencies or otherwise delegates the task of program creation.

criteria often include competing objectives among which the agency must strike a balance. Stricter emissions standards may lead to better air quality and hence lower mortality and morbidity rates. Stricter standards, however, are more costly and have consequences for employment and profitability levels.

Economists have developed several tools for policy assessment that offer systematic ways to balance competing objectives. These tools typically rely on the welfarist framework sketched in Sect. 2. Cost-benefit analysis, the most prominent of these tools, plays a central role in legal policymaking in the USA.[42] Section 6.1 outlines the theory underlying cost-benefit analysis. Section 6.2 discusses various critiques. Section 3 briefly discusses two, less ambitious frameworks: quality-adjusted life years (QALYs) and cost-effectiveness.

6.1 The Theory of Cost-Benefit Analysis

6.1.1 The Theory

Cost-benefit analysis is, in theory, a welfarist criterion that seeks to identify the policy reform that maximizes the gain in social well-being. Cost-benefit analysis thus begins with an account of individual well-being and then introduces a method of aggregating (or integrating) the well-being of each individual in society into an index of social well-being.

The theoretical account of individual well-being assumes that each agent can compare her well-being under different policies. The agent, that is, has a ranking of policies that derives from her ranking of consequences or outcomes. Cost-benefit analysis then shows how this ranking can be *represented* by a money index that identifies a dollar value to each policy such that the agent prefers policy P to policy Q if and only if the monetary value $m(P)$ assigned to policy P is larger than the monetary value $m(Q)$ assigned to policy Q.

More specifically, assume that each individual can rank all possible policies[43]; with minimal restrictions on this ranking, one can assume that the ranking is representable by a utility function $u(w, P)$ where w is the agent's wealth and P is a policy.[44]

[42] Some statutes explicitly or by interpretation incorporate a cost-benefit standard though others explicitly reject it. By executive order, agencies must submit proposed rules for a cost-benefit review by the Office of Information and Regulatory Affairs (OIRA) of the Office of Management and Budget.

[43] A purely consequentialist theory would ground the ranking of policies in the agent's underlying ranking of consequences. For purposes of this essay, however, one can take the ranking of policies as primitive. The argument thus encompasses some non-welfarist accounts of social welfare.

[44] Primarily one requires an assumption of continuity. See Debreu (1959). In this minimalist presentation, the agent's preferences are purely ordinal. Given the uncertainty of policy outcomes, a fundamentally consequentialist theory would require that the agent's preferences satisfy the axioms of subjective expected utility theory as in, for example, Savage (1954).

Thus, the agent prefers the wealth-policy pair (w, P) to the pair (w', Q) if and only if $u(w, P) > u(w', Q)$.

Fix some policy P as the status quo. Suppose that the policymaker is considering some policies $Q, R, S, \ldots$ to replace P. The policymaker transforms each agent's representation of her preferences as follows: let $m(Q)$ be defined as follow as that amount of wealth such that the agent would be indifferent between having that much more wealth under the status quo P and the alternative policy Q; i.e., $u(w + m(Q), P) = u(w, Q)$. The agent prefers Q to P, if and only if $m(Q) > 0$. The agent can thus assign a "monetary value" to each potential policy and, as she will prefer policy S to policy Q if and only if $m(S) > m(Q)$, this set of monetary values constitutes an equivalent utility function representing her underlying preferences over policies.[45]

The second stage in cost-benefit analysis aggregates these monetary representations of each agent's preferences. Specifically, the policymaker adds each agent's monetary valuation (from the same baseline policy) to yield a total net social benefit. If this net social benefit is positive, CBA states that society prefers Q to the baseline P; if it is negative, society prefers the baseline (or status quo) P to the proposed policy change Q.[46]

6.1.2 Theoretical Critiques of Cost-Benefit Analysis

Several critiques of the theory merit attention. (I address problems of implementation in Sect. 6.2) Most of these critiques attack the second step that constructs social well-being.

First, however, I consider a critique of the first step that assigns a monetary value to different policies that impose different risks of death on an individual.

Two distinct problems arise. First, the individual's valuation of the risk will depend radically on which policy the analyst chooses as the baseline. To make matters (somewhat more) concrete suppose that if policy P prevails, agent faces a risk p of death while if policy Q prevails the agent faces a risk of death of q with $p > q$. Suppose that the agent's wealth in P is v while her wealth in Q is w and that the agent prefers the pair (w, Q) to the pair (v, P). When the analyst adopts P as the baseline policy from which to construct the new monetized utility, the agent's valuation of policy Q will be no more than v as the agent's wealth in the baseline policy P serves as an upper bound to her valuations of the set of more preferred policies; the agent cannot

[45] In fact, the agent can construct a whole set of new monetary utility functions, each corresponding to a different policy that she takes as numeraire or the status quo. Consider the monetary values $m'(P)$ defined by taking policy Q as the baseline; i.e., we let $u(w, P) = u(w + m'(P), Q)$. This multiplicity of rankings will play an important role in the critique of cost-benefit analysis.

[46] The text attributes to the social planner an objective function that depends solely on the satisfaction of preferences of the citizens. A more general theory starts simply by attributing a set of preferences to the planner.

be willing to pay more wealth than she has to adopt a more preferred policy.[47] By contrast, when the analyst adopts policy Q as the baseline, the move to policy P entails that the agent faces an increased risk of death should policy P be adopted. If $q-p$ is sufficiently large, there may be no amount of wealth that the agent would be willing to accept to change to policy P.

When both the agent's willingness to pay and her willingness to accept are finite,[48] This discrepancy between the agent's willingness to pay and her willingness to accept simply raises a distributional question and a technical problem, both of which I discuss below. When the agent's willingness to accept, however, is unbounded, the discrepancy suggests a problem about the underlying valuation of risks of death, or put more positively, about the valuation of life.

Two issues arise. First, the underlying theory from which the agent's valuation of risks of death is derived assumes that the agent's valuations are linear in the risk of death (see Broome 1985). This assumption, however, is highly implausible; the agent's valuation of a policy will depend non-linearly both on the background risk and the size of the change in risk created by the policy change.

Imagine an agent forced to play Russian roulette. Will her willingness to pay to reduce the number of bullets in the gun barrel from 1 to 0 be identical to her willingness to pay to reduce the number from 6 to 5? Or, phrased differently, will her willingness to pay to reduce the number of bullets from 6 to 0 be six times the amount she would willing to pay to reduce the number form 6 to 5? These equivalences seem implausible.

Non-linearity in the valuation of risk of death has important consequences for the valuation of competing policies. Contrast a preventive policy for some disease to some policy concerning treatment of the same disease. Improving the treatment reduces a realized risk of death for those who have contracted the disease; a preventive policy, by contrast, will reduce the expected number of deaths by reducing the number of people who will contract the disease (while holding the treatment constant). Suppose the preventive and the treatment policies both lead to the same reduction in the expected number of deaths. If the valuations of risk of death are not linear in the risk, these two policies will have different social valuations. Suppose that people are willing to pay more to reduce their risk of death by a given amount when the background risk is high than when it is low. Then the social willingness to pay for the improved treatment will exceed the social willingness to pay for the improved prevention policy. It is not clear, however, that society should prefer the improved treatment to more effective prevention.

[47] By contrast, the agent's valuations of less preferred policies are not bounded below; that is, the amount of wealth she would be willing to accept in compensation for a less preferred policy can be any amount.

[48] The phrasings "willingness to pay" and "willingness to accept" suggest different allocations of an entitlement. When the analyst adopts policy P as the baseline, the agent is "entitled" only to the (higher) risk of death p and she must purchase "on the market" any decreases in risk. Conversely, when the analyst adopts policy Q as the baseline, he implicitly grants the agent an entitlement to a risk q of death. The agent must be compensated for any increases in her risk of death.

Second, an unbounded willingness to accept reflects the irreplaceable nature of life.

Consequently, money does not constitute an adequate measure of the agent's valuation of risks of death. One might say that money and life are not fully commensurable; the use of money as a yardstick for the valuation of risks of death thus distorts the agent's valuations of such risks.[49] This inadequacy of the dollar measures suggests that we need a different way to compare policies that change the risk of death.

The above inadequacy argument will apply to other "irreplaceable commodities." Many issues in health policy confront the problem of irreplaceability. Many disabilities transform an agent's life in ways that make "replacement" difficult and adequate compensation a meaningless concept. A healthy individual, after an auto accident, may become a quadriplegic. The loss of the use of her arms and legs may not be fully compensable and consequently her willingness to pay to reduce the risk of quadriplegia may not be an adequate measure of her valuation of full mobility. These complexities led to the development of cost-effectiveness analysis discussed in Sect. 6.2.

Most critiques of cost-benefit analysis, however, address the second, aggregation step in cost-benefit analysis. Consider, initially, a technical problem. Recall that one may derive many different monetary representations of an agent's preferences. These representations adopt a different policy as the status quo or the baseline from which deviations in the value of alternative policies are measured. These variations representations preserve the *rank order* of the agent's preferences over policies but the representations need not be linear transformations of each other. The ratio of the monetary difference between policies P and P' to the monetary difference between Q and Q' under one representation may not equal the same ratio under a second representation. This lack of cardinality has significant implications for aggregation. Most importantly, it entails that, using policy P as the basis for the representation, society prefers policy Q to policy P but, using policy Q as the basis for the representation of the agent's ranking, society prefers policy P to policy Q. The aggregation of individual cost-benefit assessment thus does not yield a complete social ranking.

Even when we put this difficulty to one side, however, ethical objections to cost-benefit analysis persist. These generally parallel objections to utilitarianism as, in some sense, cost-benefit analysis seeks to implement a utilitarian social welfare function. Cost-benefit analysis, for example, ignores the distributional consequences of a policy change. We cannot conclude from the fact that cost-benefit analysis recommends a move to policy P from policy Q that policy Q is Pareto superior to policy P. In the move to Q, some individuals will have gained and others will have lost.

Recall the effect of wealth on the valuation of policies. As noted above, the agent's wealth places an upper bound on the agent's valuations of policies she prefers to the baseline. This feature of the willingness to pay of an individual when coupled with the additive nature of the aggregation of individual willingness to pay has distributional

[49] Broome (1985) makes this argument.

consequences. Consider to individuals: Bill H who is homeless and has a net wealth of \$10 dollars and Bill G who lives in a 66,000 square foot home and has a net wealth of \$70,000,000,000. Let the status quo and baseline policy be *P* and let society be choosing among policies *P, Q,* and *R*. Suppose Bill H prefers *R* to *Q* to *P* and Bill G prefers *Q* to *R* to *P*. Using *P* as a baseline Bill H assigns a value of \$9.00 to *R* and \$1.00 to *Q* while Bill G assigns the values \$500,000 to *Q* and \$1,000,000 to *R*. Summing implies that society should prefer *Q* to *R,* but it is not clear that this choice is justified. It takes the monetary measures each Bill assigns to policies as strictly comparable even though the yardstick that Bill H uses differs dramatically from the yardstick used by Bill G.

Several other theoretical problems beset cost-benefit analysis. Cost-benefit analysis in common with all other evaluative criteria does not have an adequate way to address problems that arise in the evaluation of policies that affect the population of individuals. Virtually, every policy that alters the risk of death for a variety of individuals will affect the set of people that will, at some time, exist in the society. Suppose, for example, that under policy A, Lisa would, at age 18, have a child Freddy but under policy B Lisa Lisa would, at age 25, have a child Henry. How do we compare policies A and B? Cost-benefit analysis ignores the well-being of both Freddy and Henry as neither exists at the time of policy choice. But it is not clear how cost-benefit would include them if it choose. Suppose we use policy A as a baseline. In what sense does Freddy prefer policy A to policy B? Conversely, in what sense does Henry, who does not exist under policy A, prefer Policy B to policy A? These variable population problems pose significant challenges to most, if not all, ethical theories and cost-benefit analysis does not escape them.[50]

Finally, note that cost-benefit analysis may be understood as an implementation of a potential compensation test. Recall that the Pareto criterion establishes a partial social ranking of policies. Society prefers policy *P* to policy *Q* if no one in society prefers *Q* to *P* and at leat one person prefers *P* to *Q*. If some people are better off in *P* than in *Q* but others are better off in *Q* than in *P*, *P* and *Q* are Pareto non-comparable. A potential compensation test extends the Pareto criterion to all (or some) of these Pareto non-comparable policy pairs. In essence, a potential compensation says that society should prefer *P* to *Q* if those who prefer *P* to *Q* could compensate those prefer *Q* to *P*. If such compensation occurred—moving society to policy P', then a Pareto improvement would have occurred—i.e., no one prefers *Q* to P' and at least one person prefers P' to *Q*.[51] But further argument is required to support the conclusion that *P* should be socially preferred to *Q*. Saying that P' is Pareto superior to *Q* does not justify saying that *P* is socially preferred to *Q*.

[50] For a general discussion see Parfit (1984) and Broome (2004).

[51] Notice that there may be very many different ways in which such compensation could be accomplished.

Since 1981, an executive order[52] has required presidential agencies[53] to conduct a cost-benefit analysis in rulemaking proceedings. This requirement arose to rationalize decision making across agencies.

In the previous section, I outlined a few theoretical difficulties that the cost-benefit analyst faces. These executive orders highlight a number of difficulties in implementation that bear on reasoning in rule-making. These problems arise because the cost-benefit analyst does not directly observe individual preferences understood as an individual's ranking of policies on the basis of well-being.

Broadly speaking, cost-benefit analysis relies on either of two methods of elicitation of these valuations: hedonic pricing or contingent valuation. Hedonic pricing infers valuations from market behavior while contingent valuation elicits these valuations through survey instruments. Contingent valuation, by contrast, elicits valuations through survey questions or laboratory experiments on willingness to pay. Each procedure elicits valuations that are problematic.

Consider hedonic pricing first. Governmental policies generally address concerns that are only indirectly priced in a market. Consider, for example, a statute or administrative regulation that sets standards for water or air quality or establishes a national park.[54] This rule will affect the morbidity and mortality rates of many individuals. A standard requiring the removal of lead-based paint from residential structures, for example, improves health for both children and adults as exposure to lead paint leads to slowed growth, hearing problems, and brain and neuronal damage in children, to miscarriages and premature births in pregnant women, and to sensory impairment, reduced mental function, and various physical problems in adults (See http://www.epa.gov/lead/pubs/learn-about-lead.html#effects). How should we measure the value that arises from the reduction in these risks? Hedonic pricing infers the value from an examination of market behavior.

The labor market often provides an indirect measurement of some individual valuations.[55] The extent of exposure to lead varies across occupations. Compare two occupations that are identical except for the extent of lead exposure. We would expect that the occupation that imposes greater risks would have to pay a higher wage to

[52] Reagan issued the first regulation, Executive Order 12291. Subsequent administrations have continued the policy with some amendments. The executive orders, of course, do not override statutory mandates to ignore cost-benefit analysis. The Delaney amendments to the Food and Drug Act for example require the FDA not to approve any drug that creates a risk of cancer.

[53] The head of a presidential agencies serves at the pleasure of the President of the USA. Executive agencies contrast with independent agencies. Independent agencies are governed by a committee appointed in various ways. The Federal Reserve Board, the National Labor Relations Board, the FCC, and the FTC among others are independent agencies.

[54] Similar issues may arise in adjudication after an oil spill or other event that contaminates the air or water. The grounding of the *Exxon Valdez*, for example, released between 250,000 and 750,000 barrels of oil into Prince William Sound, a previously pristine body of water.

[55] The housing market might also yield a measure of at least some risks. Housing prices should differ by location. A house near a toxic waste site should cost less than one farther away. This price differential should then permit the analyst to estimate the value of the decreased risk of exposure to the toxic waste in the more distant house location. The problems discussed in the text with the labor market measure apply to the housing market measures as well.

compensate its employees for facing the greater health risks the exposure imposes. This wage differential reflects the marginal worker's valuation of the risks imposed by lead exposure.[56]

It is problematic simply to adopt this wage differential as the unit social measure of the risk. In our hypothetical, the pool of workers will sort themselves between the two occupations. Those individuals who fear the risk most will find the wage differential inadequate to compensate for them the risk; they will choose the low exposure occupation. Those with low valuations will choose the high exposure occupation. The high exposure occupation may be able to meet its needs with very few workers in which case the wage differential will be small.

In any case, the worker who chooses the high exposure occupation has done so voluntarily and has a lower valuation of the exposure than those who voluntarily chose the low exposure occupation. Most environmental standards and many health and safety standards, by contrast, govern involuntary exposures. The relevance of the valuation of the marginal, voluntary individual to the social valuation is thus unclear. The formal theory of cost-benefit analysis directs us to sum the valuations of each individual. The valuation of the marginal worker is a good proxy for that sum only if it equals the mean value of the distribution of valuations.

The marginal worker, of course, will only rarely have the mean valuation of the risk.

Indeed, the marginal worker is unlikely to have the mean valuation in the population of workers in the two occupations let alone the general population. The hedonic measure is apt to be biased relative to the general population because the marginal worker in most high exposure occupations is not apt to have the average income or wealth in the population. Formal theory predicts that an agent's valuation of a policy will depend on the agent's wealth. Wealthier individuals will place a higher valuation on the avoidance of most risks than less wealthy individuals. Wealthy individuals demand greater environmental quality because they can afford it; consequently they live in neighborhoods with clearer air and cleaner water that are located far from hazardous waste sites and other disamenities.[57]

The measurement technique used by hedonic pricing, moreover, conflates two judgments made of the worker: her judgment about the size of the differential risk across the two occupations and her judgment about the valuation of that difference. A worker might opt into the high exposure occupation because the value of reduced exposure is low or because she believes that the difference in risks to health occasioned by the differential exposures is small. Ideally, the analyst would derive

[56] Occupations generally differ on more than one dimension so that the analyst must tease out the effect of the difference in exposure to a given hazard from the effect of other occupational differences on occupational choice. The high exposure occupation for example may also offer a better parental leave policy or an otherwise more pleasant work environment. More problematically occupation O may have a lower exposure to risk R but a higher exposure to risk S than occupation P. To determine the appropriate policy with respect to risks R and S, the analyst must unbundle the worker's valuation of each risk.

[57] Even if disamenities are initially sited randomly, over time wealthier individuals will relocate to cleaner sites. See Been (1993).

distinct estimates of the worker's beliefs and of her valuation of the differential risk. The desire to control for the conflation of the two estimates in part explains the use of contingent valuation techniques.

Experimental evidence shows that individuals generally misestimate very small risks. In effect, individuals cannot distinguish among very small risks. In many instances the policymaker must choose between policies that reduce risks either to 0.0001 or to 0.00001, policies that are thus behaviorally indistinguishable.

Incorrect assessment of risks by citizens complicates the policymaker's task. In one sense, the best policy should depend on the true risks that the policy generates. If the policymaker has better information than the citizens, then she should base her cost-benefit analysis on her more accurate risk assessments and the individual valuations of the given risk. On the other hand, most policies are not self-implementing. Individuals respond to these policies on the basis of their own, inaccurate beliefs. Two responses are relevant: a policy response and a political response. Individuals will adjust their behavior to the policy.

It is known, for example, that individuals overestimate the risks of solid hazardous waste relative to the risks of air and water pollution. "Ideal" policy would thus dictate relatively more lenient regulation of hazardous waste sites, regulation that presumably would permit greater concentrations of hazardous chemicals in the soil. Suppose the policymaker enacts this ideal policy. Individuals will have both a policy response and a political response. The political response will demand more stringent regulation of hazardous waste (and correspondingly less regulation of air and water quality). The policy response will consist of actions that undermine the policy objectives. Redevelopment of brown sites, for example, might be less than the policymaker anticipated or believes desirable because the risks associated with the redevelopment will be overestimated by developers (or by the final users of those redeveloped sites).

Turn now to contingent valuation as an elicitation technique. Contingent evaluation emerged roughly at the time of the grounding of the *Exxon Valdez* in Prince William Sound.

The oil spill led to numerous law suits and the need to value the various environmental harms the spill caused. These harms ranged from commercial losses that were relatively easy to value to subsistence losses[58] and to more esoteric "environmental" harms. In particular, analysts attempted to measure the "use value" and "existence value" of an ecologically pristine Prince William Sound.

The environmental use value refers to the valuation of the pristine resource (relative to the contaminated resource) that visitors to Prince William Sound would realize. The environmental existence value refers to the valuation of the pristine resource that accrues simply because Prince William Sound exists in a pristine state; even an individual who does not visit Prince William Sound or who does not visit or ever intend to visit might attribute an existence value to it.

This example raises two sets of problems: one methodological and one substantive.

[58] Subsistence losses refers to the losses to indigenous communities of their access to wildlife that formed the basis of their diet and way of life. As these communities operated largely outside the market economy, measures used to estimate commercial losses were deemed inappropriate.

Consider the methodological issue first. Individual valuations are sensitive to the method of elicitation.[59] In *The Exxon Valdez* example, for instance, very different valuations are elicited when subjects are simply asked a straightforward question that requests a dollar value for the existence value than when subjects are asked what increase in the price of gasoline they would accept. The valuation derived from the change in the price of gasoline is significantly less than the other valuation.

Substantively, one might question the relevance of existence value to a welfarist criterion. It is not clear how the existence of a pristine ecological system (or any ecosystem) distant from the agent contributes to her well-being. The agent, of course, might have moral commitments to environmental protection and work toward promoting environmental protection might further her well-being.

Further issues arise in the implementation of the formal theory of cost-benefit analysis. A significant debate has arisen over the appropriate way to value life. The debate has arisen because various administrative rules issued by the federal government of the USA assign widely divergent implicit valuations of life.[60] This debate, however, misconstrues the object of valuation in a cost-benefit analysis. Rulemakers choose among policies, and it is the consequences of these policies that must be valued. These policies generally alter the risk of death. The valuation of the risk of death, however, does not yield a unique valuation of a life; the valuation depends on the baseline risk and the change in the risk induced by the policy.

The view that there is a fixed value of life has two sources. First, one might think that, in equilibrium, a rational government would allocate money across risks so that the value of life was equated. But cost-benefit analysis assumes that the policymaker is maximizing social welfare not the number of lives saved. Thus, when allocating a pool of money across policies that alleviate different risks, she will allocate it to equate net benefits across policies. A policy, however, may have benefits that are not measured in lives (or deaths) but in quality of lives or differential resources expended. We can conclude that a social welfare maximizing policymaker would not seek to implement policies that valued the marginal life saved equally.

Second, the view that individuals assign a fixed value to life rests on a derivation of the value of a life from preferences that can be represented as a function of money alone (see Broome 1985).

Such preferences do a poor job of capturing actual attitudes toward death. To do this, we require state-dependent preferences that are represented by a state-dependent utility function $u(w, h)$ where w is the agent's wealth and h her state of health. One possible state of health, of course, is death. When dead, the agent presumably does

[59] This problem reflects a more general problem of whether individual preferences vary with the method of elicitation. The problem initially manifested in experiments concerning choice under risk in which subjects rankings of two lotteries depended on whether they chose directly between the two lotteries or between their certainty equivalents. This preference reversal phenomenon has been extensively studied in the laboratory. For a survey, see Hsee et al. (1999).

[60] See, e.g., Zweifel and Breyer (1997) analysts calculate the implicit value of life from a regulation by dividing the valuation of the reduction in risk by the level of risk. So if an agent values a reduction in risk of death of 0.01 at $100,000, the implicit value of life is $10,000,000.

not value wealth at all.[61] She values wealth at some constant low rate. When fully healthy, the live agent values her wealth more highly than she, when dead, would value that amount of wealth in her estate. The agent prefers for any level w of wealth, the paire (w, *fully healthy*) to the pair (w, *dead*).

The agent has similar preference when considereing many other disabilities. Quadriplegics prefer more wealth to less just as paraplegics do but one might argue that, for every w, the agent prefers (w, *paraplegic*) to (w, *quadriplegic*) or (w, *sighted*) to (w, *blind*).

Consider a status quo defined by the risk r of death faced by an individual. Her state-dependent utility is then simply $(1-r)u(w, \mathit{healthy}) + ru(w, \mathit{dead})$. Consider a policy P that reduces the agent's risk of death by p. The formal theory of cost-benefit analysis allows us an equivalent representation of the agent's preferences that are indexed by money. Specifically, we can calculate the amount of money that the agent would be willing to pay to adopt policy P:

$(1-r)u(w, \mathit{healthy}) + ru(w, \mathit{dead}) = (1-r+p)u(w-m, \mathit{healthy}) + (r-p)u(w-m, \mathit{dead})$

When $m > 0$, the agent prefers the new policy to the status quo. Notice that m will depend on w, r, and p.

One might argue in response that policies affect a large population. A policy that reduces a risk of death by 0.00001 will very likely reduce the number of deaths in the USA by 3000 (assuming a population of 300,000,000). When comparing two policies, we should simply compare the number of statistical lives saved. This objection simply restates the initial objection. This comparison might be dictated if the social objective is to maximize the number of lives saved, but it does not follow when the social objective is different.

Finally, one should note that, typically, we ask how a policymaker should reason about the content of a particular policy. The argument has suggested, however, that a fully rational policymaker cannot reason policy-by-policy. She must reason instead in terms of comprehensive legislative programs.

6.2 Cost-Effectiveness

Section 6.1.2 noted that cost-benefit analysis assumes that any feature of the world that policy affects can be given a monetary evaluation. Many commodities, however, are irreplaceable, and it is not clear that irreplaceable commodities such as life or certain health statuses can be adequately evaluated with money. Cost-effectiveness analysis thus compares policies of equal monetary cost on the basis of their effects on the irreplaceable goods.

Health economists have responded to this challenge by creating a different index to measure policy effects on these irreplaceable commodities. The usual index is

[61] Prior to death, the agent may have a bequest motive but it is odd to think that a dead agent has any preferences at all.

the *QALY*, the quality-adjusted life year. Conceptually, the QALY is quite simple; people generally prefer both more years of life to fewer and, for a given life span, prefer a higher quality of functioning adjusted life year. Conceptually, the QALY is quite simple; people generally prefer both more years of life to fewer and, for a given life span, prefer a higher quality of functioning to a lower quality of functioning. Individuals typically would prefer better cognitive and physical functions to poorer ones; an agent prefers full use of arms and legs to quadriplegia or paraplegia and full mental capacity to impaired mental capacity (but see Voltaire 1977).

These methods avoid some of the theoretical difficulties faced by cost-benefit analysis but they nonetheless rest on very strong assumptions. For instance, QALYs are treated equally; lengthening the life of widowed centenarian by 10 years counts for the same as lengthening the life of a 25-year old single mother of an infant. The problems of implementation are equally if not more demanding.

7 The Logic of Adjudication

Courts resolve disputes.[62] In the course of this dispute resolution, they may also promulgate rules that will govern the behavior of other agents. Dispute resolution, unlike rule-making, is backward-looking. It *applies* extant rules rather than justify new ones.

Recall Rawls' discussion in "Two Concepts of Rules." In Rawls (1958), he argued that both retributivist and utilitarian accounts of punishment might both be valid with the former informing the practice of judicial enforcement of punishments and the latter informing the legislative enactment of punishment. A similar argument might reconcile corrective justice and deterrence accounts of tort law. This dichotomy, however, offers a more stark distinction between adjudication and legislation than actual institutions warrant. A more detailed attention to common law adjudication suggests the additional complexity.

Common law courts, in resolving a case, have several tasks. First, they must determine which legal rule or rules apply to the dispute. Second, they must determine the facts underlying the dispute. Finally, they must apply the chosen rule to the found facts.[63] Identification of the governing rule and its application, however, are often not straightforward. Generally, the court must *interpret* one or more rules because

[62] As do some administrative agencies. In the USA, for example, the National Labor Relations Board rarely engages in rule-making; it operates solely through the adjudication of labor-management disputes.

[63] Civilian courts face identical questions though the first—the identification of the applicable legal rule or rules—is somewhat obscured by the structure of judicial opinions in civil law countries, or, at least, in France. In France, opinions of the Cour de Cassation do not explain or justify the choice of the legal rule or rules that the Cour has determined govern the dispute. The logic of this choice is suppressed. In common law jurisdictions, the process of choice is more apparent as the opinion will generally refer to the various rules or precedents that might apply and then offer reasons for its choice of governing rule.

either no rule explicitly governs the dispute or because multiple rules indicating contradictory dispositions govern. Interpretation thus requires the court not simply to apply a law but also to create one.

The logic of judicial law creation is much disputed. It turns in part on one's view of the role of courts in a political system; how judges reason when making law thus depends on the constitutional design of adjudicatory institutions. The approach to judicial interpretation will vary as the understanding of adjudication varies. If adjudication has been structured as rule-instrumental,[64] then, in interpreting statutes, constitutions, or prior case law, judges should act as legislators. Judges, that is, should interpret the rule to promote the aims of the legislation. If adjudication has been structured institutionally instrumentally or systemically instrumentally, then judicial interpretation may be much less consequentialist in character.

Consider, for example, Rawls' example of punishment. Suppose the dispute for the court raises a novel issue, not clearly governed by the underlying statute. A rule-instrumental account would adopt the interpretation that best promoted the deterrence (or other *ex ante*) aims of the legislation. Constitutional design, however, might have created an institutionally instrumental court that should adopt the retributory structure as its interpretive guide.

8 Concluding Remarks

Economic analysis of law began with claims that an economic logic underlay the logic of the law. I have suggested that these claims can be interpreted either as claims about the nature of the social processes that create and apply law or as claims about the processes of legal reasoning that different agents deploy.

Economic logic as the logic of social processes deepens our understanding of how law develops. In its explanatory form, this claim is best understood as a claim that we can best understand legal phenomena by thinking like an economist. Since Adam Smith, economists have developed the insight that, though social processes are the result of the intentional actions of individuals, no person necessarily intended the result of these processes. Posner's initial claim for the efficiency of the common law made a classical Smithian claim: that the unintended result of the actions of individual judges led to a socially desirable outcome. Since Smith, of course, economists have discovered structures of decision in which individually rational decisions lead to socially undesirable outcomes.

The case for economic logic as the logic of legal reasoning is more complex for two reasons. First, much legal reasoning is normative; it seeks to identify the goals we should pursue. Economic reasoning, by contrast, is largely instrumental; it takes our aims as given.

[64] See Kornhauser (Chap. 5, part II, this volume, on "Choosing Ends and Choosing Means: Teleological Reasoning in Law"), for an account of the various forms of instrumentalism.

Thus, at best, economic reasoning can only underlie the instrumental aspect of legal reasoning, the part that determines how we achieve the goals we have.

Second, the nature of reasoning about law varies with context. Legislators, administrative agencies, and courts face distinct tasks. Economic reasoning may be more appropriate for legislators than for judges. Indeed, economic logic as social process may suggest that we might best achieve our political aims when judges reason non-economically. Similarly, whether administrators should follow the canons of economic reasoning may depend on the nature of the delegation that the legislature has made.

References

Anderson, E. 1995. *Value in ethics and economics*. Cambridge, Mass.: Harvard University Press.

Arrow, K.J. 1963. *Social choice and individual values*. New Haven, Conn.: Yale University Press.

Bebchuk, L. 1980. *The pursuit of a bigger pie: Can everyone expect a bigger slice? Hofstra Law Review* 8: 671–709.

Been, V. 1993. What's fairness got to do with it? environmental justice and the siting of locally undesirable land uses. *Cornell Law Review* 78: 1001–1085.

Benoît, J.-P., and L.A. Kornhauser. 2002. Game-theoretic analysis of legal rules and institutions. In *handbook of game theory with economic applications*, vol. 3, ed. R. Aumann and S. Hart, 2229–2269. Cambridge, Mass.: Elsevier.

Broome, J. 1985. The economic value of a life. *Economica* 52: 281–294.

Broome, J. 1990. Bolker Jeffrey expected utility theory and axiomatic utilitarianism. *Review of Economic Studies* 57: 477–502.

Broome, J. 1991. *Weighing goods*. Oxford: Blackwell.

Broome, J. 2004. *Weighing lives*. Oxford: Oxford University Press.

Buchanan, J.M., and G. Tullock. 1962. *The calculus of consent: Logical foundations of constitutional democracy.* Ann Arbor, Mich: University of Michigan Press.

Calabresi, G. 1961. Some thoughts on the logic of risk distribution and the law of torts. *The Yale Law Journal* 4: 499–553.

Calabresi, G., and D. Melamed. 1972. Property rules, liability rules, and inalienability: One view of the cathedral. *Harvard Law Review* 6: 1089–1128.

Coase, R. 1959. The federal communications commission. *Journal of Law and Economics* 2: 1–40.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

Cooter, R., and L.A. Kornhauser. 1980. Can litigation improve the law without the help of judges? *Journal of Legal Studies* 9: 139–164.

Cooter, R., D. Lane, and L.A. Kornhauser. 1979. Liability rules, limited information, and the role of precedent. *Bell Journal of Economics and Management Science* 10: 366–373.

D'Aspremont, C., and L. Gevers. 1977. Equity and the informational basis of collective choice. *Review of Economic Studies* 33: 199–209.

Debreu, G. 1959. *Theory of value: An axiomatic analysis of economic equilibrium*. New York: Wiley.

Dietrich, F., and C. List. 2004. A model of jury decisions where all jurors have the same evidence. *Synthese* 142: 175–202.

Dworkin, R.M. 1980. Is wealth a value? *The Journal of Legal Studies* 9: 191–226.

Eisenberg, M.A. 1988. *The nature of the common law*. Cambridge, Mass.: Harvard University Press.

Fernandez, P.A., and G. Ponzetto. 2008. Case law versus statute law: An evolutionary comparison. *Journal of Legal Studies* 2: 379–430.

Fleurbaey, M. 2005. Health, wealth and fairness. *Journal of Public Economic Theory* 7: 253–284.

Galanter, M. 1974. Why the haves come out ahead: Speculations on the limits of legal change. *Law and Society Review* 1: 95–160.

Gennaioli, N., and A. Shleifer. 2007. The evolution of common law. *Journal of Political Economy* 115: 43–68.

Goodman, J. 1978. An economic theory of the evolution of the common law. *Journal of Legal Studies* 7: 393–406.

Hadfield, G. 1992. Bias in the evolution of legal rules. *Georgetown Law Journal* 80: 583–616.

Harsanyi, J. 1955. Cardinal welfare, individualist ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63: 309–321.

Hsee, C.K., G.F. Loewenstein, S. Blount, and M.H. Bazerman. 1999. Preference-reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin* 5: 576–591.

Kaplow, L., and S. Shavell. 2002. *Fairness versus welfare*. Cambridge, Mass.: Harvard University Press.

Kornhauser, L.A. 1980. A Guide to the perplexed claims of efficiency in the law. *Hofstra Law Review* 3: 591–639.

Kornhauser, L.A. 1996. Notes on the logic of legal change. In *Social rules: Origin, character, logic, change,* ed. D. Braybrooke. Boulder, Col.: Westview Press.

Kornhauser, L.A. 2000. These roles for a theory of behaviour in a theory of law. Rechtstheorie 31: 197–252.

Kornhauser, L.A. 2008. Law and economics. In *The encyclopedia of the supreme court of the United States*, vol. 3, ed. D.S. Tanenhaus, Andover, N.H.: Cengage Gale Publishing.

Kornhauser, L.A. 2010. *L'Analyse Economique du Droit*. Paris: Michel Houdiard Editeur.

Kornhauser, L.A. 2013. Deciding Together. *New York University Law and Economics Working Papers* 358: 1–19. http://lsr.nellco.org/nyu_lewp/358.

Kornhauser, L.A. 2017. The Economic Analysis of Law. *The Stanford Encyclopedia of Philosophy.* ed. E. Zalta. https://plato.stanford.edu/entries/legal-econanalysis/#ParCri.

Lasser, M. 2005. *Judicial deliberations: A comparative analysis of judicial transparency and legitimacy*. Oxford: Oxford University Press.

Levi, E. 1948. An introduction to legal reasoning. *The University of Chicago Law Review* 3: 501–574.

Llewellyn, K.N. 1951. *The bramble bush. On our Law and its study.* New York: Oceana Publications. (1st ed. 1930).

Myerson, R.B., and M.A. Satterthwaite. 1983. Efficient mechanisms for bilateral trading. *Journal of Economic Theory* 29: 265–281.

Nussbaum, M. 2000. The costs of tragedy: Some moral limits of cost-benefit analysis. *Journal of Legal Studies* 29: 1005–1036.

Parfit, D. 1984. *Reasons and persons.* Oxford: Clarendon.

Posner, R.A. 1973. *Economic analysis of law.* New York: Little, Brown and Co.

Posner, R.A. 1979. Utilitarianism, economics, and legal theory. *The Journal of Legal Studies* 8: 103–140.

Posner, R.A. 1980. The ethical and political basis of the efficiency norm in common law adjudication. *Hofstra Law Review* 3: 487–507.

Priest, G. 1977. The common law process and the selection of efficient rules. *The Journal of Legal Studies* 1977 (6): 65–82.

Rawls, J. 1955. Two concepts of rules. *The Philosophical Review* 1: 3–32.

Rawls, J. 1958. Justice as fairness. *The Philosophical Review* 2: 164–194.

Richardson, H. 2000. The stupidity of the cost-benefit standard. *The Journal of Legal Studies* 29: 971–1003.

Rubin, P. 1977. Why is the common law efficient? *The Journal of Legal Studies* 6: 51–63.

Satz, D. 2010. *Why some things should not be for sale: The moral limits of markets*. Oxford: Oxford University Press.

Savage, L.J. 1954. *The foundations of statistics*. New York: Wiley.

Schauer, F. 2009. *Thinking like a lawyer. A new introduction to legal reasoning*. Cambridge, Mass.: Harvard University Press.

Sen, A.K. 1970. The impossibility of a paretian liberal. *Journal of Political Economy* 1: 152–157.

Stigler, G., and G.J. Stigler. 1966. *The theory of price*. New York: Macmillan.

Sullivan, J.V. 2007. *How our laws are made*. Washington D.C: U.S. Government Printing Office. http://thomas.loc.gov/home/lawsmade.bysec/formsofaction.html.

Temkin, L. 1993. Harmful goods, harmless bads. In *Value, welfare and morality*, ed. R.G. Frey, and C.W. Morris, 290–324. Cambridge: Cambridge University Press.

Voltaire (François-Marie Arouet). 1977. *The good brahmin*. New York: Viking Penguin. (1st ed. 1761).

Zweifel, P., and F. Breyer. 1997. *Health economics*. Oxford: Oxford University Press.

Index of Names

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0

Index of Subjects

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0