Library of Ethics and Applied Philosophy 27

Nicole A Vincent
Ibo van de Poel
Jeroen van den Hoven *Editors*

Moral Responsibility

Beyond Free Will and Determinism

Springer

MORAL RESPONSIBILITY

LIBRARY OF ETHICS AND APPLIED PHILOSOPHY

VOLUME 27

For further volumes:
http://www.springer.com/series/6230

MORAL RESPONSIBILITY

Beyond Free Will and Determinism

Edited by

NICOLE A VINCENT
Delft University of Technology, Delft, The Netherlands
Macquarie University, Sydney, Australia

IBO VAN DE POEL
Delft University of Technology, Delft, The Netherlands

and

JEROEN VAN DEN HOVEN
Delft University of Technology, Delft, The Netherlands

Springer

Editors
Nicole A Vincent
Delft University of Technology
Department of Philosophy
Faculty of Technology,
Policy & Management
Jaffalaan 5
2628 BX Delft
The Netherlands
n.a.vincent@tudelft.nl

Department of Philosophy
Faculty of Arts
Macquarie University
Sydney NSW 2109
Australia
nicole.vincent@mq.edu.au

Prof. Dr. Jeroen van den Hoven
Delft University of Technology
Department of Philosophy
Faculty of Technology,
Policy & Management
Jaffalaan 5
2628 BX Delft
The Netherlands
m.j.vandenhoven@tudelft.nl

Ibo van de Poel
Delft University of Technology
Department of Philosophy
Faculty of Technology,
Policy & Management
Jaffalaan 5
2628 BX Delft
The Netherlands
I.R.vandePoel@tudelft.nl

ISSN 1387-6678
ISBN 978-94-007-1877-7 e-ISBN 978-94-007-1878-4
DOI 10.1007/978-94-007-1878-4
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2011934879

Contents

Contributors

Diana Abad Department of Philosophy, University of Potsdam, Potsdam, Germany, diana.abad@uni-potsdam.de

Jason Benchimol Department of Philosophy, University of Washington, Seattle, WA, USA, jdbench@uw.edu

Gunnar Björnsson Linköping University, Linköping, Sweden; University of Gothenburg, Gothenburg, Sweden, gunnar.bjornsson@liu.se

Lene Bomann-Larsen Department of Philosophy, Classics, History of Art and Ideas, University of Oslo, Oslo, Norway, lene.bomann-larsen@hf.uio.no

Alexander Brown School of Political, Social and International Studies, University of East Anglia, Norwich, UK, alexander.c.brown@uea.ac.uk

Susan Dimock York University, Toronto, Canada, ON, dimock@yorku.ca

Rosemary Lowry Department of Philosophy, Eindhoven University of Technology, Eindhoven, The Netherlands, r.j.lowry@tue.nl

Seumas Miller Centre for Applied Philosophy and Public Ethics, Australian National University, Canberra, ACT, Australia; Charles Sturt University, Canberra, NSW, Australia; Department of Philosophy, Delft University of Technology, Delft, The Netherlands, seumas.miller@anu.edu.au

Ibo van de Poel Department of Philosophy, Delft University of Technology, 2600 GA Delft, The Netherlands, I.R.vandePoel@tudelft.nl

Michael Smith Department of Philosophy, 1879 Hall, Princeton University, Princeton, NJ 08544, USA, msmith@princeton.edu

Steve Vanderheiden Department of Political Science, University of Colorado, Boulder, CO, USA; Centre for Applied Philosophy and Public Ethics (CAPPE), Canberra, ACT, Australia, vanders@colorado.edu

Nicole A Vincent Department of Philosophy, Delft University of Technology, 2600 GA Delft, The Netherlands; Department of Philosophy, Faculty of Arts, Macquarie University, Sydney, NSW 2109, Australia, n.a.vincent@tudelft.nl; nicole.vincent@mq.edu.au

Chapter 1
Introduction

Nicole A Vincent and Ibo van de Poel

1.1 Beyond Free Will and Determinism

It is now well over a decade since John Fischer and Mark Ravizza (1998) – and before them, Jay Wallace (1994) and Daniel Dennett (1984) – defended responsibility from the threat of determinism. What these authors' compatibilist theories have in common is the idea that responsible agents are not those agents whose actions are un-caused, but rather those agents who possess certain competences or capacities. But defending responsibility from determinism is a potentially endless and largely negative enterprise – it can go on for as long as dissenting voices remain (i.e. indefinitely), and although such work strengthens the theoretical foundations of these theories, it won't necessarily build anything on top of those foundations, nor will it move these theories into new territory or explain how to apply them to practical contexts. To this end, instead of devoting more effort to the negative enterprise of building up even stronger fortifications against the ever-present threat of determinism, the papers in this volume address these more positive challenges by exploring ways in which compatibilist responsibility theory can be extended and/or applied in a range of practical contexts.

This book begins with two chapters that set out a finer-grained understanding of the rich and multi-faceted notion of responsibility found in contemporary debates, situated within a broadly compatibilist framework.

In philosophical discussions, responsibility is often talked about as if it were a single, unitary and generic concept, or at least that is the impression that such discussions can create. For instance, in the compatibilist literature, the topic seems to be whether determinism rules out *responsibility*. In political philosophy, luck egalitarians and their opponents debate whether people's entitlements should track their *responsibility*. And in the area of neurolaw, some people claim that "a truly scientific mechanistic view of the nervous system make[s] nonsense of the very idea

N.A Vincent (✉)
Department of Philosophy, Delft University of Technology, 2600 GA Delft, The Netherlands

Department of Philosophy, Faculty of Arts, Macquarie University, Sydney NSW 2109, Australia
e-mail: n.a.vincent@tudelft.nl; nicole.vincent@mq.edu.au

N.A Vincent et al. (eds.), *Moral Responsibility*, Library of Ethics and Applied Philosophy 27, DOI 10.1007/978-94-007-1878-4_1,

of *responsibility*" (Dawkins 2006, emphasis added; also Greene and Cohen 2004), while others deny that neuroscience challenges *responsibility* (e.g. Gazzaniga 2006; Morse 2006). If confronted head on, most philosophers would probably acknowledge as Fischer and Ravizza do that the term "'responsibility' admits of a variety of uses" (1998:2, note 1), but somehow this variety seldom comes across in anything other than footnotes and its implications are seldom drawn out.

To this end, **Nicole Vincent**'s paper "A Structured Taxonomy of Responsibility Concepts" aims to remedy this situation by saying something more substantial about this variety of uses of the term "responsibility", about how these uses relate to one another, and by explaining how we are better off for knowing about these things. First, Vincent distinguishes six different responsibility concepts which, drawing inspiration from H. L. A. Hart's terminology, she calls capacity responsibility, virtue responsibility, causal responsibility, outcome responsibility, role responsibility and liability responsibility. Sometimes "responsible" describes a kind of person, either in terms of their *capacities* ("a fully responsible person"), or in terms of their *character* virtues or lack thereof ("an irresponsible person"). At other times it describes relations between events, either in *causal* terms ("his depression was responsible for his behaviour"), or in *moral* terms ("he is responsible for that accident"). And finally, it can also refer to a person's *duties* ("these are your responsibilities"), or to how they should be *treated* ("you will be held responsible for that accident"). Second, Vincent argues that these six responsibility concepts relate to one another via justificatory relations that obtain between claims which employ them, and she explains the nature of these justificatory relations. For instance, she points out that people are normally expected to take *liability* responsibility precisely because of and usually only for those things for which they are *outcome* responsible; it is precisely because a person had certain *role* responsibilities which they subsequently violated that we tend to attribute *outcome* responsibility to them but not to others who did not have those role responsibilities; and we normally impose fewer and less weighty *role* responsibilities onto people whose *capacity* responsibility was reduced. Finally, she also argues that not only can an awareness of this rich complexity help us to systematically resolve a range of problems in scholarly as well as public debates, but that it also suggests a new set of problems that compatibilists should address. Namely, she suggests that since "responsibility" refers to a range of different ideas, compatibilists must explain in what way determinism is meant to pose a threat to each of these different ideas, and then spell out how compatibilism addresses these different challenges.

On Vincent's account, claims about what states of affairs can be legitimately attributed to a given person as things of their doing and for which they can be blamed (in her terminology, things for which they are outcome responsible) hinge crucially on whether those people acted in a way that transgressed how they ought to have acted (in her terminology, whether they breached their role responsibilities). This feature of Vincent's analysis suggests that there exists a relationship between *backward*-looking responsibility claims (claims about who is outcome responsible for the things that happened in the past) and *forward*-looking responsibility claims (claims about who had- and subsequently violated what role

responsibilities), and the explicit aim of **Ibo van de Poel**'s paper "The Relation Between Forward-Looking and Backward-Looking Responsibility" is to shed further light on this relationship. In this regard, van de Poel distinguishes two varieties of backward-looking responsibility: accountability and blameworthiness. He argues that accountability implies having to account for one's actions, but if the account given is insufficient then one will be blameworthy. Furthermore, he also argues that although accountability is sometimes based on an improperly discharged forward-looking responsibility, it may also be based on the breach of a moral duty. This is important because although the duty to account for one's actions presupposes at least three conditions – a capacity condition, a causality condition and a wrongdoing condition – van de Poel argues that the exact formulation of these conditions is different for these two routes to accountability.

However, van de Poel's paper does a lot more than just shed light on the nature of the relationship between backward and forward-looking responsibility concepts. He begins by distinguishing nine different responsibility concepts from one another – namely, responsibility as cause, as task, as authority, as capacity, as virtue, as obligation, as accountability, as blameworthiness and as liability. Although a number of these concepts map straight-forwardly onto the six concepts distinguished by Vincent, some important differences and discrepancies remain. For instance, van de Poel's concept of responsibility as *authority* has no counterpart in Vincent's analysis, but yet we often say that something is someone's responsibility, meaning that they are in charge of it (in Hohfeld's terms, that they have some kind of *privilege/liberty* or perhaps a *power* with respect to it) and not just that it is their burden to carry (in Hohfeld's terms, that it is their *duty* or *liability*) (Hohfeld 1975). Secondly, van de Poel's responsibility as *accountability* also finds no direct counterpart in Vincent's taxonomy, but yet, as he argues, it is only when a person fails to provide a satisfactory account for what has happened that we would even start thinking of blaming them if what happened was untoward. Finally, although both cite wrongdoing as a condition of (outcome- or blame-) responsibility for untoward states of affairs, while van de Poel distinguishes two different ways in which this condition can be satisfied – one via a consequentialist route that is traversed when a forward-looking responsibility is not (properly) discharged, and another via a deontic route that occurs in the face of a duty transgression – Vincent's analysis does not distinguish between these two routes since on her account this condition is satisfied simply when an agent breaches their role responsibility. Thus, in addition to offering an in-depth discussion of the relationship between backward and forward-looking responsibility concepts, van de Poel's analysis also further enriches our understanding of the subtle differences and relationships between this range of different responsibility concepts.

These first two chapters develop the idea that the word "responsibility" refers to a range of subtly different and inter-related concepts, they provide examples of how this affects a range of important practical debates, and they offer suggestions about how this bears on compatibilist responsibility theory. The discussion that follows in the next six chapters can be fruitfully understood as addressing these different facets of the notion of responsibility.

Michael Smith's paper "Beyond Belief and Desire: or, How to Be Orthonomous" provides a detailed account of an important capacity which on his account is required for responsible moral agency – namely, the capacity to be orthonomous. On Smith's account, the standard belief-desire account of action explanation is attractive in part because it explains two crucial distinctions: (i) the distinction, among the things that we do, between the active and the passive, and (ii) the distinction, among the things that we do with respect to which we are active, between those that we do intentionally and those that we do not do intentionally. However, on his account there is a crucial flaw in this standard belief-desire account of action explanation. The account is normally taken to posit four basic elements – two psychological (a desire for an end and a means-end belief), one non-psychological (a bodily movement), and a relation that holds between them (a causal relation of the right kind) – whereas in reality there are five basic elements. Namely, he argues that our being rational to the extent that we are plays a distinct explanatory role. On this account, once we acknowledge the presence of this extra psychological element in every action explanation, we put ourselves in a position to explain a third crucial distinction, one which crosscuts the second: (iii) the distinction between the things that we do rationally and those that we do not do rationally. More generally, we open the door to an understanding of ourselves as orthonomous agents – that is, as agents who have the distinctive capacity to be ruled by the right (to be rational) as opposed to the wrong (to be irrational) – and of the ways in which, when faced with temptations, we may exercise the distinctive capacity we have to be orthonomous. On Smith's account, it is our capacity to be orthonomous that explains when and why we are appropriately held responsible.

A central component in compatibilist thinking is *capacitarianism* – i.e. the idea that *responsibility tracks capacity*. But precisely how is it that changes in capacity (e.g. in a person's *mental* capacities) are meant to result in changes to responsibility (e.g. in what that person can be blamed for doing)? Furthermore, responsibility is also often viewed as something that comes in *degrees* – i.e. people's responsibility is said to be *diminished* by various mental conditions, and the *severity* of punishments imposed onto wrongdoers is also said to co-vary with (among other things) the degree of their responsibility – but exactly why does responsibility come in degrees rather than being an all-or-nothing concept? Also, is there any *relationship* between the former two observations – i.e. that responsibility comes in degrees and that it tracks capacity – and if so then what is it?

Rosemary Lowry's paper "Blame, Reasons and Capacities" answers these questions in two stages. At the first stage, Lowry argues that a person is *blameworthy* when they fail to do what they have most *reason* to do, and that what they have *reason* to do is in turn constrained by what they *can* do (i.e. by what capacities they possess and by what opportunities are available to them). This first stage sets up a chain of justificatory relations between claims about *capacity*, claims about *reasons*, and claims about *blame* (and thus about responsibility, at least for blameworthy states of affairs). At the second stage, Lowry extends Michael Smith's procedure for determining whether an agent possesses certain rational capacities, so that it can be used to measure the degree to which a person *can* do something – i.e. to measure

the degree to which they have certain capacities and opportunities. According to this procedure, a person who can perform a given act – the act for which the capacity of interest is said to be a pre-condition – in a *wide range* of possible worlds, has the requisite capacity to a *larger degree* than a person who can only perform that act in a *narrower range* of possible worlds. Thus, on Lowry's account, blame comes in degrees because reasons come in degrees, reasons come in degrees because capacities come in degrees, and the procedure which she develops helps us to measure these degrees of capacity and thus the degrees of responsibility.

Normally, and in line with Lowry's discussion, reductions in relevant mental capacities (e.g. a reduced ability to control one's conduct) result in judgments of diminished responsibility (e.g. diminished or even completely extinguished blame). This is a straight forward consequence of the above-mentioned capacitarian idea that responsibility tracks capacity. However, although reduced capacity can often be cited as an exculpating factor, sometimes it can not. For instance, under many jurisdictions, drink-drivers and other self-intoxicated parties are not permitted to cite their subsequently reduced mental capacities as a defense to an accusation of criminal responsibility. This feature of the criminal law seems like a practical manifestation of John Fischer and Mark Ravizza's *tracing principle*, according to which the exculpatory value of capacity reductions for which agents are responsible should either be *discounted* or even completely *extinguished* (Fischer and Ravizza 1998:49–51).

However, in "Please Drink Responsibly: Can the Responsibility of Intoxicated Offenders be Justified by the Tracing Principle?" **Susan Dimock** argues that the tracing principle does not in fact support the policy of disallowing self-intoxicated parties from citing their diminished mental capacities as an excuse to the accusation of criminal responsibility. On Dimock's account, there is nothing inherently culpable in the choices that are made by most defendants who eventually become intoxicated, and this is especially so in light of the fact that instead of prohibition, the state rather urges the public to "drink responsibly", and it even happily – almost complicitly – collects substantial revenues from taxation on alcohol sales. Hence, Dimock continues, there is no reason to suppose that such people are blameworthy for becoming intoxicated, and thus no plausible ground for alleging that they are responsible for their subsequent loss of capacities. Put a different way, on Dimock's account there is no culpable action to which self-intoxicated people's diminished capacities can usually be traced back, and thus the discounting function of the tracing principle should not as a general rule be applied to these cases. If Dimock is right, then what is so often taken as a textbook example of the tracing principle (i.e. self-intoxication) is not an example of it at all, and so urgent legal reform is needed in jurisdictions which currently disallow self-intoxicated people to cite their diminished capacities as a defense to the accusation of criminal responsibility, or which (as in Canada) even allow evidence of self-intoxication to be substituted for evidence of fault or *mens rea*.

The tracing principle is sometimes also cited to explain what justifies blaming people for their unintended omissions and for other instances of negligence. Negligence presents a problem because the blamed parties did not realize what they

ought to have done – this lack of knowledge is precisely what distinguishes negligence from recklessness – which means that they lacked the capacity to guide their actions by the appropriate knowledge since that knowledge was not available to them, and this lack of capacity seems in turn to undermine the attribution of blame. However, the tracing retort to this worry is that negligent parties should have realized what they ought to have done – i.e. that they are responsible for their current incapacity – because their incapacity was an outcome of some earlier instance of *non*-negligent blameworthy actions on their part – i.e. it was allegedly a consequence of their earlier *choices*. Blame for negligence is thus meant to derive on this account from blame for an earlier exercise of choice.

However, **Jason Benchimol**'s paper "The Moral Significance of Unintentional Omission: Comparing Will-Centered and Non-Will-Centered Accounts of Moral Responsibility" argues that this use of the tracing principle is deeply problematic, and moreover, that it is not even needed to explain what justifies blame for negligence. On the first point, Benchimol draws attention to four problems. First, even if we can't trace back the current negligent action to an earlier blameworthy exercise of choice, intuitively that does not seem to undermine the claim that the negligent action was itself blameworthy. Second, the search for a prior choice to which the current negligent action can be traced may in fact lead us so far back in time that any blame for the negligent action will be diluted almost to the point of extinction, since only a person with incredible powers of foresight would have recognized that their earlier choice was fraught with danger, but yet some of the most systematic examples of inattention, forgetfulness and other forms of negligence are the most worthy candidates for attracting blame. Third, this view seriously distorts the substantive content for which negligent parties are meant to be blamed and for which they are supposed to atone, since attention is drawn away from their negligence to some temporally (and possibly extremely remote) prior choice which is blameworthy on account of substantively different features. And fourth, the view that *all* blame for negligence derives from prior blameworthy choice can not make sense of why someone who *does* make a prior blameworthy choice but later *does not* act negligently is intuitively less blameworthy than someone whose prior choice is the same but who later *does* act negligently, because *ex hypothesi* all blame must attach to the earlier choice which was the same in both cases. On the second point, Benchimol argues that nothing is gained by insisting that only prior blameworthy choices can ground blame for negligence, because it is not the fact of a choice being a choice per se that makes it a proper target of moral criticism, but the fact that the evaluative attitudes which it expresses are reprehensible, and this component is already present in negligence anyway. Thus, like Dimock, though for different reasons, Benchimol too draws attention to what he believes is a common misapplication of the tracing principle.

While the previous four papers focus on the relationship between mental capacity, role responsibilities, and outcome responsibility, as well as on two instances of alleged misapplication of the tracing strategy that is often cited to explain why not all mental incapacities exculpate, the next two papers focus on the relationship between claims about what a person is outcome responsible for and their liability

responsibility, and on how recalcitrant incompatibilist doubts about the legitimacy of holding people liability responsible for what they do (i.e. legal punishment) can be overcome by drawing on contractualist ideas.

Given the distinction and the relationship that obtains between backward- and forward-looking responsibility claims which Vincent and van de Poel discuss, one way to understand the debate about luck egalitarianism in political philosophy is that it is a debate about how to justify the connection between facts about outcome responsibility (or responsibility as blame in van de Poel's terminology) and conclusions about liability responsibility (or responsibility as liability in van de Poel's terminology). Put more precisely, this debate can be understood as addressing the following two questions: (1) do conclusions about people's liability responsibility (i.e. about how specific individuals may be treated) really follow purely from facts about their outcome responsibility (i.e. about what states of affairs those individuals are responsible for bringing about); and if so, then (2) how do we determine precisely what consequences one must take liability responsibility for (i.e. how one should take responsibility) given the sort of thing for which one is allegedly outcome responsible? Serena Olsaretti (2009) and Nicole Vincent (2009) have both recently described the luck egalitarian debate in this way, and a common feature of their arguments is their insistence that to justify this transition from outcome responsibility to liability responsibility we need to draw upon further substantive normative premises – what Olsaretti calls "a principle of stakes" and what Vincent calls "reactive norms".

However, **Diana Abad**'s paper "Desert, Responsibility and Luck Egalitarianism" argues that the connection between claims about outcome responsibility and conclusions about liability responsibility can be substantially (though not completely) bridged simply by the concept of *desert*. On Abad's account, desert is a three-place relation – X deserves Y in virtue of Z – where X is the desert *subject*, Y is the desert *object*, and Z is the desert *base*. Furthermore, she also suggests that to deserve something means that it is *appropriate* to get it, where *propriety* in turn means that it is both *fitting* and *required* that the desert subject get the desert object in virtue of the desert base. Given this conceptual analysis of desert, Abad then argues that question (2) above (i.e. *how* people should take responsibility) is addressed by the *fittingness* component of the notion of propriety, while question (1) above (i.e. *whether* people should take responsibility for the things for which they are responsible) is addressed by the *requirement* component of the notion of propriety. Put another way, on Abad's account, the concept of desert provides a formal framework within which we can addresses both of these questions in a single move, since what desert entails is that it is *required* that *fitting* treatment be visited onto outcome responsible parties. However, Abad also admits that there are limits to desert as a bridging principle between outcome and liability responsibility – namely, that the *concept* of desert does not itself specify *who* deserves *what* in virtue of *what*, and nor is responsibility the *only base* upon which subjects can deserve various objects. The practical upshot of this is that the same subject can deserve different and incompatible objects (i.e. treatments or liability responsibility) in virtue of different bases, and so substantive normative premises will indeed need to be considered in the process

of reconciling these different desert claims with one another. In effect, Abad's argument entails that although desert can provide a formal link between claims about outcome responsibility and conclusions about liability responsibility, the substantive normative premises at which Olsaretti and Vincent gesture will still be needed to turn these formal links into substantive conclusions about how people may be treated.

Despite many compatibilists' confidence that compatibilism can successfully reconcile legal responsibility with a scientific world view, various authors still worry that although scientific evidence of causation may not undermine responsibility attributions, it may still never the less undermine the law's distinctly backward-looking justifications for its punitive practices. For instance, although Joshua Greene and Jonathan Cohen praise the compatibilist approach, saying that "[c]ompatibilists make some compelling arguments" (2004:1777), never the less they believe that the law's retributive practices can only be defended under libertarianism (which they find wanting). Consequently, they recommend that the law's backward-looking punitive aims should be replaced with such forward-looking punitive aims as deterrence, prevention and treatment. The worry here is not just that the law's actual punitive practices may turn out to be too lenient or too harsh, but that compatibilism simply can't provide the right kind of normative justification to warrant genuinely backward-looking punishment.

Within this context, accepting this incompatibilist challenge **Lene Bomann-Larsen**'s paper "Communicative Revisionism" explores whether contractualism can provide the needed normative foundation for a genuinely backward-looking form of punishment within a broadly compatibilist framework. Echoing the concerns of Greene and Cohen, Bomann-Larsen argues that although free will may not be relevant to legal determinations of guilt and responsibility, it does never the less seem relevant to backward-looking desert-based practices like punishment. However, instead of accepting the revisionist recommendations of authors like Greene and Cohen, who would have us abandon the search for such backward-looking justifications for punishment, Bomann-Larsen suggests that contractualism can provide the right sorts of justifications while remaining consistent with compatibilist commitments. However, although on her account this move fits well with a communicative theory of punishment which is thoroughly backward-looking, some revision to the law's actual punitive practices will still be needed because although it is plausible that contracting parties may accept some forms of hard treatment to communicate censure, most would likely reject "extreme punishment . . . which either makes reconciliation with community impossible [e.g. life imprisonment], or which imposes irreparable damage" on the offender (e.g. capital punishment). Thus, although she argues that genuinely backward-looking punitive practices can be justified in a way that respects compatibilist commitments, what must still never the less be revised is the kind and severity of some current legal punitive practices.

Finally, since much compatibilist thinking focuses on the topic of individual responsibility, this book's remaining four chapters investigate theoretical and applied problems in the area of collective responsibility. Although collective

responsibility has attracted a lot of philosophical discussion, the compatibilist literature on responsibility has tended to focus on individual responsibility. It is therefore an interesting question whether, and to what extent, compatibilist notions of responsibility can account for cases of collective or joint responsibility.

A good starting point for such a discussion is **Alex Brown**'s contribution on "Moral Responsibility and Jointly Determined Consequences". Brown offers a critical examination of some aspects of Fischer and Ravizza's compatibilist theory of responsibility. Fischer and Ravizza offer both a negative argument for compatibilism – i.e. that the inevitability of certain events or consequences does not rule out the possibility of an agent's moral responsibility for those consequences – and a positive argument – i.e. that moral responsibility for consequences depends on action-responsiveness. With respect to the first, Brown highlights a flaw in Fischer and Ravizza's account of simultaneous over-determination of consequence-universals. With respect to the second, he points out that if, as Fischer and Ravizza propose, action-responsiveness were a necessary condition of an agent's being morally responsible for certain consequences, then an act or omission that only jointly determines a consequence could not attract moral responsibility. Given these problems, Brown argues that in addition to an account of moral responsibility for cases of individual action, we also need accounts of moral responsibility and action-responsiveness which are designed specifically to handle cases of joint action.

In his contribution Brown discusses three types of cases of joint determination of consequences:

- *Cumulative*: The case Brown discusses is *Revolutionaries*, where two revolutionaries independently and unaware of each other try to kill the same governmental official by shooting him. Since both are not trained shooters, only the combined, cumulative effect of their shootings kills the mayor.
- *Joint enterprise*: For example, two *Fundamentalists* cooperate to kill a country leader. Although only one of them pulls the trigger, we are inclined to hold them both responsible for the leader's death.
- *Probabilistic*: Brown's example is a *Firing Squad*, consisting of six good shooters, but only one of the six rifles being loaded. The deserter who is executed is sure to be killed, but the soldier who actually killed the deserter is selected randomly. We are inclined to hold the six soldiers jointly responsible.

In none of these examples is Fischer and Ravizza's condition of action-responsiveness met. As Brown points out, this condition, among other things, requires that the action of the person held responsible is *sufficient* to bring about the consequence, but yet in none of the above cases does this obtain. To deal with this problem, Brown proposes an alternative notion of action-responsiveness for cases of joint responsibility: "An individual's act or omission is jointly action-responsive with respect to a consequence C if and only if it along with the acts or omissions of at least one other individual is part of a process of type P which is sufficient for C to obtain and a different consequence would have obtained if it along with the other

acts or omissions had not occurred and all other triggering events that would have been sufficient to cause C are not in play."

In his paper "Joint Responsibility without Individual Control: Applying the Explanation Hypothesis", **Gunnar Björnsson** further explores cases of what Brown calls cumulative determination of consequences. The example he discusses is, however, in one respect distinguishably different from Brown's *Revolutionaries*. In *Revolutionaries*, the individual contributions are not sufficient for the overall effect, but they are necessary. In Björnsson's example, *The Lake*, they are neither sufficient nor necessary. In *The Lake*, three people pour an equal amount of solvent into a lake with the consequence that the fish are killed. In this example, two amounts of solvents are enough to kill the fish. Björnsson argues that in cases like this most people have a strong intuition that the individuals are responsible, but it is not clear how this intuition is to be explained. In cases like *The Lake*, individuals jointly cause a detrimental effect without being individually in control of the effect. Since the individual contributions are neither necessary nor sufficient for the effect, accounts of responsibility in terms of "difference making" do not work in cases like this. By working through a number of variations on *The Lake* example, Björnsson also points out that accounts in terms of causal involvement or causal facilitation do not work either. Existing accounts of joint responsibility in terms of joint actions, joint intentions or social ties fail due to another feature of *The Lake* example – namely, that the individuals are unaware of each other's existence, and hence we can't even plausibly claim that their responsibility stems from their membership in a relevant social group.

To address these problems, Björnsson offers a reconstruction and explanation of our intuitions in cases of cumulative actions or omissions. To this end, he employs his own "explanation hypothesis", which he developed in an earlier paper. This hypothesis states that the motivational structure of an agent should be an integral part of the explanation for why the (undesirable) outcome occurred, for that agent to be responsible for the outcome. Björnsson argues that in cases like *The Lake* we consider the three individuals individually responsible "because their motivational structures are part of a significant explanans only taken together with the motivational structures of the other two." Focusing on only one of them is not only unsatisfactory because the individual contributions do not make a difference to the outcome, but also because it neglects the role of the other two, who, from an explanatory point of view, played exactly the same role.

Steven Vanderheiden focuses on still further kinds of cases of joint responsibility. Vanderheiden asks whether persons can reasonably be held morally responsible for harmful consequences that result from the acts or omissions of their nation or society, even if they as individuals try to avoid contributing toward those consequences. The example he discusses is global climate change. The 1992 UN Framework Convention on Climate Change holds nations responsible for the climate-related harm that they cause through their greenhouse gas emissions. Vanderheiden argues that this implies holding citizens of those nations responsible (in the liability responsibility sense) even if some of them did not contribute to the harm (i.e. even though they were apparently not outcome responsible),

which makes this responsibility attribution seem unreasonable. Vanderheiden argues that the attribution is nevertheless just because all citizens participate in a culture that permits or even encourages consumption patterns that directly contribute to global warming. Moreover, all citizens – even those who actively oppose current government policies – benefit from the consumption patterns, and hence on his account all citizens are therefore responsible although not necessarily to the same degree.

The example of climate change that Vanderheiden discusses is somewhat similar to the pattern of joint enterprise discussed by Brown. It has, however, some features that are distinctively different from cases like *Fundamentalists*. In *Fundamentalists*, all participants actively and willingly cooperate to attain the effect, whereas Vanderheiden is also interested in those participants that do not actively or willingly cooperate. The key question for Vanderheiden is thus what would be required to "extricat[e] oneself from responsibility for harm caused by group actions." One option he considers is voting for candidates or policies that aim at avoiding contributing to global warming, but he replies that such voting alone is not sufficient to avoid responsibility. This is especially not sufficient if this voting is likely to be ineffective in terms of overall policies, and if the dissenting voters continue to profit from the collective benefits of the harmful practices. If ineffective voting would be enough to generate an excuse, then it would invite moral free riding: enjoying the collective benefits while bearing none of the responsibility for creating these benefits through ineffective "resistance". Thus, following David Miller, Vanderheiden suggests that the only sincere opposition is to refuse to partake in the benefits of society's harmful practices. But entirely refusing such benefits may be practically impossible for citizens of affluent societies as some of the benefits are public goods and one is born and raised in affluent circumstances that are themselves the result of those unjust practices. This is not to deny that one can extricate oneself from some responsibility, but, if we follow Vanderheiden, not from *all* responsibility for the negative consequences of climate change. Interestingly, Vanderheiden's conclusion has a distinct compatibilist flavor: one can be responsible for consequences even if these consequences are inevitable.

The final paper by **Seumas Miller** provides a methodologically-individualistic theoretical framework for reasoning about collective responsibility – and in particular, about collective responsibility for epistemic actions – and it demonstrates this framework's usefulness by investigating the topic of our collective responsibility for harms due to human-induced global warming and climate change. Miller begins by outlining a theory of *joint action* as the action of individuals within stratified organizational frameworks that define specific role responsibilities for constituent individuals, or what he calls "multi-layered structures of joint action characteristic of organizational action". His stated explicit aim throughout this chapter is "to display the continuity between individual and institutional moral responsibility for actions: the continuity between, for example, the moral responsibility for harmful climate change of individual citizens and that of governments". However, on Miller's account many important issues in the debate about collective responsibility for harms due to human-induced climate change hinge on *epistemic* factors that in

his view have received insufficient attention in the literature, and so to this end he adapts the previously-developed theory to take account of *joint epistemic action*. An account of *collective moral responsibility* for joint (epistemic and behavioural) action (and omissions) is then developed, and it is subsequently put to use to explain who has, and in what degrees, "both retrospective responsibility for causing harm and also prospective responsibility for addressing the problem in terms of mitigation and/or adaptation".

This book's chapters deal with a range of theoretical problems discussed in classic compatibilist literature – e.g. the relationship between responsibility and capacity, the role of historical tracing in discounting the exculpatory value of incapacities, and the justifiability of retributive punishment. But instead of motivating their discussions by focusing on the alleged threat that determinism poses to responsibility, these chapters' authors have animated their discussions by tackling practical problems which crop up in contemporary debates about responsibility in the hope of both applying and extending compatibilist thought. For instance, how is the narrow philosophical concept of responsibility that was defended from the threat of determinism related to the plural notions of responsibility present in everyday discourse, and how might this more fine-grained understanding of responsibility open up new vistas and challenges for compatibilist theory? What light might compatibilism shed, and what light might be shed upon it, by political debates about access to public welfare in the context of responsibility for one's own health, and by legal debates about the impact of self-intoxication on responsibility. Finally, and perhaps most importantly, does compatibilist theory, which was originally designed to cater for analysis of individual actions, scale to scenarios that involve group action and collective responsibility – for instance, for harms due to human-induced climate change – or must compatibilism be modified in some way to handle collective responsibility scenarios and problems? Although the range of topics covered by the papers in this volume is broad, what ties them together is their authors' commitment to using the foundations of compatibilist theory to address important moral, political and legal questions about responsibility, and to develop this otherwise largely-negative branch of philosophy in a distinctly positive and practical direction.

Acknowledgments The papers in this volume are a selection of papers presented at the international conference *Moral Responsibility: Neuroscience, Organization & Engineering*, that was held in Delft, The Netherlands, on August 24–27, 2009, as well as some invited contributions. This volume was composed as part of the research programs "The Brain and The Law" and "Moral Responsibility in R&D Networks", both supported by the Netherlands Organization for Scientific Research (NWO). The editors express their gratitude to Malik Aleem Ahmed for helping them to prepare this manuscript for publication despite our constantly-shifting deadlines and a never-ending stream of "another thing please" requests. Special thanks also go to Antony Duff, Walter Glannon and Neil Levy for providing timely, insightful and helpful feedback – often on successive drafts of the same papers – as well as to Maja de Keijzer and Nicoline Ris from Springer and Springer's anonymous referees. Ibo van de Poel is grateful to the Netherlands Institute for Advanced Study (NIAS), for providing him with the opportunity, as a Fellow-in-Residence, to work on this volume during his stay in the academic year 2009–2010. And Nicole Vincent is grateful for the generous financial and moral support of the Philosophy Department at Delft University

of Technology, as well as the 3TU Centre for Ethics and Technology (3TU.Ethics), but also to Saskia Polder-Verkiel for helping her to plan the aforementioned conference, for making things happen when the pressure was on, and for her feedback on oral presentations of many of the papers contained in this volume.

References

Dawkins, Richard. 2006. "Let's all Stop Beating Basil's Car." Accessed February 12, 2007. http://www.edge.org/q2006/q06_9.html#dawkins.

Dennett, Daniel Clement. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge, MA: MIT Press.

Fischer, John Martin and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge, UK: CUP.

Gazzaniga, Michael S. 2006. "Facts, Fictions and the Future of Neuroethics." In *Neuroethics: Defining the Issues in Theory, Practice, and Policy*, edited by Judy Illes, 141–48. New York, NY: Oxford University Press.

Greene, Joshua and Jonathan Cohen. 2004. "For the Law, Neuroscience Changes Nothing and Everything." *Philosophical Transactions of the Royal Society of London* 359:1775–85.

Hohfeld, Wesley Newcomb. 1975. "Rights and Jural Relations." In *Philosophy of Law*, edited by Joel Feinberg and Hyman Gross, 357–67. Belmont, CA: Wadsworth Publishing Company.

Morse, Stephen J. 2006. "Moral and Legal Responsibility and the New Neuroscience." In *Neuroethics: Defining the Issues in Theory, Practice, and Policy*, edited by Judy Illes, 33–50. New York, NY: Oxford University Press.

Olsaretti, Serena. 2009. "Responsibility and the Consequences of Choice." *Proceedings of the Aristotelian Society* (Hardback) 109(1pt2):165–88.

Vincent, Nicole. 2009. "What Do You Mean I Should Take Responsibility for My Own Ill Health?" *Journal of Applied Ethics and Philosophy* 1:39–51.

Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

Chapter 2
A Structured Taxonomy of Responsibility Concepts

Nicole A Vincent

Abstract This paper distinguishes six different responsibility concepts from one another, and it explains how those concepts relate to each other. The resulting "structured taxonomy of responsibility concepts" identifies several common sources of disputes about responsibility, and it suggests a procedure for resolving such disputes. To demonstrate their utility, this taxonomy and procedure are then used to illuminate debates in two familiar contexts.

2.1 Introduction

In philosophical discussions, responsibility is often talked about as if it were a single, unitary and generic concept, or at least that is the impression that such discussions can create. For instance, the compatibilist literature considers whether determinism rules out *responsibility* (Wallace 1994; Fischer and Ravizza 1998; Pereboom 2001; Dennett 2003). In political philosophy, luck egalitarians and their opponents argue about the extent to which people's entitlements should track their *responsibility* (Dworkin 1981; Cohen 1989; Rakowski 1991; Anderson 1999; Arneson 2000). And in the area of neurolaw, some people claim that "a truly scientific mechanistic view of the nervous system make[s] nonsense of the very idea of *responsibility*" (Dawkins 2006, emphasis added; also Greene and Cohen 2004), while others deny that neuroscience challenges *responsibility* (e.g. Morse 2006; Gazzaniga 2006). If confronted head on, most philosophers would probably acknowledge that the term "'responsibility' admits of a variety of uses" (Fischer and Ravizza 1998:2, note 1), but somehow this variety seldom comes across in anything other than footnotes.

This paper aims to remedy this situation by saying something more substantial about this variety of uses of the term "responsibility" and by explaining how we are better off for knowing about it. Section 2.2 distinguishes six different senses of the term "responsibility" – six different responsibility concepts – and it introduces some terminology to keep these concepts disambiguated. Section 2.3 explains how these

N.A Vincent (✉)
Department of Philosophy, Delft University of Technology, 2600 GA Delft, The Netherlands

Department of Philosophy, Faculty of Arts, Macquarie University, Sydney NSW 2109, Australia
e-mail: nicole.vincent@mq.edu.au; n.a.vincent@tudelft.nl

N.A Vincent et al. (eds.), *Moral Responsibility*, Library of Ethics and Applied Philosophy 27, DOI 10.1007/978-94-007-1878-4_2,

responsibility concepts relate to each other – namely, via justificatory relations that obtain between claims which employ them. Together, the first two sections describe what I call a *structured taxonomy of responsibility concepts* (STRC), and the following two sections explain why this taxonomy is useful and not just theoretically neat. Section 2.4 argues that the STRC helps us to identify fifteen distinct sources of disputes about responsibility as well as a corresponding procedure for resolving such disputes, while Section 2.5 demonstrates the utility of this taxonomy and procedure by applying them to two familiar contexts in political philosophy and in tort law.

2.2 Six Concepts[1]

A close examination of ordinary language use reveals that the word "responsibility" refers to a number of different though related ideas. To see this, consider the following parable about Smith the ship captain, originally developed by H.L.A. Hart[2]:

> (1) Smith had always been an exceedingly *responsible* person, (2) and as captain of the ship he was *responsible* for the safety of his passengers and crew. But on his last voyage he drank himself into a stupor, (3) and he was *responsible* for the loss of his ship and many lives. (4) Smith's defense attorney argued that the alcohol and his transient depression were *responsible* for his misconduct, (5) but the prosecution's medical experts confirmed that he was fully *responsible* when he started drinking since he was not suffering from depression at that time. (6) Smith should take *responsibility* for his victims' families' losses, but his employer will probably be held *responsible* for them as Smith is insolvent and uninsured.

The word "responsibility" is used here in at least six different ways, each of which suggests a subtly different responsibility concept, and I will now give each of these concepts a name so that the upcoming discussion can proceed without the ambiguity inherent in using the generic term "responsibility".[3]

First, there is a claim about his *virtue responsibility* – Smith was normally a dependable person who took his duties seriously and did the right thing. To call somebody "responsible" in this sense is to say something good about their character, reputation or intentions, as exemplified by their history which testifies to their manifest commitment to doing what they take to be right. The opposite of calling someone "responsible" in this sense is to call them "irresponsible" (Vincent 2009).[4]

---

[1] The material contained in this section is an elaboration of Vincent (2010:80–82).

[2] This parable is an adapted version of Kutz (2004:549), who in turn derived his version from Hart (1968:211). Hart did not give the captain a name, but I find it helpful to do so.

[3] Although I use H.L.A. Hart's terminology to preserve continuity with his work, there are differences between my and his terminology. Also compare to Chapters 1 and 3 by van de Poel, this volume.

[4] Gary Watson's *aretaic* sense of responsibility is very similar (2004). Also compare (e.g. Haydon 1978; Williams 1995; Bovens 1998; Duff 1998:291; Williams 2008).

Second, there is a claim about Smith's *role responsibility* – as the ship's captain Smith had certain duties to various parties, both on and off his ship. I do not mean to imply, by using the word "role", that we only have role responsibilities in virtue of our institutional, social or conventional roles, nor that we can settle who has which responsibilities simply by examining their roles. As Garrath Williams (2008) points out, duties can arise from plural sources and not just from formal roles. Furthermore, it is also conceivable that a single individual may be subject to a range of conflicting demands at the same time – some related to their various roles (we all wear many hats), and others to "the imperatives of basic human decency" (Williams 2008:467) – and that a conscientious person must find the right way to balance them (Williams 2008:459). Used in this second sense, "responsibilities" refers to duties – to what a person should (not) or ought (not) to do.[5]

Third, there is a claim about his *outcome responsibility* – it is alleged that various states of affairs, such as the loss of the ship and many of its passengers and crew, are rightfully attributable to him as something that he did, and perhaps even as something for which he is blameworthy. Normally, only agents are responsible for things in this sense – for instance, a chair or table can't be outcome responsible for anything – but sometimes legal entities (e.g. corporations) are also treated as agents and are claimed to be responsible for things in this sense. Furthermore, the word "outcome" should not be taken to imply that agents can only be responsible for outcomes, since they are also responsible in this same sense for their actions. I take this to be the sense which philosophers usually have in mind when they talk about responsibility.[6]

---

[5] Unlike Robert Goodin (1986; or 1987) who distinguishes "task responsibilities" from other duties, I use "responsibilities", "role responsibility" and "duties" interchangeably to refer to the various things which we should or ought to do.

[6] I use Stephen Perry's term "outcome responsibility" (2000:555) because it captures the idea of a form of responsibility which looks backwards in time to states of affairs (outcomes or actions) that occurred in the past, and for which the person in question is blameworthy (if what they are responsible for is bad) or perhaps praiseworthy (if it is good), but others have given this concept different names. For instance, Hart calls it "causal responsibility" (1968:212), though I find this name unhelpful since it runs together at least two distinct ideas – i.e. the *normative* concept of responsibility (e.g. see Kutz (2004:555); citing Wallace (2002)) and what I take to be its causal component. Fischer and Ravizza call it "moral responsibility" (1998), and they too distinguish between moral responsibility for actions and for outcomes; though I am not fond of this name either since although it captures the inherently *normative* nature of responsibility, it fails to adequately differentiate between our *forward*-looking moral responsibilities (our "role responsibilities" comprise some of these, and what I will shortly call "liability responsibility" comprises the rest of them) and the *backward*-looking moral responsibility which I am presently calling "outcome responsibility". Also, Peter Cane calls this concept "historical responsibility" (2004:162); Thomas Scanlon calls it "responsibility as attributability" (1998:248); Gary Watson calls it the "attributability" sense of responsibility (2004:263–65); and Garrath Williams and Antony Duff call it "retrospective responsibility" (Williams 2008:457, 459, 460 and 467; Duff 1998). Finally, although Herbert Honoré also uses the label "outcome responsibility" (e.g. Honoré 1999), he uses it to refer to the concept which below I call "liability responsibility".

Fourth, there are two claims about *causal responsibility* – Smith's defense lawyer alleged that Smith's aberrant behaviour was caused by the alcohol and/or by his depression. Used in this way, the word "responsibility" is a synonym for words like "cause" and "condition", and we could re-phrase what the defense lawyer said without loss of meaning as the claim that the alcohol and depression caused (or that they were conditions of) Smith's aberrant behaviour. However, people's *actions* are also very often cited as causes or conditions of various outcomes, and although in doing this we pick out that person's actions as particularly significant in the production of those outcomes, this is not yet the same as the full moral allegation that they are (in the terminology introduced in the previous paragraph) outcome responsible. Hence, we might say that causal responsibility is a thinner and less morally imbued concept than outcome responsibility.

Fifth, there is a claim about his *capacity* responsibility – since Smith was not suffering from depression at that time, the prosecution alleged that his mental capacities were fully intact, which meant that his moral agency was unimpaired. The capacities in question are usually conceived of as the so-called "cognitive" and "volitional" capacities of folk psychology – as Hart describes them, "the capacity to understand what [we are] required … to do or not to do, to deliberate and to decide what to do, and to control [our] conduct in the light of such decisions" (1968:218) – and typical cases of agents who lack responsibility in this sense (i.e. who are not (yet) fully responsible) are children and the mentally ill. When capacity responsibility is conceptualized in this mentalistic way, it can be usefully contrasted with the concept of virtue responsibility introduced first above, since a fully (capacity) responsible person may at the same time be a very (virtue) irresponsible person, and someone who is not yet a fully (capacity) responsible person may still be a very (virtue) responsible person (Vincent 2009). Although capacities can come in degrees – for instance, one might be more or less intelligent, or have more or less strength of will to resist temptation – we often set thresholds that must be reached before a person is considered fully responsible. However, although this example cites a *mental* kind of capacity, it is plausible that non-mental factors – e.g. a person's physical strength or the tools at their disposal – may also affect their capacity responsibility. When a person's non-mental capacities are diminished though, we would not usually say that they are not fully responsible, but rather we might simply acknowledge that they lack the corresponding capacity.

And finally, the parable ends with comments about *liability* responsibility – about who will now be held responsible (and how) for what has happened. In this case, financial liability is cited – that is apparently one way to take responsibility – but perhaps Smith should also apologise to the bereaved families and go to prison to properly take due personal responsibility for what happened. When "responsibility" is used in this way, it is usually coupled with another word – i.e. *take* responsibility or *hold* responsible – and it refers to the things that someone must do, or how they should be treated, to set things right.

I take the fact that "responsibility" can be used in so many different ways, to show that this word refers to a "syndrome" of concepts – i.e. to multiple concepts

that share a common word – rather than to a single, unitary or generic concept. Furthermore, to my mind at least all of these uses of "responsibility" are equally legitimate – none of them seems merely metaphorical, or like a misuse of the word – and I think that very little if anything is gained by claiming that some of these senses of "responsibility" are more primary while others are more derivative.

2.3 Relations Between These Six Responsibility Concepts[7]

Hart's parable also suggests that claims which employ these concepts stand in certain justificatory relations with respect to one another. Here are three examples: Smith should now do certain things (i.e. take liability responsibility) precisely because of his outcome responsibility; it is precisely because he had certain role responsibilities (which he subsequently violated) that we attribute outcome responsibility to him for the ship's loss, but not to (e.g.) the ship's chef since he did not have them; and the parable also hints that Smith's outcome responsibility may have been reduced had he been suffering from depression at the time, as that would have reduced his capacity responsibility. In fact, moving beyond Hart's parable, it is plausible that a number of other relations also obtain between claims that employ these different responsibility concepts, and these relations can be expressed using a structure diagram of the sort that is sometimes used to chart relations that obtain between premises and the conclusions which they are intended to support in philosophical arguments:

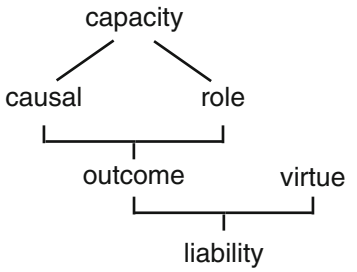

Fig. 2.1 Lines represent justificatory relations between connected responsibility concepts

Three groups of justificatory relations are expressed in this diagram – (1) that claims about outcome responsibility are derived from claims about causal and role responsibility, (2) that claims about capacity responsibility bear on claims about causal and role responsibility, and (3) that claims about liability responsibility are derived from claims about outcome and virtue responsibility – and in what follows I will explain why I think that these relations obtain, as well as characterise the nature of these relations in some detail.

[7] The material in this section combines parts of Vincent (2010:82–85, 2011:80–82, Forthcoming: Section 3.8.). Diagram originally from Vincent (2010:82).

2.3.1 Outcome Responsibility from Causal and Role Responsibility

First, claims about a person's *outcome* responsibility seem to depend on prior claims about their *causal* responsibility and their *role* responsibility.

To see why this might be so, imagine that you stumble upon Jones' dead body while strolling through a forest. It seems that Jones died of a gunshot wound to the head, and that his body was then hastily concealed beneath the bush where you found him. Consider now what processes we might engage in to determine who is responsible (in the outcome responsibility sense) for Jones' death.

Our first question would probably be "Who dunnit?", and so in the beginning our inquiry would focus on discovering such things as who was where at the time of the crime, how they behaved, and what consequences their behaviour brought about. If insufficient information was available then we would probably gather up witnesses and suspects, conduct a line-up to identify prime suspects, and then interrogate some people; in court, both witnesses and suspects might eventually testify, and physical evidence such as finger prints, spent bullet cartridges, DNA samples and so on might also be collected, examined and presented. Many epistemic barriers may stand in the way of answering the *who dunnit* question, but once these puzzle pieces are put together we may discover that Brown is the one who shot Jones dead – or put another way, that Brown is causally responsible for Jones' demise.

But to establish that Brown is outcome responsible for Jones' death, we need to show more than just that she "dunnit". Rather, given causal indeterminacy – i.e. the fact that any outcome is a result of many causal contributions – we must also show that her causal contributions were of particular significance, and the way that we commonly do this is by looking at whether Brown acted contrary to how she should have acted – or put another way, whether she violated her role responsibilities. For instance, Brown might have shot Jones in *self-defence* when he ambushed her on her stroll through the forest, and as long as what she did is viewed as a reasonable response – e.g. not an unwarranted use of extreme force – then this should suffice to establish that Jones was at fault for his own death (because he should not have attacked her), and thus Brown's causal responsibility would not translate into a finding that she was outcome responsible for his death. Alternatively, suppose that Brown was out hunting in a well-known, sign-posted and cordoned-off area of the woods which was only supposed to have other hunters in it, all of whom wore brightly coloured clothing and who knew each others' locations through GPS devices, and that there was simply no way for her to know (nor any reason to suspect) that Jones – a thrill-seeking prankster who liked to frustrate hunters by hiding in bushes and scaring away their game – was hiding in those bushes. In this case it is again likely that Jones' rather than Brown's actions will be viewed as the salient causes of his own demise, and hence that despite Brown's causal responsibility, Jones again would be picked out as the person who is outcome responsible for his

own death, because Brown acted *reasonably* (her actions were not unduly risky) whereas Jones did not (his actions were too risky).[8]

The point of these examples is that to turn a finding of causal responsibility into the fully-fledged moral accusation of outcome responsibility, the party whose actions causally contributed to the said outcome must also have violated their role responsibilities in acting like that. As Joel Feinberg puts it, "blame-fixing" requires both a *genuine causal relationship* between the thing that does the causing and the thing that is allegedly caused, as well as a relevant *moral element* (1970:207). Hence, this is why I suggest that claims about a person's *outcome* responsibility gain support from claims about their *causal* responsibility and their *role* responsibility.[9]

2.3.2 Capacity Responsibility to Causal and Role Responsibility

Second, claims about a person's *capacity* responsibility seem to bear on what may legitimately be said about their *causal* responsibility and *role* responsibility.

Regarding the relationship between capacity responsibility and causal responsibility, suppose that Brown's defense had been that she shot Jones while sleepwalking. If this "automatism" defense were accepted as a truthful account of what happened, then the way in which it might work in her favour is by denying that her body movements even count as instances of her actions. Such body movements would be conceptualized as something that *happened to* Brown rather than as something that *she did* – her body movements would not be conceptualized as genuine instances of actions, let alone as her actions – and that in turn would undermine the claim that her actions were causally responsible for Jones' death.

On the other hand, suppose that Brown's defense had been that she shot Jones because God commanded it – a symptom of her severe mental illness. This defense would attempt to get her off the hook by alleging that she can not be blamed because due to her delusion she lacked the pertinent cognitive or volitional capacities – e.g. that she lacked the capacity to know what she was doing, that what she was doing was wrong, or to control her conduct. This is also how we think about children and others whose mental capacity falls below the minimum threshold of fully responsible moral agency – i.e. their reduced capacity is taken as a reason to excuse them, or to simply expect less of them in the first place, and thus subsequently to morally blame them for less. Put another way, this sort of defense presupposes that people are blameworthy when their actions breach their role responsibilities (this is just

[8] My point is not that we would necessarily make this particular evaluation about the risks involved, but rather that *if* we made that evaluation *then* Smith would be deemed responsible.

[9] Gary Watson also argues that we would not usually say that someone was responsible for a bad outcome unless their actions were "contrary to one of [their] responsibilities" (2004:274)

conceptual analysis), and that what role responsibilities a person is subject to is something that depends at least in part on what capacities they (ought to) possess.[10]

In broad terms, people with greater capacities are usually expected to conform to higher standards – i.e. we are *capacitarian* in the sense that we generally hold that *responsibility tracks capacity* – and this expectation can be explained in either a *positive* or a *negative* way. In the positive explanation capacity *generates* duties. The idea is that we ought to do what we have most reason to do, and what we can and can't do (presumably along with a range of many other things) generates the reasons that we have to do various things. An inference is thus first made from what *capacities* I have to what I have *reason* to do, and then another inference is made from what I have most *reason* to do to what I *ought* to do – i.e. we move from *capacity* claims via *reasons* claims to *ought* claims.[11] On this account, if I can not save a child from drowning – perhaps because I do not know that they are drowning, or because I can not swim, or because I do not have a rope to throw to them – then it is simply not true that I ought to save them (unless I am responsible for the fact that I can not do this – see note 10 above). The reason why I would not be blameworthy for not saving them is because I was not in the first place even subject to that saving duty. On the other hand, in the negative explanation capacity *regulates* duties. The idea here is that regardless of the source of our duties, on this second view our *in*capacities can excuse departures from those duties. On this latter account, the three cited considerations – i.e. I don't know that the child is drowning, I can't swim, or I have no rope – do not extinguish the saving duty, but rather they provide an excuse for departing from it. The reason why I would not be blameworthy on this second account is because although I did have the saving duty, my incapacity provided an excuse for departing from it.

Two advantages of the negative explanation are that only it has an explicit place for excuses (and perhaps justifications) which play a prominent role in much ordinary and legal thinking about responsibility, and arguably it also more adequately captures the rich structure of practical reasoning in which some considerations *discount*, *undermine* and *invalidate* (rather than just *outweigh* or *extinguish*) other considerations. Never the less, I suspect that both views of the relationship between capacity responsibility and role responsibility will generate the same conclusions about when someone is outcome responsible for some state of affairs, and since I find the positive explanation simpler, in what follows my discussion will be framed

[10] The "ought to" clause captures the idea, expressed for instance in Fischer and Ravizza's *tracing* principle (1998:48–51), that diminished capacities for which a person is responsible have at most only a discounted exculpatory value – people who cause accidents while voluntarily intoxicated are typically not excused for what they did despite their reduced capacities. For contrast, see Chapters 6 and 7 by Dymock and Benchimol, this volume.

[11] There is room in both inferences for other considerations too: my capacity (e.g. to swim) may be a reason to do various things (e.g. save the drowning child, join the swimming team, etc.), and other reasons (e.g. that I am running late for work) may compete with the capacity-based reasons for determination of what I ought to do. See Chapter 5 by Lowry, this volume, for a discussion of the relationship between *can* claims and *ought* claims via *reasons* claims.

in terms of it rather than in terms of the negative explanation.[12] Thus, the relationship which I think obtains between capacity responsibility and role responsibility, and which is expressed in the above diagram, can be stated as follows. The idea is not that we can read off a person's responsibilities simply from an assessment of their capacities (this would be an instance of *can implies ought*), but it is rather that in determining what responsibilities a person has we should, among other things, consider what capacities they (ought to) possesses – i.e. the idea is that *can*, taken together with a range of other considerations, implies *ought*.

2.3.3 *Liability Responsibility from Outcome and Virtue Responsibility*

Finally, claims about *liability* responsibility – that is, about who should be held responsible, and about how (the kind and extent) they should be held responsible – seem to be affected by claims about that person's *outcome* responsibility and by the degree of their *virtue* responsibility or lack thereof.

This is particularly easy to see in the criminal law where punishment (one form of liability responsibility) is normally only imposed onto those parties who were previously established as outcome responsible for the corresponding offence, and the kind and extent of their punishment normally depends at least in part on the seriousness of the sort of thing for which they were outcome responsible (e.g. homicide, arson, theft, etc.) and on the degree of their outcome responsibility for it. But we find similar thinking beyond the criminal law too. For instance, whoever broke the vase is normally expected to make up for it; if two people were responsible for breaking it (suppose it fell to the floor when they knocked the table as they chased each other down the hall) then the extent of each one's liability responsibility will usually track the degree of each one's outcome responsibility; and if the vase was a precious antique then their liability will undoubtedly be greater than if it had been a mass-produced IKEA flower pot.[13]

But another consideration that also seems to play a role is what the person has previously been like – i.e. whether they have previously had a "clean slate" and been virtue responsible individuals, or whether this is merely another example of their

[12] Peter B. M. Vranas (2007:181) argues that the positive or generative account of the relationship between capacity and obligation is more plausible than the second or regulative account, because there are good reasons to suppose that if I really could not have done something, then I really can not legitimately be expected to have done that thing in the first place.

[13] The main exception to this is vicarious liability – i.e. when one person is held responsible for what another person did – but this can either be explained away and criticized as a departure from an otherwise legitimate moral norm, or we could check whether the specific instance might perhaps be justified. For instance, we might check whether the party that will be held vicariously responsible failed to supervise the other party as they ought to have, and that they are after all outcome responsible for what the other person did which means that no injustice is done *qua* holding them liability responsible. This is presumably one reason why parents are held responsible for their children's actions.

generally irresponsible character as testified to by their history of similar actions. Nicola Lacey (2007:240, 245–47) notes that within the criminal law considerations of what a person's character is like can play a mitigating (good character) or an aggravating (bad character) role at sentencing.[14] And similarly, beyond the criminal law it is again a common enough practice to treat more harshly those people who should have learned from their past actions but didn't (i.e. those who have a history of being irresponsible), and to treat less harshly and maybe even to forgive those for whom this was their first infraction or who have even been model citizens. In other words, a person's *virtue* responsibility can be a mitigating factor, and its lack can be an aggravating factor, by modulating the kind and degree of treatment that we impose upon them in order to hold them (liability) responsible for what they did.

2.3.4 Norm Setting and Substantive Evaluations

Lastly, something which isn't depicted in the diagram at the top of Section 2.3 is that norm setting, substantive evaluations and our aims also play important roles in relating different kinds of responsibility claims to one another, and in this subsection I will offer a brief discussion of the role that such considerations play.

Firstly, just how much and what kind of mental capacity it takes to be a fully capacity responsible person, is at least in part a norm setting exercise. In some places a person is deemed to be fully responsible by the time they turn 18 years of age – the assumption being that by then people have matured and possess a *sufficient* degree of the right kinds of mental capacities – but the sufficiency threshold could also be set at a lower or a higher age. Admittedly, we do not have complete freedom to set this threshold wherever we like, because how much and what kind of capacity it takes to be fully responsible depends at least in part on what sorts of expectations will later be thrust onto those who fall into the "fully responsible" category – "fully responsible" is relative to a context. But given that (as I am about to argue) the expectations that we have of one another are themselves informed by further instances of norm setting and ultimately-contestable substantive evaluations, this might in the end only weakly constrain the norms that we choose in regards to capacity responsibility.

Secondly, just how much care a person must take to avoid being deemed negligent in the event of an accident – i.e. to avoid the claim that they breached their role responsibility – is also a norm setting task. Whether this or that amount of care is *sufficient* is something that is up to us to determine – i.e. it is a norm that we must set. Admittedly, the norms that we set in this regard are not arbitrary either, since they arguably reflect our commitment to (e.g.) efficiency. For instance, whether A is negligent for causing B's losses depends on whether A took *reasonable* precautions, and according to the Learned Hand test what is reasonable in turn depends on the costs involved in taking those precautions weighed against the risk-depreciated savings

[14] The so-called "three strikes laws" demonstrate how a judgment that someone lacks virtue responsibility can be an aggravating factor at sentencing.

that those precautions were intended to yield – i.e. the cheapest cost avoider is the one who should have taken the precautions, and that could just as well be B as A (e.g. Landes and Posner 1987; Posner 1996). However, such comparisons presuppose ultimately-contestable substantive evaluations of people's different interests – i.e. reasonable people can disagree about whether my interest in safety (depreciated by the unlikelihood of me being harmed in our interaction) is more or less valuable than the resources that you would have to forego to take the aforementioned precautions. Thus, whether a person will be seen to be in breach of their role responsibilities or not, is also something that depends on the norms that we set and on the evaluations that we make.

Finally, conclusions about liability responsibility – i.e. about how someone should be treated in order to take due responsibility (or, put another way, to be properly held responsible) for what they have done – are also significantly affected by the norms that we choose. For instance, whether the appropriate kind and degree of punishment for theft is a fine, incarceration, ten lashes of the whip, twenty lashes, amputation of the hand that stole the item, or execution of the offender, is not something that can be read off straight forwardly from a consideration of what they are outcome responsible for, from the degree of their outcome responsibility, and from the degree to which they are a virtue (ir)responsible person. Rather, this too is a thoroughly norm setting exercise since we must decide for ourselves what treatment is appropriate or fitting for a given offence. Admittedly, this norm setting exercise is also not completely unconstrained – for instance, how a person should be treated on account of what they have done depends in part on what we take to be our aims in treating them in that way (e.g. retribution, deterrence, reform or even something else)[15] – and so we do not have completely free reign on what norms we may set. Furthermore, in order for a person to intelligibly be a legitimate subject of retribution, deterrence or reform, they must possess certain mental capacities – that is for instance why some people who do not oppose capital punishment never the less express qualms about executing the mentally ill (e.g. Latzer 2003; Eisenberg 2004), and it may simply be futile to attempt to deter or to reform people that lack certain mental capacities – and so this too is another constraint on how we might eventually hold someone (liability) responsible[16] – i.e. as with all of the other instances of norm setting mentioned in this section, this norm setting exercise also can not be an expression of our pure whim. However, once we do settle on a particular aim[17] – for instance, retribution – and once we confirm that the party slated for punishment possesses a sufficient measure of the right kinds of mental capacities to be a legitimate

[15] I discuss these points in greater depth in Vincent (2010:91–93).

[16] This suggests that there exists a more direct relationship between capacity responsibility and liability responsibility than what is depicted in the diagram at the top of Section 2.3. However, since this relationship is mediated via our aims, and our aims are not themselves depicted in the above diagram (the diagram after all only sets out to depict the relations that obtain between different responsibility concepts), I have therefore chosen to only describe the relationship between capacity responsibility and liability responsibility here.

[17] . . . or a collection of aims, since we might after all be pluralists in this regard . . .

subject of retribution, *we* will still have to make the difficult decision about what kind of punishment is fitting or appropriate – this is something that we will have to determine and not something that we can discover.[18] Thus, none of these constraints eliminate the fact that conclusions about liability responsibility will also be affected in important ways by the norms which we *choose* or set.

The norms that we set, the evaluations that we make, and the aims that we have also play important roles in relating different kinds of responsibility claims to one another – i.e. they also affect the justificatory relations that obtain between claims that employ different responsibility concepts. But since these relations were not depicted in the diagram at the top of Section 2.3, this sub-section therefore described some of the more prominent roles that they play in our thinking about responsibility.

2.4 The Utility of the STRC

Taken together, the six responsibility concepts and the justificatory relations that obtain between them constitute the core elements of the STRC. The STRC aims to be a "one stop shop" that catalogues, characterizes, distinguishes and relates to one another six different senses of the term "responsibility", and it also suggests some terminology to help us avoid ambiguity in debates about responsibility.

For a simple example of how the STRC helps us to avoid ambiguity, consider the text of a notice that hangs in Café Doerak, my favourite bar in the Dutch city of Delft:

> The management of this establishment is not responsible.

This notice is terribly ambiguous, and one might imagine two people engaged in a frustrating argument about it, simply because each understands it differently, though neither realizes that this is so. But the text of this notice could be helpfully re-written, replacing the generic "responsibility" with the terminology of the STRC, and then something like the following six distinct messages would emerge:

> CAPACITY: The management of this establishment are not (yet) psychologically mature.
> VIRTUE: The management of this establishment are not dependable and might be reckless.
> ROLE: The management of this establishment have no responsibilities towards its clients.
> CAUSAL: The management didn't causally contribute to losses suffered on these premises.
> OUTCOME: The management can't be blamed for whatever happens on these premises.
> LIABILITY: The management won't pay for any losses suffered on these premises.

But is this practically useful? To the extent that at least some responsibility disputes arise through misunderstanding borne out of ambiguity – i.e. because the disputants are not actually using the word "responsibility" in the same way, and hence they are not really even engaged in a genuine dispute but are instead talking

[18] Abad (Chapter 8, this volume) disagrees – she argues that only some kinds of responses (i.e. liability responsibility) are fitting or appropriate given what someone has done (i.e. given their outcome responsibility).

past one another – the STRC can be useful by helping us to avoid such ambiguity and thus to avoid such misunderstandings.

However, in what follows I will also argue that the STRC can help us to identify fifteen distinct sources of disputes about responsibility, and that it also suggests a handy procedure for systematically resolving such disputes.

2.4.1 Fifteen Sources of Disputes About Responsibility

I have argued that responsibility is not a single, unitary and generic concept, but that it is rather a syndrome of at least six different concepts. However, if this is right then there must be at least six different kinds of responsibility questions that we might ask – i.e. one for each of the STRC's responsibility concepts (in what follows, *P* stands for "person" and *O* stands for their action or an outcome of that action):

(1) Is *P* (outcome) responsible for *O*?
(2) Were *P*'s actions (causally) responsible for *O*?
(3) What were *P*'s (role) responsibilities, and were they breached?
(4) Is *P* a fully (capacity) responsible person?
(5) Is *P* a (virtue) responsible or an (virtue) irresponsible person?
(6) How should *P* be held (liability) responsible for *O*?

Thus, a responsibility dispute might arise whenever people give diverging answers to any of these six responsibility questions. Put another way, for each of the concepts in the STRC diagram shown at the top of Section 2.3, there exists an opportunity for a responsibility dispute to arise, and this gives rise to the first six possible sources of disputes about responsibility. Furthermore, given that justificatory relations obtain between these six responsibility concepts – i.e. that different kinds of responsibility claims stand in relations of justification with respect to one another – if two people give diverging answers to one responsibility question, then they may also give diverging answers to responsibility questions related to connected concepts that are depicted either below or above the corresponding concept in the STRC diagram.

However, disputes about responsibility can also arise because in a pluralistic society reasonable people may disagree about what norms we ought to adopt, about the absolute or relative value of different interests, or about our aims. As I explained in Section 2.3.4. above, norm setting, substantive evaluations and the aims that we pursue play a crucial role in several places within the STRC, and for each of these places there exist multiple opportunities for disagreements to arise.

First, in regards to capacity responsibility, just how much or what kind of mental capacity is required for "full" responsibility, is at least partly a norm-setting issue. Thus, parties might disagree about someone's status as a fully responsible person because (7) they have different views about what kinds of capacities are required for responsible moral agency, (8) they disagree about how much of that capacity a person must have to be fully responsible, or (9) there might be a factual disagreement about just how much of a given capacity that person actually has.

Second, given my account of the relationship between capacity responsibility and role responsibility – i.e. that the capacities which we (ought to) have are among the grounds of the reasons that we have to do various things, and that what we have most reason to do is what we have a role responsibility to do – a dispute might also arise because (10) disputants disagree about what reasons the possession of a particular capacity gives rise to, and thus about what role responsibility that party might have, or (11) they might believe that the person in question had a role responsibility to not jeopardize, or perhaps to develop, a particular capacity, which they failed to fulfil. Furthermore, to the extent that claims about role responsibility are subject to norm setting and substantive evaluations, people may also have diverging opinions about (12) just how much care a person must take to avoid being assessed as negligent in the event of an accident, and about (13) the value of the competing security and liberty interests that were at stake in the given situation (i.e. about the value of what the victim stood to lose and what the injurer had to forego to take sufficient care).

Finally, in regards to liability responsibility, further disagreements may also arise depending on (14) what disputants take to be the aim(s) of holding people responsible,[19] and (15) what they take to be appropriate or fitting treatment for a given transgression.

Reasonable people can disagree about the matters that are raised in points (7–15) above, and such disagreements (either tacit or explicit) can also result in a broader responsibility dispute.

2.4.2 A Procedure for Resolving Disputes About Responsibility

The previous section's observations suggest a two-step procedure for resolving disputes about responsibility. First, for a given dispute, we should *identify its subject matter*, which is another way of saying that we should ascertain which of the six responsibility questions – i.e. points (1–6) – the disputants answer differently. Second, for each of the identified disagreements, we should *examine the norms and evaluations – i.e. points (7–15) – related to the corresponding responsibility concept, and the connected concepts – i.e. points (1–6) above – depicted above that difference in the STRC diagram (as well as the related norms and evaluations)*, since that is where we will find the differences from which the dispute ultimately springs and which must therefore be reconciled to resolve that dispute. Rather than mechanically spelling out how this procedure is meant to be used, I will now proceed to the next section which provides an example of this.

2.5 The STRC in Action

This last section will demonstrate the utility of the STRC and the two-step procedure that was described above. Section 2.5.1 will demonstrate how the STRC can make it easier to discern the role which different responsibility claims play in luck

[19] . . . or, if we have plural aims, about the relative importance of our different aims . . .

egalitarian debates; and Section 2.5.2 will demonstrate how this two-step procedure can suggest argumentative strategies in courtroom debates about responsibility.

2.5.1 Luck Egalitarianism

Take two groups of people. In the first group are alcoholics who develop liver cirrhosis and need a liver transplant, heavy smokers who develop emphysema and need a lung transplant, and those with an unhealthy diet and a sedentary lifestyle who develop diabetes and need a range of ongoing medical treatments. In the second group are those whose livers were destroyed through hepatitis contracted by ingesting accidentally contaminated food, those whose lungs were destroyed through hereditary conditions like cystic fibrosis, and those who were born diabetic – and all of them also need identical transplants and medical treatments.

Because livers, lungs and the public health budget are limited resources, some people miss out on the transplants and medical treatments which they need. And many arguments have been offered both *in support of* and *against* the view that we may, or even that we ought to, give preferential treatment to people in the second group since allegedly they are not as responsible for their own ill health (they are allegedly merely unfortunate) as those in the first group (they are allegedly imprudent) – i.e. that if we must choose between two otherwise-identical candidates, one from each group, then we should favour a person from the second group. But rather than attempting to summarize this rich debate in the next few paragraphs, with undoubtedly unsatisfactory results, let me instead cite just three examples of the sorts of considerations which are sometimes raised (some better than others), and then comment on how the STRC helps us to better understand the role which they are intended to play in this debate:

(A) The addictive nature of alcohol and tobacco, and the proliferation of unhealthy eating options, as well as a lack of public parks in cities where people could exercise, are often cited to protect the first group's interests.
(B) The harshness of discriminating between people on this basis, and the fact that our interest in health is in some way different and deserving of special consideration, are also sometimes cited to support the first group's interests.
(C) But others claim that the first group's bad health is due to their own shortcomings as people – i.e. because they are weak and have bad habits like gluttony and over-indulgence – and that since such character faults should not be condoned, we may therefore give preferential treatment to the moral innocents in the second group.

Consider now where, within the STRC diagram shown at the top of Section 2.3, and within the related norm setting, substantive evaluations and aims – i.e. points (1–15) that were listed above – these different sets of considerations are meant to have an impact.

The considerations raised by example (A) address the top-left-hand quadrant of the STRC diagram. Presumably, those who cite the addictive nature of alcohol and tobacco do this to establish that alcoholics' and smokers' outcome responsibility for their own ill health is diminished because addictive substances have detrimental effects on people's capacity responsibility. On this view, people lack the capacity to stop drinking excessively once they are addicted, and thus their resultant liver cirrhosis is not truly a consequences of their own free choice – i.e. it is not something for which they are fully outcome responsible. But a different argumentative strategy is employed by those who cite the proliferation of fast food chains, and the relative difficulty of maintaining a healthy diet and getting sufficient exercise in modern crowded cities. Here, the aim is to show either that people do not in fact have as great a role responsibility to eat a healthy diet and to exercise as what we presumed, or at least that they have legitimate excuses for eating unhealthy food and getting insufficient exercise (because it is allegedly more difficult to eat a healthy diet and to exercise in modern crowded cities), or to shift at least some of the blame for this group's ill health onto another group (perhaps those who advertise and sell unhealthy food, and to urban planners who design such health-unfriendly cities).

Second, the considerations raised by example (B) map onto the middle of the bottom-most part of the STRC diagram. People who raise such considerations do not deny that the first group might be more outcome responsible for their own ill health, but rather they allege that a range of normative considerations also bears on what conclusions about this group's liability responsibility may be derived from claims about their outcome responsibility. For instance, the "harshness" objection suggests that norms of common decency rile against deriving these sorts of conclusions about how people may be treated (about their liability responsibility) from facts about what they are allegedly responsible for (about their outcome responsibility); while the other consideration draws attention to the exceptional value of people's health in an attempt to justify departing from the default rule which stipulates that distributive decisions should normally track people's outcome responsibility. The STRC thus helps us to notice that people who raise considerations like those at (B) do not necessarily take the mere fact that someone *is responsible* for their own ill health as a sufficient reason to conclude that they should therefore *take responsibility* for it, or that they should take responsibility for it *in some specific way*, since in their view a range of normative considerations (e.g. a range of those mentioned in Section 2.3.4 above) also bears on whether and how they should take this responsibility.[20]

Finally, the considerations raised by example (C) map onto the bottom-right part of the STRC diagram. People who raise such considerations attempt to stigmatize

[20] Since different responsibility concepts appear in claims about *taking* responsibility as opposed to claims about the things for which someone allegedly *is* responsible – the former are *forward*-looking while the latter are *backward*-looking – claims about the former can't be derived through mere logical entailment from claims about the latter. Put another way, the fact that someone *is* responsible for some state of affairs (e.g. their bad health) does not yet tell us how that person should take responsibility for it. Abad (Chapter 8, this volume) however disagrees with me on this point.

alcoholics, smokers and the obese, by characterizing them as lacking in moral rectitude, and they insinuate that their bad character (a legitimate target of our criticism) rather than their incapacity (something for which we may not be able to criticize them, unless perhaps via tracing they are outcome responsible for that incapacity) is the true cause of their ill health. On their account, people in the first group lack virtue responsibility rather than having diminished capacity responsibility – they view the alcoholic, the smoker and the obese person as someone with character flaws not as someone with capacity deficits – and they maintain that on account of their character flaws we may legitimately treat them in a less privileged manner.

Viewed through the lens of the STRC, these three sets of considerations play very different roles in this debate. The first set of considerations informs a debate about whether people in the first group are more outcome responsible for their own ill health than people in the second group. The second set of considerations informs a debate about whether we might be able to block the transition from claims about outcome responsibility to claims about liability responsibility. While the third set of considerations informs a debate about whether the correct way to characterize people in the first group is that they have capacity deficits or character flaws, and whether the latter characterization could justify treating these people in a less privileged way. The STRC is, among other things, a theoretical framework, and viewing the debate about luck egalitarianism through the lens of this framework can make it easier to discern precisely how the many different claims that are made in such debates are supposed to bear on the broader issues.

Thus, an important reason why the STRC is useful is because it can help us to navigate our way around such complex debates, but also because it can help us to introduce these debates to novices (e.g. students) in a more structured manner.

2.5.2 Law Suits

Consider now another example, this time aimed at demonstrating the utility of the two-step procedure outlined above:

> You drive to a popular shopping centre but the car park is very busy. After driving around for what seems like ages you finally find an empty spot, but a Porsche parked in the next spot over is straddling the dividing line and partially hanging over into your spot. Resolute to get your shopping done, you carefully maneuver your car into the empty spot, but alas the fit is so tight that you can't even open your door to get out. Annoyed at the Porsche driver's inconsiderateness, you reverse out just a dash too quickly and leave a dinted and scratched Porsche in your wake. Distraught by the prospect of a hefty repair bill you flee the scene, though an anonymous witness jots down your number plate and leaves a note for the Porsche owner, and you are now in court being sued for the cost of repairs.

Suppose that you have no scruples, and that you are quite prepared to lie to avoid being landed with the hefty bill for repairs. Given your aim of avoiding liability responsibility, which is located at the bottom of the STRC diagram, a number of defense strategies are encountered as we work our way up this diagram's branches.

First, traveling one level up the left branch of the STRC diagram, we reach the suggestion that you could deny your outcome responsibility for the damage to the Porsche. After all, if you were not outcome responsible for it, then that would be an effective defense to the claim that you should now accept liability responsibility for it. But how might you support your denial of outcome responsibility?

Second, traveling another level up the left branch of the STRC diagram, we reach the suggestion that you could say "It wasn't me!" – i.e. you could deny your causal responsibility for the damage. If you were not causally responsible for that damage then that would undermine the claim that you were outcome responsible for it, and there are at least two ways of denying your causal responsibility. On the one hand, you could deny that you ever even acted like that – i.e. that you maneuvered your car into that parking spot. On the other hand, you could admit that you acted like that, but then deny that in the process you dinted and scratched the Porsche. In both cases you would be pitching your word against the witness' word, and presumably you have more of an incentive to lie than the witness does, and perhaps your lies might come undone if the security cameras in the car park are suitably positioned to capture footage of you dinting and scratching the Porsche. But never the less, this is still an argumentative strategy that is open to you.

Third, instead of denying your causal responsibility, you could also travel up the right branch of the STRC diagram from outcome responsibility, and try to avoid at least some of the outcome responsibility for those dints and scratches by shifting some of the blame onto others – i.e. by arguing that others also violated their role responsibility. For instance, you could try to shift some of the blame onto the Porsche owner by pointing out that had they not parked their car in such an *inconsiderate way* – in a blameworthy way – then the accident might not have occurred. Alternatively, you could argue that the Porsche owner is outcome responsible for the *excessive portion* of the cost of repairs, because it was after all their folly to put such an extravagantly expensive car in a dangerous public place. Finally, you might even try to shift some of the blame onto the shopping centre management, by arguing that they are at least partially at fault since they made the car parking spaces *too tight*. The broad aim of this argumentative strategy would be to show that although your faulty actions might indeed be among the causes of the dints and scratches, so too are others' faulty actions – i.e. the Porsche owner's inconsiderateness, their folly in leaving such an expensive and fragile item in a dangerous public place, and the shopping centre's stingyness with space – and hence that those others are at least partially outcome responsible (and should thus bear a proportionate amount of liability responsibility) for the accident.

Admittedly, it is not clear that any of these arguments would ultimately work. But facts differ from case to case, and it is plausible that in other cases a judge might rule in your favour. For instance, if the reason why the cost of repairs to the other car was so great was because its owner paid millions to Pablo Picasso to paint their last masterpiece on the duco of their car (and the masterpiece is now scratched), then the judge might indeed make the substantive evaluation that this is pure folly on their part and award them damages consistent only with how much the repairs to a reasonably-priced car would have cost.

The point of this example is not that any of these strategies would necessarily work, but it is rather to demonstrate how the two step procedure outlined above can provide hints about the sort of arguments that you might be able to use in a debate about responsibility.

2.6 Conclusion

As I suggested in the opening paragraph, responsibility is often talked about as if it were a single, unitary and generic concept. However, in this paper I have argued that there are at least six different responsibility concepts, that certain justificatory relations obtain between them, and that knowing about these distinctions and relations is not only interesting but also useful. But although Section 2.5.1 explained how this understanding of responsibility sheds new light on the luck egalitarian debate, and elsewhere (Vincent 2010) I have written about how this understanding illuminates the neurolaw debate which was also cited above in the opening paragraph,[21] I have not yet said anything about how compatibilist responsibility theory might be illuminated by it. Thus, in closing, I will now briefly comment on how the understanding of responsibility developed in this paper bears on compatibilist responsibility theory.

Philosophical compatibilism aims to reconcile moral responsibility with determinism – i.e. to show how responsibility might still be possible even if everything that happens, including everything that we do, is completely caused by prior events. For instance, on John Martin Fischer and Mark Ravizza's account, whether a person is responsible for what they do depends not on whether they could have done otherwise – i.e. freedom from causation is not needed for moral responsibility – but on whether their actions issued from their own moderately reasons-responsive mechanisms.[22] However, if moral responsibility is indeed not a single thing but it is rather a collection of at least six different things, then a question that should now be asked is how compatibilism reconciles each of these six different kinds of responsibility with determinism? For instance, how is determinism meant to pose a threat to the idea that some people have capacity responsibility while others do not, or that some people are responsible in the virtue sense while others are irresponsible? Alternatively, what light might compatibilism shed on what role responsibilities people can be legitimately expected to discharge, or on how people should be held liability responsible for what they do? Such questions have not been explicitly addressed in the compatibilist literature because that literature has all too often treated responsibility as a single thing, but if responsibility is indeed more of a syndrome of things than a single thing then attention should now be paid to these different questions.

---

[21] In that paper I argue that rather than asking whether *neuroscience* is relevant to *responsibility*, we should ask how the *range of different neuroscientific techniques and technologies* can help us to address the *six different kinds of responsibility questions* that we might ask – i.e. the first six questions listed at the top of Section 2.4.1. above.

[22] It also depends on a range of historical factors discussed by Fischer and Ravizza under the heading of the "tracing approach" (1998:48–51).

In a recent paper, John Fischer and Neal Tognazzini argue that many of the long-standing responsibility disputes in philosophy are plausibly due to the fact that the disputants actually have different things in mind when they utter the word "responsibility", and hence they suggest that:

> Clarity now demands that compatibilists and incompatibilists about determinism and moral responsibility specify which face of moral responsibility is at issue in their theses[, since s]omeone may well think that determinism is compatible with [one sense of "responsibility"] even though it is not compatible with [another sense]. (2011:398)

What compatibilists must now do, given that "responsibility" refers to a range of different ideas, is to explain in what way determinism is meant to pose a threat to each of these different ideas, and to spell out how compatibilism addresses these different challenges. To the extent that in ordinary discourse the term "responsibility" refers to a range of different concepts, this would be useful since it would explain how compatibilist responsibility theory can be applied in a range of practical contexts that use the term "responsibility" in these different ways – i.e. to help us resolve the sorts of problems that are encountered in public policy, in courtrooms and in other non-philosophical contexts.

References

Anderson, Elizabeth. 1999. "What Is the Point of Equality?" *Ethics* 10(2):287–337.

Arneson, Richard J. 2000. "Luck Egalitarianism and Prioritarianism." *Ethics* 110(2):339–49.

Bovens, Mark. 1998. "Two Concepts of Responsibility." In *The Quest for Responsibility: Accountability and Citizenship in Complex Organisations*, 22–42. Cambridge: Cambridge University Press.

Cane, Peter. 2004. "Responsibility in Law and Morality: Book Symposium, Author's Introduction." *Australian Journal of Legal Philosophy* 29:160–63.

Cohen, G.A. 1989. "On the Currency of Egalitarian Justice." *Ethics* 99(4):906–44.

Dawkins, Richard. 2006. "Let's All Stop Beating Basil's Car." Accessed February 12, 2007. http://www.edge.org/q2006/q06_9.html#dawkins.

Dennett, Daniel C. 2003. *Freedom Evolves*. New York, NY: Viking.

Duff, R.A. 1998. "Responsibility." In *Routledge Encyclopedia of Philosophy*, edited by Edward Craig, Vol. 9, 290–94. New York, NY: Routledge.

Dworkin, Ronald. 1981. "What Is Equality? Part 2: Equality of Resources." *Philosophy & Public Affairs* 10(4):283–345.

Eisenberg, Leah. 2004. "Medicating Death Row Inmates so They Qualify for Execution." *Virtual Mentor* 6(9).

Feinberg, Joel. 1970. "Sua Culpa." In *Doing & Deserving: Essays in the Theory of Responsibility*, edited by Joel Feinberg, 187–221. Princeton, NJ: Princeton University Press.

Fischer, John Martin and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Fischer, John Martin and Neal A. Tognazzini. 2011. "The Physiognomy of Responsibility." *Philosophy and Phenomenological Research* 82(2):381–417.

Gazzaniga, Micheal. S. 2006. "Facts, Fictions and the Future of Neuroethics." In *Neuroethics: Defining the Issues inTheory, Practice, and Policy*, edited by Judy Illes, 141–48. Oxford: Oxford University Press.

Goodin, Robert E. 1986. "Responsibilities." *Philosophical Quarterly* 36:50–56.

Goodin, Robert E. 1987. "Apportioning Responsibilities." *Law and Philosophy* 6:167–85.

Greene, Joshua and Jonathan Cohen. 2004. "For the Law, Neuroscience Changes Nothing and Everything." In *Law & the Brain*, edited by Semir Zeki and Oliver Goodenough, 207–26. New York, NY: Oxford University Press.

Hart, H.L.A. 1968. "IX. Postscript: Responsibility and Retribution." In *Punishment and Responsibility*, 210–37. Oxford: Clarendon Press.

Haydon, Graham. 1978. "On Being Responsible." *The Philosophical Quarterly* 28(110):46–57.

Honoré, Tony. 1999. "4. The Morality of Tort Law: Questions and Answers." In *Responsibility and Fault*, edited by T. Honoré, 67–93. Portland, OR: Hart Publishing.

Kutz, Christopher. 2004. "Chapter 14: Responsibility." In *Jurisprudence and Philosophy of Law*, edited by Jules Coleman and S. Shapiro, 548–87. Oxford, UK: Oxford University Press.

Lacey, Nicola. 2007. "Space, Time and Function: Intersecting Principles of Responsibility Across the Terrain of Criminal Justice." *Criminal Law and Philosophy* 1(3):233–50.

Landes, William M., and Richard A. Posner 1987. *The Economic Structure of Tort Law*. Cambridge, MA: Harvard University Press.

Latzer, Barry. 2003. "Between Madness and Death: The Medicate-to-Execute Controversy." *Criminal Justice Ethics* 22(2):3–14.

Morse, S.J. 2006. "Moral and Legal Responsibility and the New Neuroscience." In *Neuroethics: Defining the Issues in Theory, Practice, and Policy*, edited by Judy Illes, 33–50. Oxford, UK: Oxford University Press.

Pereboom, Derk. 2001. *Living Without Free Will*. Cambridge: Cambridge University Press.

Perry, Stephen R. 2000. "Loss, Agency, and Responsibility for Outcomes: Three Conceptions of Corrective Justice." In *Philosophy of Law* (6th Edition), edited by Joel Feinberg and Jules Coleman, 546–59. Belmont, CA: Wadsworth/Thompson Learning.

Posner, Richard A. 1996. "Lecture Two: The Common Law." In *Law and Legal Theory in England and America*, edited by Richard A. Posner, 39–67. Oxford: Oxford University Press.

Rakowski, Eric. 1991. *Equal Justice*. New York, NY: Oxford University Press.

Scanlon, Thomas. 1998. "Chapter 6: Responsibility." In *What We Owe to Each Other*, 248–94. Cambridge, MA: The Belknap Press of Harvard University Press.

Vincent, Nicole. 2009. "Responsibility: Distinguishing Virtue from Capacity." *Polish Journal of Philosophy* 3(1):111–26.

Vincent, Nicole. 2010. "On the Relevance of Neuroscience to Criminal Responsibility." *Criminal Law and Philosophy* 4(1):77–98.

Vincent, Nicole. 2011. "Madness, Badness and Neuroimaging-Based Responsibility Assessments." In *Law and Neuroscience, Current Legal Issues, Volume 13*, edited by Michael Freeman, 79–95. Oxford: Oxford University Press.

Vincent, Nicole. Forthcoming. "Enhancing Responsibility." In *Legal Responsibility and Neuroscience*, edited by N. Vincent. Oxford: Oxford University Press.

Vranas, Peter B.M. 2007. "I Ought, Therefore I Can." *Philosophical Studies* 136(2):167–216.

Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

Wallace, R. Jay. 2002. "Précis of Responsibility and the Moral Sentiments." *Philosophy and Phenomenological Research* 64(3):680–81.

Watson, Gary. 2004. "Two Faces of Responsibility." In *Agency and Answerability*, edited by Gary Watson, 260–88. Oxford: Oxford University Press.

Williams, Bernard. 1995. "Voluntary Acts and Responsible Agents." In *Making Sense of Humanity*, edited by Bernard Williams, 22–34. Cambridge, UK: Cambridge University Press.

Williams, Garrath. 2008. "Responsibility as a Virtue." *Ethical Theory and Moral Practice* 11(4):455–70.

Chapter 3
The Relation Between Forward-Looking and Backward-Looking Responsibility

Ibo van de Poel

Abstract This contribution discusses the relation between forward-looking and backward-looking responsibility. The notion of forward-looking responsibility I focus on is, following Robert Goodin, that of seeing to it that a certain state of affairs obtains. In addition, I focus on two types of backward-looking responsibility: accountability and blameworthiness. I argue that accountability only entails blameworthiness if the agent cannot cite certain reasonable excuses like ignorance and compulsion (which were already mentioned by Aristotle). It is further argued that accountability can both be based on not properly discharging a forward-looking responsibility and on the breach of a duty that caused a certain negative consequence. I show that in both cases three general conditions need to apply in order to hold an agent reasonably accountable: a capacity condition, a causality condition and a wrongdoing condition. The exact content of these conditions, however, depends on whether accountability is based on a forward-looking responsibility that is not properly discharged or on the breach of a duty.

3.1 Introduction

Traditionally much of the philosophical literature on responsibility has focused on backward-looking responsibility. The main question has been: when, under what conditions, is it proper to blame someone for a certain action or a certain outcome? More recently, a number of authors have discussed forward-looking notions of responsibility, either on consequentialist grounds (e.g. Goodin 1995) or on the basis of virtue ethics or care ethics (e.g. Ladd 1991; Williams 2008). In this contribution, I will explore the relations between forward- and backward-looking notions of responsibility. My focus will be on notions of *moral* responsibility, i.e. responsibility that is grounded in moral considerations, rather than legal or organizational notions of responsibility. Moreover, I will focus on responsibility for consequences (states of affairs) rather than responsibility for actions, although I will understand the notion of "consequence" broadly, as also encompassing for example a person possessing certain virtues.

I. van de Poel (✉)
Department of Philosophy, Delft University of Technology, 2600 GA Delft, The Netherlands
e-mail: I.R.vandePoel@tudelft.nl

N.A Vincent et al. (eds.), *Moral Responsibility*, Library of Ethics and Applied Philosophy 27, DOI 10.1007/978-94-007-1878-4_3,

My approach in this paper is largely conceptual in nature, i.e. I will proceed by conceptually distinguishing and clarifying different notions of responsibility and their relations. In doing so, I will build on accounts of responsibility that are offered by other thinkers and as they can be found in daily language. At some points, I will make choices with respect to how I understand the relevant terms and their relations that are not strictly conceptual but imply a substantial normative choice. Such choices are at some points inevitable if one wants to sketch a coherent picture. Nevertheless, I have tried to provide an account that is general and abstract enough to accommodate different substantive notions of responsibility.

This paper is structured as follows. I start with an overview of different meanings of responsibility, suggesting five major normative notions of responsibility, three of which are primarily backward-looking (accountability, blameworthiness and liability), and two which are primarily forward-looking (responsibility as virtue and as the moral obligation to see to it that something is the case). I will then restrict my analysis to three of the notions: accountability, blameworthiness and the moral obligation to see to it that something is the case. After suggesting a relation between these three that makes sense prima facie, I will add further complexities.

3.2 Notions of Responsibility

The term "responsibility" comes in different meanings and senses. It is therefore useful to distinguish some of the main meanings so that we know what we are talking about. We might to this end distinguish the following meanings of responsibility[1]:

1. *Responsibility-as-cause*. As in: the earth quake is responsible for the death of 100 people.
2. *Responsibility-as-task*.[2] As in: the train driver is responsible for driving the train.
3. *Responsibility-as-authority*.[3] As in he is responsible for the project, meaning he is in charge of the project.

[1] Hart (1968:210–37) was probably the first to distinguish four main senses of responsibility: role-responsibility, causal-responsibility, liability-responsibility and capacity-responsibility. The additional senses I distinguish in addition are related to these, but have in certain respects an (importantly) different meaning as explained. All the additional senses can indeed also be found in the literature on responsibility (e.g. Casey 1971; Baier 1972; Ladd 1982; Zimmerman 1988; Lucas 1993; Bovens 1998; Cane 2002; Duff 2007; Williams 2008; Davis forthcoming). Hart discusses blameworthiness as a component of moral liability, but I think that it is conceptually clearer to distinguish both notions.

[2] This is what Hart calls role-responsibility.

[3] This may also be called responsibility-as-office or responsibility-as-jurisdiction. It refers to a realm in which one has the authority to make decisions or is in charge and for which one can be held accountable.

4. *Responsibility-as-capacity*, i.e. as the ability to act in a responsible way. This includes for example the ability to reflect on the consequences of one's actions, to form intentions, to deliberately choose an action and act upon it.
5. *Responsibility-as-virtue*, i.e. as the disposition (character trait) to act responsibly. As in: he is a responsible person.
6. *Responsibility-as-(moral)-obligation*, to see to it that something is the case. As in: he is responsible for the safety of the passengers, meaning he is responsible to see to it that the passengers are transported safely.
7. *Responsibility-as-accountability*, i.e. as the (moral) obligation to account for what you did or what happened (and your role in it happening).
8. *Responsibility-as-blameworthiness*. As in: he is responsible for the car accident, meaning he can be blamed for the car accident happening.
9. *Responsibility-as-liability*. As in: he is liable to pay damages.

The first four meanings are more or less descriptive: responsibility-as-cause, role, authority and capacity describe something that is the case or not. The other five are normative. They imply a normative evaluation, as in responsibility-as-virtue and responsibility-as-blameworthiness, or a prescription, as in responsibility-as-obligation to see to it that something is the case and in responsibility-as-liability (to pay damages, offer excuses, put the situation right, etcetera). Responsibility-as-accountability seems to imply both an evaluation as well as a prescription; the agent is supposed to account for something because an action or outcome can be laid at her feet.[4]

Some of the normative meanings are related to, or rely on the descriptive meanings of responsibility. Responsibility-as-virtue is closely related to responsibility-as-capacity. But whereas the latter only refers to the ability to act responsibly, responsibility-as-virtue refers to the actual disposition, also surfacing in actions, to be a responsible person. Similarly responsibility-as-obligation to see to it that something is the case is closely related to responsibility-as-task.[5] Tasks are typically often formulated in terms of seeing to it that something is the case. The difference is that not every task or role defines a moral obligation. So, while it might be said that Nazi engineer Eichmann had the task (responsibility) that the Jews were effectively transported to the concentration camps, it does not follow that he had a (moral) obligation to see to it that they were effectively transported. In fact, since the transport was part of an immoral plan, aiming at the extinction of the Jews, he might even have had a moral obligation to see to it that they were not effectively transported.

Although responsibility-as-accountability, responsibility-as-blameworthiness and responsibility-as-liability are in meaning not directly related to one of

[4] Sometimes responsibility-as-accountability may be understood in a descriptive sense, as in cases in which one is accountable on the basis of certain organizational or legal rules. In such cases, responsibility-as-accountability seems often closely related to responsibility-as-task.

[5] Goodin (1995) in fact calls such responsibilities task-responsibilities, but – as pointed out in the text – I think there is an essential difference between task-responsibility in the sense I use the term and a (moral) obligation to see to something.

the descriptive notions, it is often assumed that responsibility-as-cause and responsibility-as-capacity are preconditions for holding someone accountable or liable or for blaming someone. Blameworthiness, in turn, is sometimes seen as a condition for moral liability. It is also often believed that responsibility-as-task or responsibility-as-authority may lead to responsibility-as-accountability, especially if the former responsibilities are not properly discharged.

The first two normative meanings are primarily forward-looking (prospective) in nature. This is most obvious for responsibility as the obligation to see to it that something is the case; it relates to something that is, usually, not yet the case. Responsibility-as-virtue is often primarily understood as being forward-looking (e.g. Ladd 1991; Bovens 1998); it relates to responsibilities an agent actively assumes and to a certain attitude rather than to blame (or praise). Nevertheless, one could well argue that a responsible person is one who is willing to account for his actions and who accepts blame and liability where that is due (Williams 2008).

Responsibility-as-accountability, blameworthiness and liability are backward-looking in the sense that they usually apply to something that has occurred. Nevertheless accountability and liability have a forward-looking (prescriptive) element in the sense that the agent is supposed to do something (in the future): to account for his actions, to pay damages and the like.

Now that we have distinguished different meanings of responsibility, it is possible to be a bit more precise about the main question of this paper, i.e. the relation between forward-looking and backward-looking responsibility. I am mainly interested in the relation between the various normative notions of responsibility. As we have seen two of these notions are primarily forward-looking, i.e. responsibility-as-virtue and as the obligation to see to it, and three are primarily backward-looking, i.e. accountability, blameworthiness and liability. Below, I will focus on the forward-looking notion of having to see to it that something is the case. This notion is more directly related to concrete actions and state of affairs than responsibility-as-virtue and therefore more obviously related to the backward-looking notions of responsibility. When I talk below of forward-looking responsibility I mean to refer to responsibility as the (moral) obligation to see to it that something is the case rather than to the virtue notion unless stated otherwise. With respect to the backward-looking notions I will focus on accountability and blameworthiness. Before saying more on the relation between these three notions, it is useful to say a bit more on the relational nature of the concept of responsibility.

3.3 Responsibility as a Relational Concept

In most of its meanings, the notion of responsibility refers to a relation between at least two entities. The most basic form this relation takes is:

(1) *A is responsible for X*

In which A is some agent and X can refer to actions, state of affairs (outcomes), tasks or realms of authority. Two meanings of responsibility, however, seem to

resist this conceptualization, namely responsibility-as-capacity and responsibility-as-virtue.[6] Capacity and virtue are better understood as a "property" or characteristic of the agent A rather than as a relation between an agent A and some X. This is not to say that it never makes sense to particularize responsibility-as-capacity or responsibility-as-virtue to a particular X. We might, for example, say that someone is a responsible parent but an irresponsible engineer (both in the virtue sense). However, for both responsibility-as-capacity and responsibility-as-virtue it makes sense to say that A is responsible full stop, while that appears impossible for all other meanings.

Responsibility can also be understood as a triadic relational concept. Duff (2007:23–30) argues that normative notions of responsibility are best understood according to the following scheme[7]:

(2) *A is responsible for X to B*

In which B is some agent, usually different from A. In cases of forward-looking responsibility, (2) reflects the fact that we may have specific responsibilities to different people. Professionals like engineers, for example, have different responsibilities to their employer, to their colleagues, to their clients and to the public.[8] What they owe to their clients is different from what they owe to their employer or to the public; their responsibilities to these different groups of agents may even conflict.

More generally (2) may be seen as a reflection of the fact that forward-looking responsibilities may arise from the specific relations we have with specific people (cf. Scheffler 1997). This is not to deny that we may also have responsibilities to ourselves or general responsibilities. These may be seen as special cases in which B = A, or in which B is humanity (or morality or God if one wishes). In the case of forward-looking responsibility (2) might then be understood as follows:

(3) *A is forward-looking responsible for X to B means that A owes it to B to see to it that X*

How is (2) to be understood for backward-looking notions of responsibility like accountability and blameworthiness? Duff (2007:23) suggests that B is "a person or body who has the standing to call me to answer for X." More generally, B may be any agent who can fittingly, i.e. fairly or reasonably, hold A responsible for X. This may be expressed as follows:

---

[6] According to Duff (2007:23), responsibility-as-capacity can be explained in relational terms as "the capacities that are necessary if one is to answer for one's actions." However, this conflates the conceptual nature of responsibility-as-capacity with its being a precondition for other relational concepts of responsibility.

[7] In fact, responsibility-as-task is also usually understood as a triadic relation.

[8] In fact, we may also have different responsibilities in the different roles we have, for example as teacher, as colleague and as parent, and also these responsibilities may conflict.

(4) *A is backward-looking responsible for X to B means that it is fitting for B to hold A responsible for X*

Holding responsible here includes holding accountable or blaming. Duff suggests that the ones to which I owe something (the agent B in (3)) is the same as the one for whom it would be appropriate to hold me backward-looking responsible (the agent B in (4)). This suggestion, however, seems to me false. I may have a forward-looking responsibility to future generations to limit my emissions of greenhouse gases with an eye to global warming, but it does not follow that they are the only ones who can hold me accountable (or blameworthy) for not limiting my emissions of greenhouse gases. In fact they may never be able to hold me accountable because future generations do not exist yet, and when they (hopefully) exist in the future I may no longer exist. There may nevertheless be others, including I myself, who can properly hold me accountable for my emissions of greenhouse gases.

It has indeed been suggested that in cases of moral backward-looking responsibility, in contrast to legal, organizational or social responsibility, it is in principle appropriate for anyone to hold me accountable (or blameworthy) under certain conditions. The reference to B in other words is superfluous for moral responsibility. In fact, Strawson's conceptualization of backward-looking responsibility as the fittingness of certain reactive attitudes can be understood along the following lines (Zimmerman 2009; Strawson 1962):

(5) *A is responsible for X means it is fitting to adopt some reactive attitude toward A in respect of X*

Or in a formulation that closely resembles a proposal by Wallace (1994:92):

(6) *A is responsible for X if and only if it is fitting to hold A responsible for X*

Formulations (5) and (6) suggest that the fittingness of reactive attitudes is independent from the specific agent B, so that the reference to B becomes superfluous. An interesting criticism of formulations like (5) and (6) is offered by Kutz (2000: 17–65). He admits that in cases of moral responsibility, it might be appropriate for anyone to hold me accountable or to express certain reactive attitudes. However, what reactive attitudes are appropriate or under what conditions it is fitting to hold some agent A responsible may well depend on the specific relation between A and B, so that the reference to B is not superfluous.

A specific case is the situation when B = A. Consider the following example.[9] During a departmental meeting I make a statement or an argument that turns out

[9] This is my example. Kutz provides other examples, including examples in which it seems appropriate for the agent A to assume responsibility rather than, as in my example, there being a degree of freedom in taking responsibility or not. My suggestion is then not that the phenomenon of taking responsibility, to which I draw attention below, exhausts the relational nature of responsibility. Rather it draws attention to an aspect of responsibility that is also not fully grasped by formulations like (4).

to be insulting for one of the other participants in the meeting. Let's suppose that my statement cannot in general be considered insulting or inappropriate nor could I have known that my words would be insulting. Because of this excusable ignorance, it seems inappropriate for other participants in the meeting to hold me responsible (blameworthy) for what turned out to be an insult. For the same reasons I am morally allowed not to blame myself nor should I feel guilty. Nevertheless, it would not be inappropriate if I felt guilty and would offer her excuses. In fact doing so might, depending on the exact circumstances, be laudable or virtuous. This example suggests two things. First, I can appropriately take responsibility for some X even if it would be inappropriate for others to hold me responsible for that X. Second, I have a degree of choice in taking responsibility. Of course, I cannot reasonably take responsibility for everything nor can I reasonably escape responsibility for some things. However, within the bounds of reason and morality, I have some freedom for taking responsibility for certain things or not.

The possibility of taking responsibility seems to undermine formulations (5) and (6). In fact, it is also not fully captured by (4) because, as the example suggests, there may be situations in which it would be both appropriate, i.e. rationally and morally allowable, not to take responsibility and to take responsibility. Whether an agent is, or rather becomes, responsible in such situations depends on the volitional choice of that agent.

Although the phenomenon of taking responsibility is interesting and important, I will neglect it in the rest of this paper. I will focus below on conditions under which it is appropriate to hold A backward-looking responsible (in the different senses of the term) independent from characteristics of B or the relation between A and B. The reason is that I am interested here mainly in the relation between forward-looking and backward-looking responsibility and that I think that it is better not to burden a first-order characterization of that relation with having to account for the relation between A and B as well. Another way of characterizing the account that I will offer is to say that the account that I shall offer will presuppose a standard or default relationship between A and B (cf. Kutz 2000:30) in which A and B are both members of the moral community while abstracting from any further specific relations between A and B, so that the reference to B becomes superfluous because B stands for any member of the moral community.

3.4 The Relation Between Forward-Looking and Backward-Looking Responsibility: A Suggestion

Let me now turn to the main question of this paper: the relation between forward-looking and backward-looking responsibility. The previous sections make it possible to be more precise about what I mean with that. I will be focusing on accountability and blameworthiness as backward-looking notions of responsibility and on the forward-looking notion as the obligation to see to it that. In all these cases the focus will be on responsibility for consequences. I will further understand accountability and blameworthiness in line with (6). This means that I will be focusing on conditions under which it is fitting (appropriate, reasonable) to hold

an agent A responsible (accountable, blameworthy) for an outcome X. I will neither pay attention to agents taking accountability or blame nor to the relation between the agent being held responsible (agent A) and the agent holding responsible (agent B). With respect to forward-looking responsibility I will assume that agents (can) have such responsibilities, whatever their exact source or ground, and will inquire what that means for their backward-looking responsibilities.

For a first order answer to the question of the relation between forward-looking and backward-looking responsibility, I turn to a real-life example. The answer is a brief account that Leslie Robertson, the structural engineer who designed the Twin Towers in New York, provided when he was asked whether he felt guilty about the fact that the towers did not stand longer after they had been hit by two planes in a terrorist attack on September, 11 2002:

> The responsibility for arriving at the ultimate strength of the towers was mine. The fact that they did not stand longer could be laid at my feet. Do I feel guilty about . . . the fact that they collapsed? The circumstances on September, 11 were outside of what we considered in the design. . . .. If I knew then what I know now they would have stood longer, of course.[10]

I think that the different elements of his account can be interpreted as follows:

- "The responsibility for arriving at the ultimate strength of the towers was mine." This refers to responsibility as the obligation to see to it that.
- "The fact that they did not stand longer could be laid at my feet." This can be interpreted as: It would be proper (for others) to hold me accountable for that fact.
- "Do I feel guilty about . . . the fact that they collapsed?" This could be interpreted as: "Would it be proper to blame me for it?"
- "The circumstances on September, 11 were outside of what we considered in the design. . . .. If I knew then what I know now they would have stood longer, of course." This can be interpreted as giving an account (invoking certain excuses like nonculpable ignorance) that shows why it would not be proper to blame him.[11]

This example then suggests a first order answer to my question along the following lines:

> If A has the obligation to see to it that some state of affairs X is the case and X happens not to be the case, A is accountable for "not X". If A is not able to give a satisfactory account for "not X", A is also blameworthy.

---

[10] Excerpt from documentary "Why the Twin Towers collapsed" broadcasted by Discovery Channel.

[11] It is very interesting and indeed impressive that on the video of the interview Robertson starts nodding when he poses himself the question "Do I feel guilty?" One interpretation would be that in his non-verbal expressions (the nodding) he answers the question "Do I take the blame?" (He obviously feels very bad about what happened if not guilty), while in his verbal expressions he answers the question "Would it be proper for others to blame me?" His answer to the first question seems affirmative and to the second not.

Below, I will further explore this account, try to make it more precise and adjust it where that might turn out to be necessary. In doing so, I will start at the end of the chain and then work backwards from there.

3.5 Blameworthiness

As indicated earlier, much of the traditional literature on responsibility focuses on responsibility-as-blameworthiness. A range of authors have suggested a set of conditions that need to be satisfied in order for an agent A to be blameworthy for a state of affairs X (e.g. Feinberg 1970; Thompson 1980; Wallace 1994; Bovens 1998; Fischer and Ravizza 1998; Corlett 2001; Swierstra and Jelsma 2006). Conditions that are often mentioned include[12]:

- *Moral Agency*: the agent A is a moral agent, i.e. has the capacity to act responsibly. (responsibility-as-capacity)
- *Causality*: the agent A is somehow causally involved in X. This can be because A causally contributed to X. (responsibility-as-cause)
- *Wrongdoing*: The agent A did something wrong. (On some accounts, the occurrence of X *simpliciter* may constitute the wrongdoing).
- *Freedom*: The agent A was not compelled to bring about X.
- *Knowledge*: A knew, or at least could have known, that X would occur and that this was undesirable.

Although these general conditions can be found in many accounts, there is much debate about at least two issues. One is the exact content and formulation of each of the conditions. For example, does the freedom condition imply that the agent could have acted otherwise? The other is whether these conditions are individually necessary and together sufficient in order for an agent A to be blameworthy for X. One way to deal with the latter issue is to conceive of the mentioned conditions as arguments or reasons for holding someone responsible (blameworthy) for something rather than as a strict set of conditions (Davis forthcoming).

Despite these debates, the above list of conditions is helpful to distinguish between blameworthiness and accountability. My suggestion is that some of the above mentioned conditions are possible excuses (reasons) that can be used by an agent that is accountable for something in order to show that she is not blameworthy.[13] Others are rather preconditions for being accountable.

---

[12] Also control is sometimes mentioned as condition, but see Sher (2006).

[13] A similar suggestion can be found in Hart (1968), Wallace (1994) and Duff (2007). Although Hart and Duff do not distinguish between blameworthiness and liability, they suggest a similar relation between accountability (or answerability) and liability as I do between accountability and blameworthiness. Wallace makes a distinction between A- and B-conditions for responsibility: "B-conditions make it fair to hold people morally to blame . . . while A-conditions make it fair to hold people morally accountability" (Wallace 1994:118) His A-conditions focus on when it is in general

The freedom and knowledge condition are both arguments that can be used in an account to excuse oneself.[14] Typically, both are already mentioned by Aristotle as possible reasons why someone is not to be blamed for her actions or the consequences of these (Aristotle, *The Nicomachean Ethics*, book III, chapters 1–5).[15] The capacity and causality condition on the other hand are typically conditions for holding someone accountable. We do not hold people, or other entities, accountable if we do not have reason to believe that they have the capacity to act responsibly[16]; without this capacity they would in fact not be able to provide an account at all, or so it seems. We also do not hold people accountable if we believe that they are completely causally disconnected from the state of affairs X we are concerned about. There should be at least a suspicion of causal involvement or the ability to causally influence the outcome X by the agent A. Although capacity and causality are conditions for holding someone accountable, they may also function as arguments in the account given. We might suspect a causal connection, but the agent might be able to show in her account that we are wrong. Similarly, the agent might argue that she temporarily lacked, for circumstances beyond her control, the capacity to act responsibly and therefore is not to be blamed.

How does the wrongdoing condition fit into this picture? In general it seems that we hold people not only accountable for bad things, but also for neutral things, like a reimbursement, and even for good outcomes, for example to judge whether a certain price would be deserved. What is common to these cases of accountability is that a judgment is made whether a certain treatment is deserved. In the case, we are interested in here, the question is whether blame is deserved. Such accountability for blame is sometimes called answerability. In the words of Hart:

> The original meaning of the word "answer", like that of the Greek "αποχρινέσθαι" and the Latin *respondere*, was not that of answering questions, but that of answering or rebutting accusations or charges, which, if established, carried liability to punishment or blame or other adverse treatment (see O.E.D., *sub. tit. "answer"*). . . . a person who fails to rebut a charge is liable to punishment or blame for what he has done, and a person who is liable to punishment or blame has had a charge to rebut and failed to rebut. (Hart 1968:265)

In as far as wrongdoing is a condition for blameworthiness, at least a reasonable suspicion of wrongdoing is a precondition for reasonably holding someone

---

fair to hold people accountable (cf. Wallace 1994:154), this is my first condition (moral agency); his account seems to assume wrongdoing implicitly (e.g. Wallace 1994:156). My conditions for accountability also include conditions for when it is fair to hold someone accountable for a *specific outcome*.

[14] Wallace (1994:136–47) mentions four types of excuses: (1) inadvertence, mistake or accident, (2) unintentional bodily movements, (3) physical constraint and (4) coercion, necessity and duress. The first is a case of non-culpable ignorance (referring to the knowledge condition), the others of compulsion (referring to the freedom condition).

[15] Typically many other authors have treated those two conditions as the conditions for being at fault, suggesting that these are conditions for blameworthiness rather than for accountability.

[16] This capacity might be understood in terms of reason-responsiveness (Fischer and Ravizza 1998) or reflective self-control (Wallace 1994).

Table 3.1 Conditions for accountability and blameworthiness

Conditions for accountability	Possible excuses to avoid blameworthiness (if held accountable)
Capacity	Ignorance (knowledge)
Causality	Compulsion (freedom)
Wrongdoing	(capacity, causality, wrongdoing)

accountable. The reason for this is that accountability shifts the burden of proof for blameworthiness: the agent is now to show – by giving an account – that she is not blameworthy. Such a shift in the burden of proof seems only reasonable if there is a reasonable suspicion of wrongdoing.

To summarize: the suggestion is that an agent A is accountable for X if A has the capacity to act responsibly (has moral agency), is somehow causally connected to the outcome X (by an action or omission) and there is a reasonable suspicion that agent A did somehow something wrong. A may then provide an account that she is not blameworthy. In this account she can refer to the knowledge and freedom conditions, but possibly also to the other three conditions. Table 3.1 summarizes this result.

3.6 Accountability

We now have two suggestions for when an agent A is accountable for X. One was given at the end of the previous section: A is accountable if A has the capacity to act responsibly, is causally involved in X and did something wrong. The other was given earlier: A is accountable for X if A had to see to it that "not X" (i.e. A was forward-looking responsible for "not X") and X happens to be the case. It might well be possible to combine these accounts. We could try to argue that in order for an agent to be forward-looking responsible for X that agent needs to posses moral agency (i.e. the capacity to act responsibly) and A needs to be able to causally influence the occurrence of X. In addition, we could try to argue that the occurrence of X constitutes wrongdoing given A's responsibility to bring about "not X".

Nevertheless, I think it is better not to merge the two suggestions completely. The reason for this is that I think it would tie accountability too closely to the notion of forward-looking responsibility as the obligation to see to it that. Typically, we do not only hold people accountable for not discharging their forward-looking responsibilities but also for not meeting other moral obligations, preeminently for not discharging their duties. One could deal with this by arguing that doing our duty is part of our forward-looking responsibilities. This would, however, eventually stretch forward-looking responsibility such as to include all of our moral obligations. I think this would water down the notion of forward-looking responsibility too much. It is therefore better to distinguish between moral obligations that refer to actions (duties) and ones that refer to outcomes (forward-looking responsibilities).

To do so, I will follow Goodin's characterization of forward-looking responsibility. He conceives of both duties and responsibilities as prescriptions of the general form (Goodin 1995:82):

A ought to see to it that X

Where A is some agent, and X is some state of affairs. For duties, X takes the form:

A does or refrains from doing ϕ

Where ϕ is some specific action. For responsibilities, X does not refer to specific actions of A. X can refer to states-of-the-world, to states-of-mind of A, to A possessing certain virtues, to actions of other agents, as long as it does not include specified actions of A. Goodin suggests that the exercising of forward-looking responsibility:

> . . . require[s] certain activities of a self-supervisory nature from A. The standard form of responsibility is that A *see to it* that X. It is not enough that X occurs. A must also have "seen to it" that X occurs. "Seeing to it that X" requires, minimally, that A satisfy himself that there is some process (mechanism or activity) at work whereby X will be brought about; that A check from time to time to make sure that that process is still at work, and is performing as expected; and that A take steps as necessary to alter or replace processes that no longer seem likely to bring about X. (Goodin 1995:83)

According to Goodin these self-supervisory activities are "genuine responsibilities" because "they are injunctions that mandate goals and very general classes of activities, rather than specific actions" (Goodin 1995:83). For Goodin the crucial distinction between duties and responsibilities is the discretionary component built into the latter (Goodin 1995:84). However, also most duties have a discretionary component. Such duties do not prescribe specific actions but rather forbid general classes of actions or prescribe actions with certain properties. There are usually several ways in which one can abide by duties like "tell the truth" or "do not lie". Duties are therefore often best seen as constraints on actions rather than as strict prescriptions.

Nevertheless, there is a sense in which responsibilities are different from duties and it is related to the presence of a discretionary component. The difference is that responsibilities do not require the agent to achieve the outcome X by her own actions. Responsibilities can be delegated whereas duties cannot. If I have the duty to tell the truth it is not enough that somebody else tells the truth or that the truth surfaces in some other way. Each of these does not count as fulfilling the duty; the duty can only be fulfilled by an action of mine. This is different in the case of responsibilities. Consider the following example. Suppose that I have a forward-looking responsibility to see to it that the door of the classroom is closed before my lecture commences. Initially, I can just wait and see whether somebody closes it. If so, my responsibility has been discharged and I can start my lecture. If not, I can ask one of my students to close the door. If he indeed closes the door my responsibility has been discharged. If not, I can decide to close the door myself in order to discharge my responsibility.

The example illustrates two points. One is that fulfilling my responsibility does not require that X is achieved by an action of mine. The other point is that responsibility requires, as also suggested by Goodin, some action on my part of a supervisory nature: I have to see to it that X is achieved. This supervisory activity refers, contrary to what Goodin believes, not to a responsibility but to a duty. This is so because the supervision is to be done by me and cannot be delegated.[17]

The above is helpful in further specifying what it means to say that A did or did not discharge her forward-looking responsibility to see to it that X. Although this responsibility is aimed at realizing X, the occurrence of X is not the main criterion whether A actually fulfilled her forward-looking responsibility. The reason is that X may occur even if A did not see to it that X because X may be caused by something else. So even if A did not discharge her responsibility X may be the case. Moreover, X may not be realized even if A saw to it that X; X may for example not come about due to circumstances that A could neither foresee nor control. Discharging responsibility thus basically means exercising one's (self-) supervisory duties to see to it that X.

As Goodin suggests, both forward-looking responsibilities and duties can lead to what he calls blame-responsibility (and to accountability I would add). Here we can make sense I think of the two suggestions made earlier. There is not one route to accountability and blameworthiness but rather two. One route, call it the consequentialist one, is rooted in A's forward-looking responsibility for X; the other, call it the deontic one, is rooted in some moral duty that is transgressed. On the consequentialist route, an agent is basically accountable for a state of affairs X if that agent was forward-looking responsible for "not X" but did not see to it that "not X". On the deontic route, an agent may said to be accountable if she transgressed a duty D (and blameworthy if she was not acting under compulsion or ignorance). The latter is however accountability for actions rather than for consequences. But suppose that the transgression of D results in X, where X is an undesirable state of affairs; here it might be said that A is accountable for X if A is a moral agent who performed an action that transgressed D, and this action caused or causally contributed to X.

It seems then that the two suggestions with which we started this section are actually two routes to accountability for consequences. Although it is important to see the differences between the routes, I think it can be argued that both routes are based on the same general conditions for accountability, i.e. capacity, causality and wrongdoing, although the latter two conditions have a different meaning in the different routes (see Table 3.2). In the deontic route, wrongdoing consists in the breach of a duty D; the causality conditions boils down to this breach of D (rather than something else) causing a state of affairs X (Feinberg 1970:222). For the consequentialist route, the wrongdoing consists in the not discharging of one's supervisory duties to see to it that "not X."

[17] It might in specific circumstances be possible to delegate some supervision, but the agent cannot delegate away all responsibility and still has a duty to supervise the supervision, et cetera.

Table 3.2 Two routes for accountability

Accountability for X	Consequentialist route	Deontic route
Capacity	Ability to act responsibly	Ability to act responsibly
Wrongdoing	Forward-looking responsibility for "not X" not discharged	Transgression of duty D
Causality	Ability to causally influence the occurrence of X	Transgression of D caused X

How does the causality condition fit into the consequentialist route? My suggestion is that it is a precondition for A's forward-looking responsibility for X (which in turn, in this route, is a precondition for accountability).[18] I do not think it makes sense to ascribe forward-looking responsibility for X to agents who are not able to exercise some causal influence over the occurrence of X.[19]

3.7 Conclusions

The account that I have developed in this paper suggests a number of things. First, it suggests, contrary to what some other authors have suggested (e.g. Vedder 2001; Duff 2007:30–31), that one can be backward-looking responsible (accountable or blameworthy) for a state of affairs without having been forward-looking responsible for preventing that state of affairs. Second, it suggests that it may be useful to distinguish between accountability and blameworthiness. Accountability refers to what some see as a primary meaning of responsibility as answerability and as being able to provide a justification for one's action. Such answerability or justification do not, however, yet imply blameworthiness, which is also often seen as a main connotation of responsibility. Distinguishing between accountability and blameworthiness therefore avoids conceptual confusion. In addition, I have suggested a relation between the two in terms of the traditional conditions for responsibility. Third, I have suggested that accountability may both be rooted in a duty transgressed as well as in a forward-looking responsibility not discharged. Although there are two routes to accountability, both seem to share the basic conditions for accountability: capacity, causality and wrongdoing, even if these conditions have a different content in the two routes.

[18] An alternative would be to require, in analogy with the deontological route, that the transgression of the supervisory duties implied by A's responsibility for "not X" caused X rather than that something else. Apart from that this implies a counterfactual that may be difficult to establish ("what if A had lived by her responsibility?"), this seems me intuitively implausible. If A had to see to it that "not X" and did not fulfill her responsibility, this seems enough reasons to hold A accountable if X occurs. This is different in the deontological route because there the duty D does not contain a direct reference to X.

[19] Note that this causality condition is much weaker than the one in the deontic route.

There are various ways in which the account that I have proposed can be extended. One is by including the notion of moral liability. It has been suggested that one can be liable without being blameworthy (Davis forthcoming). My framework suggests the further question whether one can also be liable without being accountable. Another direction in which the account can be extended is conceiving of responsibility as a triadic rather than as a dyadic relation and by incorporating the role of responsibility-as-virtue and the phenomenon of "taking responsibility".

Acknowledgments This paper was written as part of the research program "Moral Responsibility in R&D Networks", which is supported by the Netherlands Organization for Scientific Research (NWO) under grant number 360-20-160. An earlier version of this paper was presented at the International Conference on Moral Responsibility: Neuroscience, Organization & Engineering, Delft, August 24–27, 2009. I would like to thank the conference participants, my co-workers in the project Moral Responsibility in R&D networks, Sven Ove Hansson, and Michael Davis for comments on earlier versions. I am grateful to NIAS, the Netherlands Institute for Advanced Study, for providing me with the opportunity, as a Fellow-in-Residence, to rewrite and finish this paper during my stay in the academic year 2009–2010.

References

Baier, Kurt. 1972. "Guilt and Responsibility." In *Individual and Collective Responsibility: Massacre at My Lai*, edited by Peter A. French, 35–61. Cambridge, MA: Schenkman.

Bovens, Mark. 1998. *The Quest for Responsibility. Accountability and Citizenship in Complex Organisations*. Cambridge: Cambridge University Press.

Cane, Peter. 2002. *Responsibility in Law and Morality*. Oxford: Hart Publishing.

Casey, John. 1971. "Action and Consequences." In *Morality and Moral Reasoning: Five Essays in Ethics*, edited by John Casey, 155–205. London: Methuen.

Corlett, J. Angelo. 2001. "Collective Moral Responsibility." *Journal of Social Philosophy* 32: 573–84.

Davis, Michael. forthcoming. ""Ain't No One Here But Us Social Forces": Constructing the Professional Responsibility of Engineers." *Science and Engineering Ethics*. Online first. doi: 10.1007/s11948-010-9225-3.

Duff, Antony. 2007. *Answering for Crime: Responsibility and Liability in the Criminal Law*. Oxford: Hart Publishing.

Feinberg, Joel. 1970. *Doing & Deserving: Essays in the Theory of Responsibility*. Princeton, NJ: Princeton University Press.

Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Goodin, Robert E. 1995. *Utilitarianism as a Public Philosophy*. Cambridge: Cambridge University Press.

Hart, H.L.A. 1968. *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford: Clarendon Press.

Kutz, Christopher. 2000. *Complicity: Ethics and Law for a Collective Age*. Cambridge; New York: Cambridge University Press.

Ladd, John. 1982. "Philosophical Remarks on Professional Responsibility in Organizations." *International Journal of Applied Philosophy* 1:58–70.

Ladd, John. 1991. "Bhopal: An Essay on Moral Responsibility and Civic Virtue." *Journal of Social Philosophy* 32:73–91.

Lucas, J.R. 1993. *Responsibility*. Oxford: Oxford University Press.

Scheffler, Samuel. 1997. "Relationships and Responsibilities." *Philosophy and Public Affairs* 26:189–209.

Sher, George. 2006. “Out of Control.” *Ethics* 116:285–301.
Strawson, Peter. 1962. “Freedom and Resentment.” *Proceedings of the British Academy* 48: 187–211.
Swierstra, Tsjalling and Jaap Jelsma. 2006. “Responsibility Without Moralism in Techno-Scienctific Design Practice.” *Science, Technology & Human Values* 31:309–32.
Thompson, Dennis F. 1980. “Moral Responsibility and Public Officials: The Problem of Many Hands.” *American Political Science Review* 74:905–16.
Vedder, Anton. 2001. “Accountability of Internet Access and Service Providers: Strict Liability Entering Ethics?” *Ethics and Information Technology* 3:67–74.
Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.
Williams, Garrath. 2008. “Responsibility as a Virtue.” *Ethical Theory and Moral Practice* 11: 455–70.
Zimmerman, Michael J. 1988. *An Essay on Moral Responsibility*. Totowa, NJ: Rowman & Littlefield.
Zimmerman, Michael J. 2009. “Responsibility, Reaction and Value.” Paper read at International Conference on Moral Responsibility: Neuroscience, Organization & Engineering, August 24–27, 2009, at Delft University of Technology, the Netherlands.

Chapter 4
Beyond Belief and Desire: or, How to Be Orthonomous

Michael Smith

Abstract The standard belief-desire account of the explanation of action is inadequate to the task of explaining even the very simplest actions. We must suppose instead that three psychological states are in play when we explain action, not just two: desire, belief, and the exercise of the capacity to be instrumentally rational. Once we enrich our understanding of action explanation to acknowledge the causal role played by an agent's exercise of his rational capacities, much richer accounts of action explanation come into view, accounts that highlight the many different ways in which agents' actions can be explained by their rational capacities. Of special interest are cases in which agents' actions are explained by their failure to exercise their rational capacities, where these are capacities that they possess, and cases in which their actions are explained by their failure to exercise their rational capacities, where these are capacities that they do not possess. Richer accounts of action explanation such as these suggest a distinctive story about the conditions under which people are responsible for wrongdoing, a story with surprising implications for our understanding of what it is for an agent's moral beliefs to be justified.

4.1 Introduction

The standard belief-desire account of the explanation of action, at least in the form in which that account was put forward by Donald Davidson, is inadequate to the task of explaining even the very simplest actions (Davidson 1963). If some version of the standard account is correct, then we must suppose that it is a variation on the version put forward by Carl G. Hempel (1961) prior to Davidson. According to Hempel, three psychological states are in play when we explain action, not just two as Davidson supposes. Desire and belief are part of the explanation of every action, but so too is the capacity to be instrumentally rational, a capacity that is but one among many capacities rational agents possess (Smith 2009).

Once we enrich our understanding of action explanation to acknowledge the causal role played by an agent's exercise of his rational capacities, much richer accounts of action explanation come into view, accounts that highlight the many

M. Smith (✉)
Department of Philosophy, 1879 Hall, Princeton University, Princeton, NJ 08544, USA
e-mail: msmith@princeton.edu

N.A Vincent et al. (eds.), *Moral Responsibility*, Library of Ethics and Applied Philosophy 27, DOI 10.1007/978-94-007-1878-4_4,

different ways in which agents' actions can be explained by their rational capacities. Of special interest are cases in which agents' actions are explained by their failure to exercise their rational capacities, where these are capacities that they possess, and cases in which their actions are explained by their failure to exercise their rational capacities, where these are capacities that they do not possess (Smith 2003). Richer accounts of action explanation such as these suggest a distinctive story about the conditions under which people are responsible for wrongdoing, a story with surprising implications for our understanding of what it is for an agent's moral beliefs to be justified.

In the first section of the paper I explain why, and in what way, we need to go beyond the standard belief-desire account of action explanation. In the second section I outline the much richer kinds of explanation of action that come into view once we go beyond the standard account in the way suggested, and I describe the conditions of moral responsibility that they suggest. I also provide further support for this story about the conditions of responsibility by bringing out the similarities between it and the conception of responsibility found in the criminal law. And then in the third and final section I outline some of the surprising implications of this conception of responsibility.

4.2 Beyond the Standard Belief-Desire Account of the Explanation of Action

Consider a very simple case of action. Suppose that John non-instrumentally desires more than anything else to get muscly and believes with as much certainty as he believes anything that he can get muscly by exercising. Does it follow that he will exercise, if he does anything at all intentionally?

According to a principle that Donald Davidson accepts, this does follow. The principle is this:

> P1: If an agent wants to do x more than he wants to do y and believes himself free to do either x or y, then he will intentionally do x if he does either x or y intentionally. (Davidson 1970)

In our example, John satisfies this condition, so by Davidson's lights it follows that he will exercise if he does anything intentionally at all. I take it that this is why Davidson thinks that just two psychological states figure in the explanation of action. All we need to know in order to know what John will do intentionally, if he does anything intentionally, is what he desires and believes, or so Davidson seems to think. However a moment's reflection makes it clear that this isn't so.

Given P1, what should we say about the case in which an agent wants x exactly as much as he wants y, but wants each of these more than he wants anything else, and believes himself free to do various things, including either x or y? If P1 tells us everything we need to know, then all we could say is that he will do either x or y, if he does anything intentionally. But as we know from reflection on Buridan's Ass type cases, this isn't all that we can say. When a rational agent goes to the

supermarket and is confronted by three identical boxes of cereal, desires most to take one or another box, but desires to take each box exactly as much as he desires to take the others, he may still take (say) the one in the middle intentionally. This is because rational agents possess the capacity to just *pick* an alternative when their antecedent desires and beliefs leave them indifferent. They choose one intentionally despite the fact that they don't antecedently desire it more (Raz 1999:100).

To tell us everything we need to know in order to know what an agent will do intentionally, if he does anything intentionally, P1 would therefore need to be supplemented with an account of the role of this distinctive capacity that rational agents possess to pick or chose, a capacity whose exercise explains why they are not flummoxed, suffering the counterpart of starving to death in Buridan's Ass cases. This is therefore a psychological state of great normative significance, and it is also one whose exercise turns out to be empirically tractable. According to one study, for example, when the choice is between three or four or five identical items, as in the case of choosing a box of cereal in a supermarket, rational agents tend to avoid the endpoints, opting for the item in the middle (Christenfeld 1995).

It might be objected that this is all confused. If rational agents in supermarkets intentionally choose items in the middle of a row of identical items, then it follows that they must have at least some non-instrumental desire for things in the middle, a non-instrumental desire that breaks the alleged tie among agents' non-instrumental desires for objects in middle, those on the right, and those on the left. It might be thought that this follows from what it is to desire something. If a desire is just a disposition to choose, then there is no conceptual space for the idea of an additional capacity to choose. The agents in question might not desire the one in the middle more *antecedently*, but they do when they act.

But a little reflection suggests that this is not really an objection to what's been proposed. If the so-called desire for things in the middle only manifests itself in circumstances in which a tie needs to be broken between alternatives that can't be discriminated between by an agent's other non-instrumental desires – if it is the non-instrumental desire, when an agent's other non-instrumental desires leave a choice underdetermined, for those things in the middle, and if it is constitutive of being rational that agents have some such non-instrumental desire when a tie needs to be broken – then there is only a verbal difference between the suggestion that agents have such desires, and the suggestion that they have the capacity to choose an alternative when their desires leave them indifferent between alternatives, a capacity that they tend to exercise by going for the thing in the middle. The reply to the objection thus concedes everything that is at issue.

Once we allow that an additional role may be played in intentional action by the exercise of an agent's rational capacities, as in Buridan's Ass cases, the question immediately arises whether there are other situations in which a role is played by an agent's exercise of his rational capacities. And the answer is that there are. Almost ten years prior to Davidson, Hempel had put forward his own version of the standard account of action explanation. In the course of doing so, he had pointed out that whenever an agent acts on his desires and beliefs, he must also exercise a distinctive rational capacity to put his desires and beliefs together.

Consider once again our example. Suppose that John does non-instrumentally desire to get muscly more than he desires anything else and believes that he can get muscly by exercising with more certainty than he believes anything else. If he is instrumentally irrational he will not form the instrumental desire to exercise, and, absent the formation of that instrumental desire, he won't exercise (Hempel 1961:266–67). For desire and belief even to begin to play the role that Davidson describes in P1, we therefore need to suppose that a ubiquitous role is played by yet another psychological state. Hempel himself calls this state the agent's being rational, but in fact the psychological state in question is both more specific than this, and we have to understand the role that it plays in a certain way.

If an agent with non-instrumental desires and beliefs is to act at all, he must *have and exercise* the capacity to be *instrumentally rational*. It would not be enough for him merely to have the capacity to be rational, where this is a capacity that he may or may not exercise. To return to our example, even if John does have the capacity to be instrumentally rational, if he does not exercise it on the occasion, then he still will not form the instrumental desire to exercise, and, absent the formation of that desire, he will not exercise. Moreover, not just any old rational capacity will do the job. It wouldn't be enough if the agent exercised his capacity to form his beliefs in the light of the available evidence, for example. Indeed, the exercise of that capacity isn't even necessary for an agent to act. Having means-end beliefs is enough. How he came by his means-end beliefs is neither here nor there.

To see more precisely what the distinctive causal role is that's played by an agent's possession and exercise of his capacity to be instrumentally rational, we need to consider a slightly less simple case of action explanation, a variation on the case that we have discussed thus far. Suppose that John has a non-instrumental desire to get muscly and that he believes there are two ways in which he could do so. He believes that he could get muscly by exercising a lot, and he also believes that he could get muscly by exercising less, but taking pills as well. If he does it by exercising a lot, then it will take longer to get muscly, whereas if he does it by using the combination strategy, then he will get muscly sooner, but once the musculature is achieved, it will last equally long either way.

If John is as confident about one of these causal claims as he is about the other – equally confident that exercising a lot will cause him to get muscly and that exercising less and taking pills will cause him to get muscly – then, assuming that he doesn't care whether he gets muscly sooner or later, it follows that, if he were fully instrumentally rational, he would be indifferent between these options. His instrumental desire to exercise a lot and his instrumental desire to exercise less and take pills would be equally strong. He would be in a Buridan's Ass situation, and would therefore need to just pick an option.

But now suppose that John is equally confident about both strategies and that he opts for the combination strategy. If John's possession and exercise of the capacity to be fully instrumentally rational is part of the explanation of his pursuit of the combination strategy, then we already know something about what he would have done if the option of exercising and taking pills hadn't been available to him. John

would have exercised, notwithstanding the fact that it would take him longer to get muscly because, being fully instrumentally rational, he has an instrumental desire just to exercise waiting in the wings to produce action should it turn out that the combination strategy isn't available.

If, however, John is less than fully instrumentally rational – if, say, he has a tendency not to form instrumental desires when gratification is significantly delayed – then we have no grounds for supposing that he would have exercised if the option of taking pills hadn't been available to him. For though it follows from the fact that he exercises and takes pills intentionally that he is at least somewhat instrumentally rational, there is no reason at all to suppose that he is sufficiently instrumentally rational to have formed the desire simply to exercise as well and have it on standby. Indeed, if he has a tendency not to form instrumental desires when gratification is significantly delayed, there is every reason to suppose that he isn't sufficiently instrumentally rational to have formed that desire. Which counterfactuals are true of John thus depends on which explanatory hypothesis is correct.

Here, then, are the questions we must ask. Does John exercise and take pills because he is fully instrumentally rational and picks? Or does he exercise and take pills because, though he is less than fully instrumentally rational, since he had the option of exercising and taking pills, he didn't have to delay gratification? This is an empirical question, one whose answer is fixed by whatever psychological states are the causal antecedents of John's action. If John's action is caused by his possession and exercise of the capacity to be fully instrumentally rational, then he would have just exercised if the combination strategy hadn't been available. But if John's action is caused by his possession and exercise of a diminished capacity to be instrumentally rational, then he wouldn't have just exercised if the combination strategy hadn't been available.

Let's sum up the discussion thus far. We have seen that the standard belief-desire account of the explanation of action, at least in the form proposed by Davidson, is inadequate. We must suppose, with Hempel, that agents possess not only desires and beliefs, but also the capacity to be instrumentally rational to some extent. Moreover we must also suppose that their possession and exercise of this capacity plays its own distinctive explanatory role, complementary to the role played by their desires and beliefs, whenever agents act. We must also suppose that agents possess other rational capacities as well, capacities like the capacity to pick an alternative when antecedent desires and beliefs leave them otherwise indifferent.

I take this to be sufficient reason to move to a Hempelian, rather than a Davidsonian, conception of the standard account of action explanation. An agent's actions are explained by psychological states of three kinds, not just two: his desires, his beliefs, and the exercise of his rational capacities. But once we acknowledge that an agent's possession and exercise of rational capacities plays a distinctive role in the explanation of his actions, our eyes are opened to the possibility of much richer accounts of action explanation. These richer accounts in turn suggest a way in which we might begin to flesh out the conditions of responsibility.

4.3 The Nature of Responsibility

Consider once again the case in which John exercises and takes pills because, though he is less than fully instrumentally rational, he didn't have to delay gratification. It turns out that there are two possibilities here, depending on whether we suppose that John has a diminished capacity to be instrumentally rational which he fully exercises, or an undiminished capacity to be fully instrumentally rational that he fails to exercise on the occasion. There are therefore two corresponding further explanations of John's behaviour, depending on which of these possibilities is realized. In the first, John exercises and takes pills because he lacks the capacity to be fully instrumentally rational. In the second, he exercises and takes pills because, though he has that capacity, he fails to exercise it.

What's especially striking about these two further explanations is that they bear their relationship to ascriptions of responsibility more or less on their face. It follows from the very nature of responsibility that an agent who fails to act permissibly because he lacks the rational capacities required to act in that way is not responsible for failing to act permissibly. He is not responsible because his incapacity serves as an excuse. This is why children, those who are deranged, and those with volitional deficiencies like Obsessive Compulsive Disorder (OCD) are so often excused when they act wrongly. Children, the deranged, and those with OCD lack certain rational capacities, so when they act wrongly because they lack these capacities – and note that this needn't be true every time they act wrongly – they are thereby excused. Ascriptions of responsibility for wrongdoing are assignments of fault and these agents are not at fault.

By contrast, an agent who acts impermissibly because he fails to exercise rational capacities that he possesses is responsible for failing to act in that way. He is responsible precisely because he has no excuse. This is why someone who suffers from (say) weakness of will isn't treated like a child, someone who is deranged, or someone with OCD. Those who suffer from weakness of will have the capacity to will otherwise, but fail to exercise it. When they act impermissibly they are therefore liable to be held responsible because they are expected to exercise their capacity to will otherwise. The fact that they don't exercise their capacities is the problem, not an excuse. Fault is properly assigned to them.

Just to be clear, note I do not intend these claims to express a substantive moral view. They are meant to express conceptual claims, or, if you prefer, metaphysical claims, about what it is for an agent to be responsible. It is a priori that an agent is responsible for wrongdoing just in case he acts impermissibly without justification or excuse, and it is similarly a priori that when an agent's wrongdoing is explained by his lacking certain rational capacities, he has an excuse. This is why I said earlier that I took myself to be spelling out the nature of responsibility. These claims spell out *internal* correctness conditions of responsibility ascriptions, not substantive moral commitments.

There is, of course, a substantive moral view according to which we should treat agents in the way we typically treat responsible agents – we should, for example, punish them – only if they are responsible. In the theory of punishment, this is

the view held by retributivists. Others disagree. They hold that we should sometimes treat agents in the way we typically treat responsible agents even when they aren't responsible. In the theory of punishment, this is the view held by utilitarians. They believe that the fact that it would maximize happiness to punish someone (say) is always a good reason to do so, whether he is responsible or not.

Moreover, again just to be clear, the claim that an agent is excused of a wrong that he has done isn't a substantive moral claim either. In particular, it is not, and does not imply, the claim that the agent in question should be left free to do whatever he pleases. To say that an agent is excused of wrongdoing is simply to say that the wrong he did was not his fault. But even when the wrong that someone does is not his fault, his acting wrongly in the circumstances in which he did might still provide others with grounds for coercing him. Those who hold different substantive moral views can and should agree with this.

For example, retributivists and utilitarians can and should agree that someone who does wrong, but who is excused of that wrongdoing because he is deranged, may not be someone who should be left free to do what he pleases. There may be a justification for using coercive means to restrain him if he won't listen to reason. The crucial point is simply that the justification for coercing him could not be that he did something that was his fault. The justification would have to be that (say) what he did, together with his being deranged, shows that he is a danger to himself and others (this is the sort of justification that might be given by those attracted to retributivism on Kantian grounds, though of course this is no part of retributivism itself), or that coercing him would maximize happiness (this is the sort of justification that might be given by a utilitarian).

If what has been said so far is along the right lines, then this suggests a way in which we might proceed in order to fully spell out the conditions of responsibility. We might proceed by coming up with an exhaustive list of the rational capacities whose possession and exercise would be necessary for agents to be responsible when they act impermissibly. We have already seen that at least two such capacities would be required: the capacity to pick an alternative when our non-instrumental desires and means-end beliefs underdetermine the choice between alternatives and the capacity to put our non-instrumental desires together with our means-end beliefs. Are there any others?

A capacity suggested by the foregoing discussion is the capacity to form beliefs in the light of the evidence available to us. Someone who harms another in the course of satisfying some instrumental desire he has is not excused of wrongdoing merely because he had no idea that harm would result. Ignorance is no excuse because we are expected to exercise such capacities as we have to access relevant evidence, and then to form our beliefs on the basis of that evidence. But an agent who was literally *incapable* of forming the belief that harm would result from something that he does would be excused. If he lacks the capacity to access the relevant evidence, or the capacity to form beliefs in the light of that evidence, then the harm that he causes is not his fault. If he has, but simply fails to exercise, these capacities, however, then the harm he causes is his fault.

What about an agent's non-instrumental desires? Are there capacities to form non-instrumental desires that agents must possess if they are to be responsible when they act impermissibly? This is an issue on which there are deep divisions within philosophical psychology. Some theorists line up behind Hume who thinks that the non-instrumental desires that move us to action are "original existences" (Hume 1740). They think that the answer therefore has to be "No". Others line up behind Kant who thinks that we have the capacity to allow only those non-instrumental desires that accord with universal laws of reasons to motivate us (Kant 1786). They think that the answer has to be "Yes". But without addressing the issue that divides these theorists head-on, note that there is at least one reason for supposing that the answer must be "Yes", a reason that can be appreciated by followers of both Hume and Kant alike.

According to the doctrine in meta-ethics known as *Judgement Internalism*, if an agent believes that he ought to φ, then is motivated to φ, at least insofar as he is rational (Smith 1994:63–84). Given that an agent's motivations are constituted by his non-instrumental desires and means-end beliefs, *Judgement Internalism* implies that if an agent believes that his φ-ing is a basic wrong – that is, if he believes that it is a wrong simply in virtue of its being a φ –ing – then, insofar as he is rational, he will have a non-instrumental aversion to φ-ing. The capacity to be rational thus mediates between an agent's beliefs about the things that would be basic wrongs to do and his non-instrumental aversions. For an agent to fail to have a non-instrumental aversion to doing what he believes it would be a basic wrong to do therefore implies irrationality on his behalf, where, as above, this irrationality might therefore be grounded in two quite different sorts of fact about him.

On the one hand, the agent's irrationality might be grounded in an incapacity to acquire non-instrumental aversions that accord with his beliefs about basic wrongs. In commonsense terms, this would be for him to do wrong because he has no capacity for self-control. He knows what he should do, but he can't get himself to want to do it. If this is the form that his irrationality takes, then, as before, if he acts wrongly, he has an excuse, for his inability to control himself constitutes his excuse. His acting wrongly is not his fault. Alternatively, though the agent possesses the capacity to control himself, his irrationality might be grounded in his failure to exercise this capacity. If this is the form that his irrationality takes, then, as before, he has no excuse, for his wrongdoing is his fault.

Moreover, note that there will evidently be complicated mixed cases. If an agent does something wrong because he is (say) crazed on drugs, but he didn't lack the capacity for self-control when he took the drugs, and he knew at that earlier time that taking drugs would cause him to become deranged and do something wrong, then though there is a sense in which he does wrong because he lacks the capacity for self-control – when he was crazed, he couldn't control himself – his doing wrong isn't grounded in his lack of self-control in the way it would have to be to constitute an excuse. This is because his doing wrong can be traced to his earlier failure to exercise the capacity that he possessed for self-control when he took the drugs. His wrongdoing may therefore still be his fault.

So far we have focused on rational capacities that mediate between an agents' beliefs about basic wrongs and their non-instrumental desires. But *Judgement Internalism* implies that additional *cognitive* capacities will also be required if agents are to be responsible for acting impermissibly. We have already seen that when it comes to belief formation, agents who have capacities to access relevant evidence, and then to form their beliefs on the basis of that evidence, are expected to do so. When it comes to exercises of self-control, what is especially important is therefore that agents exercise their capacity to access relevant evidence about which acts are wrong, and that they exercise their capacity to form beliefs on the basis of that evidence. The upshot is thus that even those who do have the capacity for self-control, and who exercise that capacity, may still not be responsible for wrongdoing if they are not responsible for the moral beliefs on which they act.

For example someone who doesn't know that what he is doing is wrong will be excused for his wrong-doing if his ignorance results from an inability to know that what he is doing is wrong: that is to say, he will be excused if he is ignorant because he lacks either the capacity to access the available evidence, or the capacity to form beliefs in the light of that evidence. However if he doesn't know that what he is doing is wrong because he has, but fails to exercise, his capacities to access available evidence, or to form beliefs in the light of that evidence, then he has no excuse. His wrongdoing is his fault because he could and should have exercised his capacity to know what's right and wrong.

Let's step back for a moment. What's remarkable about this account of the internal correctness conditions of ascriptions of responsibility isn't just that it can be derived entirely a priori by reflection on what the rational capacities are whose exercise might play a role when an agent acts, but also that it bears a striking similarity to conceptions of criminal responsibility that we find in the law. This shouldn't really be surprising, given that legal conceptions of criminal responsibility have a distinctively retributivist flavour. But the fact that it is so adds additional credence to the account of responsibility I have been sketching. Here are some examples, just to drive the point home.

The minimum age at which someone can be held criminally responsible in Australia is ten, as younger children are deemed to be incapable of knowing the difference between right and wrong (Australian Government 2005). Australian law also includes the doctrine of *Doli Incapax*. According to this doctrine, though children between the ages of ten and fourteen may possess the capacity to know the difference between right and wrong, they are presumed not to. The presumption therefore has to be proved mistaken before a criminal case can proceed. Both of these ideas fit very smoothly with the idea that agents who are incapable of knowing that what they are doing is wrong, children being a prime example, are excused when they act wrongly.

The law on insanity as a defence in criminal cases builds on a related idea. The law on insanity was developed in *Queen v M'Naghten* in 1843 when Daniel M'Naghten approached a man who he believed to be Sir Robert Peel, the then Prime Minister of England, and shot him in the back, so killing him. When M'Naghten was tried for the man's murder it emerged that he firmly believed that Peel was out to

kill him. After testimony from medical experts, M'Naghten was found not guilty by reason of insanity. The decision caused such a controversy that the House of Lords asked the Lords of Justice to formulate a strict definition of when insanity could be used as a defence against criminal charges. According to the definition, now known as the "M'Naghten Rule", insanity is a defence only if:

1. At the time that the act was committed
2. the defendant was suffering from a defect of reason, from a disease of the mind, which caused
3. the defendant to not know
   a. the nature and quality of the act taken or
   b. that the act was wrong. (Hall 2008:226–27)

The M'Naghten Rule is widely accepted in jurisdictions influenced by English law. In crucial respects, the idea behind the M'Naghten Rule is much the same as before. An agent who has lived for long enough to develop the capacity to know what's right and wrong, but who suffers from some "disease of the mind" that destroys that capacity, is excused of wrongdoing if his wrongdoing is the result of his incapacity to know either what he is doing or that what he is doing is wrong.

Critics of the M'Naghten Rule insist that its exclusive focus on cognitive incapacities results in too narrow a conception of the insanity defence. In some jurisdictions, it is therefore supplemented with what is known as the "Irresistible Impulse Test". The Irresistible Impulse Test, developed in a decision by the Alabama Supreme Court in the USA in *Parsons v State* in 1887, holds that even if an accused party could tell right from wrong, he may still be excused by reason of insanity:

A. if mental disease caused the defendant to so far lose the power to choose between right and wrong and to avoid doing the alleged act that the disease destroyed the defendant's free will, and
B. if the mental disease was the sole cause of the act. (Lippman 2009:279)

The Irresistible Impulse Test thus also fits very smoothly with the account of the conditions of responsibility sketched in the previous section. What's relevant to the Irresistible Impulse Test is the capacity for self-control: that is, an agent's capacity to form non-instrumental aversions to doing those things he believes to be basic wrongs. If an agent lacks the capacity for self-control, and acts because he lacks that capacity, then he is excused because his act is not his fault.

Let's sum up. Not only are actions are always explained, inter alia, by agents' exercise of their capacity to be instrumentally rational, a capacity they might possess to a greater or a lesser extent, but many actions are also explained by agents' other rational capacities, sometimes by their exercise of these capacities, and sometimes by their failure to exercise them. These other rational capacities include the capacity to access available evidence; the capacity to form beliefs in the light of that evidence, both beliefs about means to ends and also beliefs about what's right and wrong; and

the capacity to exercise self-control, that is, the capacity to form non-instrumental desires and aversions in the light of beliefs about which acts are right and wrong.

Cases in which actions are explained by agents' failure to exercise their rational capacities are in turn of two types. There are those in which their actions are explained by a failure to exercise rational capacities that they possess, and there are those in which their actions are explained by the fact that they do not possess those capacities in the first place. This is important when it comes to ascriptions of responsibility, because it is part of the internal correctness conditions of such ascriptions that when agents act wrongly because they lack some relevant rational capacity, they are excused, whereas when they act wrongly because they have, but fail to exercise, some relevant rational capacity, they are not excused. Distinctions widely made within the criminal law give some support to these ideas.

4.4 Implications

At the very beginning I said that we need to move beyond Davidson's version of the standard story of action explanation, and that, when we do, a distinctive story emerges about the conditions under which people are responsible for wrongdoing, a story with surprising implications about the justification of an agent's moral beliefs. Let me now spell out some of these implications.

In "Sanity and the Metaphysics of Responsibility", Susan Wolf describes an agent whose responsibility is seriously in doubt.

> JoJo is the favourite son of Jo the First, an evil and sadistic dictator of a small, undeveloped country. Because of his father's special feelings for the boy, JoJo is given a special education and is allowed to accompany his father and observe his daily routine. In light of this treatment, it is not surprising that JoJo takes his father as a role model and develops values very much like Dad's. As an adult, he does many of the same sorts of things his father did, including sending people to prison or to death or to torture chambers on the basis of whim. He is not coerced to do these things, he acts according to his own desires. Moreover these are desires he wholly wants to have. When he steps back and asks 'Do I really want to be this sort of person?' his answer is resoundingly 'Yes,' for this way of life expresses a crazy sort of power that forms part of his deepest ideal. . .In light of JoJo's heritage and upbringing — both of which he was powerless to control — it is dubious at best that he should be regarded as responsible for what he does. It is unclear whether anyone with a childhood such as his could have developed into anything but the twisted and perverse person that he has become. (Wolf 1987:53–54)

Wolf doesn't explicitly describe JoJo in the detailed terms that we have seen are crucial for understanding his responsibility. It is, however, easy to see how different ways of filling in the details of his story would affect his responsibility, and while certain of these ways of filling the story deliver unsurprising results, others deliver results that are much more surprising.

Here is one way in which JoJo's story might be filled in. JoJo's father, Jo the First, developed a talent for the manipulation of other people for his own purposes when he was a young boy. It was this talent for manipulation that enabled him to become ruler of the small, undeveloped country, in which he grew up. Once he became

leader, Jo the First saw to it, sometimes by manipulation, but when manipulation failed, by whatever means were necessary, that all of those who lived in the country did exactly what he wanted. No one spoke ill of him, not even in private, for fear of dire consequences.

When JoJo was born, Jo the First fixated on him. He saw in JoJo someone who could see to it that no one would speak ill of him even after he died. He therefore gave JoJo a special education which consisted of humiliating him and then building him back up by making him believe that the only way he could have any worth at all was by getting his father's approval, something that he could do by emulating his father's behaviour, singing his praises, and generally seeing to it that others did nothing that his father wouldn't like. When people criticized his decision to home-school his son, he had them silenced. JoJo too therefore developed a talent for the manipulation of other people for his own purposes – many of which were of course Jo the First's purposes – when he was a young boy.

After Jo the First died, JoJo took over as ruler of the country, doing many of the same sorts of things his father had done, including sending people to prison or to death or to torture chambers on the basis of whim. He did all of this willingly, constantly singing the praises of his father and seeing to it that no one ever said anything to challenge the official view of his father as a great man. Since he was following in his father's footsteps, this meant that he had to see to it that people treated him as a great man as well. When he had a son, he saw in him exactly what his father had seen in him, and decided to give him the same education that he had received. When people criticized his decision to homeschool his son, he had them silenced. Brutally.

If we fill in the details of JoJo's story in this way, then Wolf is surely right that his responsibility is seriously in doubt, as JoJo seems to have been brainwashed to do his father's bidding. His belief that his father is a great man, the belief which sustains his desire to emulate his father's actions, is kept in place not by his assessment of the evidence available to him for his father's greatness, but rather by his need for his father's approval. JoJo seems to lack the capacity to form beliefs about whether his father's, and hence his own, actions are right or wrong in the light of the evidence available to him. He therefore seems not to be responsible for reasons similar to the reasons why children and the insane are not responsible.

Wolf herself points out that JoJo is similar to those who fall under the M'Naghten Rule (Wolf 1987:55). Indeed, she suggests that we should suppose that JoJo is "insane" in an "admittedly specialized sense":

> [A]lthough like us, JoJo's actions flow from desires that flow from his deep self, unlike us, JoJo's deep self is itself insane. Sanity, remember, involves the ability to know the difference between right and wrong, and a person who, even on reflection, cannot see that having someone tortured because he failed to salute you is wrong plainly lacks the requisite ability. (Wolf 1987:56)

There are, of course, differences between JoJo and those who are criminally insane. To return to the M'Naghten Rule, JoJo's brainwashing doesn't seem to have caused a "disease of the mind", unless, of course, we are using the term "disease"

in a highly metaphorical sense. (I note in passing that the M'Naghten Rule itself doesn't define what a disease of the mind is, and that in some jurisdictions it is interpreted very broadly.)

On the other hand, however, JoJo's brainwashing plainly has caused him to have a mental problem, a problem whose upshot is a lot like the upshot of M'Naghten's disease of the mind. M'Naghten's disease of the mind caused him to be insensitive to evidence as to whether or not Peel was out to kill him in his formation of his belief that this is what Peel was out to do. JoJo's belief that his father's actions are not wrong is insensitive to evidence as to whether or not his actions are right or wrong because his brainwashing has caused his desire for his father's approval to sustain his belief independently of evidence. Perhaps this similarity is all Wolf needs to be right that JoJo is insane "in a specialized sense". We will return to this point below.

It might be helpful if we think in terms of diagrammatic representations. The responsible agent's actions can be represented as in Fig. 4.1.

In Fig. 4.1, the "→"s represent either a relation that is knowledge conducive (1, 2), or the exercise of some relevant rational capacity (3, 4, 5, 6), or causation of a kind that sustains differential explanation (7, 8) (for more on this, see Smith 2004, 2009). In these terms, what's crucial about JoJo, at least when we fill in the details of his story as we did above, is that he lacks a crucial rational capacity that the responsible agent possesses and exercises: specifically, he lacks the capacity at junction 4. This suffices to excuse him from wrongdoing when his wrongdoing is explained by his lack of this capacity. He is excused, in such cases, because his wrongdoing is not his fault.

There is, however, the following rather different way in which the details of JoJo's story might be filled in. As he grew up, JoJo was given a special education, both formal and informal, by his father, Jo the First. Jo the First was brutal, but also articulate and larger than life, much like one of the main characters in a Quentin

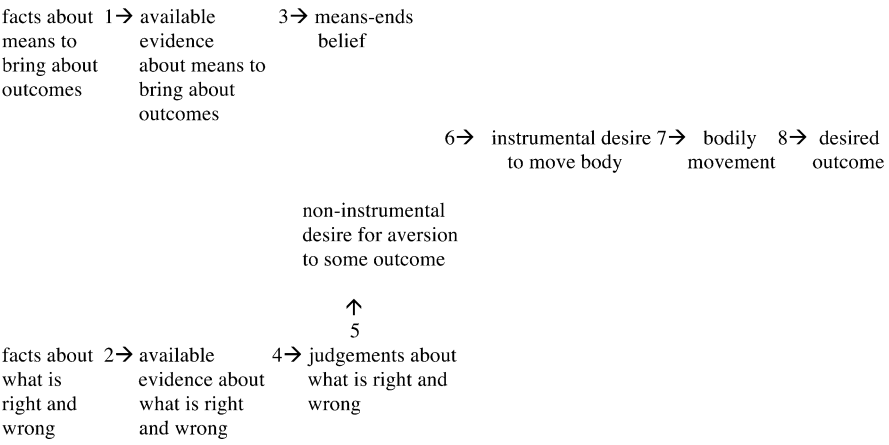

Fig. 4.1 The responsible agent

Tarantino movie. As a result, JoJo came to have the rather idiosyncratic belief that someone who has the sort of power his father has, exercised in the stylish way in which he exercised it, thereby has the right to do whatever he has to do in order to maintain that power, no matter what the consequences are for those he doesn't really care about. JoJo read widely, something his father encouraged, so he realised that his view about his father's entitlements was idiosyncratic. Though this was initially a cause of some cognitive dissonance, one day JoJo came across some books by a German philosopher that contained an elaborate statement and defence of ideas that imply that something like what he had come to believe about his father was in fact true. After reading the German philosopher's work for himself, he concluded that, idiosyncratic though his ideas were, there was also something deeply intuitive about them.

JoJo's interest in philosophy led him to apply to graduate school at an Ivy League university in the USA where the world's leading expert on the German philosopher taught. JoJo's admission file was so strong that he was given a special scholarship. He eventually wrote a dissertation defending his own unusual moral views, drawing on their similarities to those of the German philosopher's, via a wide reflective equilibrium argument, a dissertation that eventually became a celebrated monograph. JoJo's monograph was much discussed by academic philosophers, and also much discussed in the pages of the New York Review of Books and on NPR.

While many reviewers thought that the arguments JoJo gave were spot on, others thought that the position itself, though internally consistent, and though consistent with everything else JoJo believed, depended on basic premises that were themselves manifestly implausible. JoJo's conclusions could be true only if some of the things that they themselves believed were true were false, but they did not think that anything JoJo said had provided them with sufficient reason to reject the truth of what they believed. Some of these reviewers, ignorant of JoJo's history, went so far as to say that it just as well that JoJo was a harmless academic, as by this time he had secured a professorship of his own at a prestigious left-wing university in northern California.

JoJo saw nothing unusual in the fact that his colleagues had such starkly different opinions about his work. He had come to the view years earlier that there are no philosophical theories on any subject matter that command universal assent. As he saw things, all philosophical theories are defended via wide reflective equilibrium arguments of the kind he had given, and this meant that the very deepest philosophical disagreements amounted to disagreements about fundamental premises: that is, they were disagreements about which beliefs are, and which are not, supposed to survive such an argument. He went on to become something of an academic celebrity, universally admired for his charm and wit and intelligence and loyalty to his graduate students and close colleagues, but also feared by those who experienced how ruthlessly dismissive he could be of those with whom he saw no profit to engage.

After Jo the First died, JoJo seized the opportunity to put his ideas into practice on a much larger scale. He returned home to take over as ruler of his country. He did many of the same sorts of things his father had done, including sending people

to prison or to death or to torture chambers on the basis of whim. He did all of this willingly, constantly singing the praises of his father and seeing to it that no one ever got away with challenging the official view of his father as a great man. When people criticized his father, or him, he would send them a copy of his monograph and a long list of references to papers in academic journals in which philosophers wrote at length defending the essentials of his views. When he eventually had a son, and people criticized his decision to homeschool him, he would remind them of that one of their heroes, John Stuart Mill, was himself homeschooled by his father. If they persisted with their criticisms, and became disruptive or unpleasant, he had them silenced. Brutally, but stylishly.

If we tell JoJo's story in this way, is it plausible to suppose that he is not responsible for his wrongdoing? By contrast with the earlier telling of his story, it doesn't seem that JoJo's beliefs are the product of brainwashing or wish fulfilment. They are rather the product of deep thought and rational assessment. Indeed, when we tell the story in this way, JoJo *seems* to be at least as rational as anyone we are likely to meet, more rational than most. Moreover he seems to be exceptionally diligent in his exercise of his rational capacities. Given his education and his dedication to the academic enterprise, we might even be tempted to suppose that JoJo's moral beliefs, though false, are as justified as anyone's could be. JoJo thus doesn't seem to lack any rational capacities. But if he lacks no rational capacities, then how could he not be responsible?

What this way of telling JoJo's story teaches us, I think, is that we need to get much clearer about what's happening at both junctions 2 and 4 in Fig. 4.1. In order to do this, it will be helpful if we first of all think about junctions 1 and 3. Imagine someone who lacks peripheral vision, and to whom it therefore seems that there are no objects in his immediate environment when in fact there are. He therefore regularly acquires beliefs about things he can do that are false: for example, he regularly acquires the belief that it is safe to cross the road, when in fact crossing the road would cause him to be hit by a car. Is such a person responsible for his false beliefs? This question would not be easy to answer in practice, but we know how to answer it in theory.

Assuming that the person we are imagining didn't cause his own lack of peripheral vision, he certainly isn't responsible for its seeming to him that there are no dangerous objects in his environment when there are, because he can't help how things seem to him. That is just a given, a function of his perceptual system. Of course, since experience has presumably taught him that he shouldn't trust how things look to him in forming his beliefs about how things are, he may well be responsible for not pausing to ask whether things really are, in every detail, the way they look to him to be. But it is an empirical question whether human beings really do have the capacity to resist the natural inclination to form perceptual beliefs on the basis of perceptual appearances during the hustle and bustle of daily life. Perhaps the connection between perceptual appearances and perceptual beliefs at such times is so immediate that that kind of second-guessing simply isn't realistic. Either human beings don't have such a capacity, or, if they do, it is a capacity that it would be very difficult for them to exercise.

If this is right, then the person we are imagining may not be responsible for his false beliefs at all, or his responsibility might be seriously mitigated. He is not responsible if he is incapable of having the world seem to him to be the way it really is and he non-culpably finds himself unable to resist the natural tendency to believe that things are how they seem to him to be on the basis of the indirect evidence available to him. His responsibility is mitigated if, though he has the latter capacity, it would be very difficult for him to exercise it during (say) the hustle and bustle of daily life. We might put the same points more explicitly in terms of the language of Fig. 4.1 as follows. Focus on the case in which he isn't responsible at all. The person who lacks peripheral vision has two problems. First, perceptual evidence about how things are in certain regions of his immediate environment isn't available to him because things don't seem to him to be the way that they are. And second, indirect evidence about how things are in those regions – for example, indirect evidence that, for all he knows, there is something in those regions of his immediate environment – though available to him, isn't evidence to which he has the capacity to be sensitive during the hustle and bustle of daily life. He therefore isn't responsible for (say) his false belief that it is safe to cross the road because his false belief isn't his fault.

Let's now consider what to say about JoJo in the light of this. Focus on junctions 2 and 4. JoJo acquired the false belief that Jo the First was within his rights to do the brutal things he did in the way in which children usually acquire their moral beliefs, that is, by an informal process of socialization. When he became an adult, however, he questioned whether his beliefs were true, and he concluded that they were. He reached this conclusion in two ways. It both seemed to him that they were true – that is, what he believed was, he thought, deeply intuitive – and, furthermore, after thinking long and hard about questions in moral philosophy, he came up with a reflective equilibrium argument for a theory that entailed the truth of the things that he believed. So is JoJo responsible for his false belief? The answer to this question bears certain similarities to the answer we just gave about the person who lacks peripheral vision.

JoJo also has two sorts of problems. First, direct evidence of his father's wrongdoing isn't available to him, because the things that seem permissible to him aren't permissible. His father seems to JoJo to have the right to brutalize people, when in fact he has no such right. This is strictly analogous to what we said about the person who lacks peripheral vision. Second – and this is a difference between JoJo's case and that of the person who lacks peripheral vision – indirect evidence of his father's wrongdoing isn't available to him either. For in order to access indirect evidence of the wrongness of his father's actions, JoJo would have to be able to construct a theory that entailed that his father's acts were wrong via an attempt to get his beliefs into a wide reflective equilibrium. But he can't. When he succeeds in his attempt to get his beliefs into a wide reflective equilibrium, the theory that he comes up with entails that his father's acts are not wrong.

We can put the same point more simply as follows. For JoJo to be able to access evidence of his father's wrongdoing, there would have to be something that he believes, or something that he feels, or some way that things seem to him to be, that doesn't square with his father's having a right to brutalize people. Absent some

such psychological hook, JoJo will be unable to reason himself to the conclusion that his belief that his father has a right to brutalize people is false because there will be nothing for him to reason from. But there are no such psychological hooks in JoJo. His beliefs, his feelings, the ways things seem to him to be, all of these things square with his belief that his father has a right to brutalize people. The upshot is that JoJo isn't responsible for his false belief that his father has a right to brutalize people. He isn't responsible because he cannot access evidence to the contrary. His false belief isn't his fault.

I said at the beginning that the distinctive story we've told about the conditions under which people are responsible for wrongdoing has surprising implications for the justification of an agent's moral beliefs. These implications are implicit in the conclusions we have just drawn from the second way of filling in the details of JoJo's story. As we have seen, even though JoJo succeeds in getting his beliefs into a wide reflective equilibrium, his moral beliefs are false. Should we suppose that his moral beliefs are justified? There are two ways we could go in answering this question. On the one had, we might suppose that the description of the wide reflective equilibrium procedure itself just is an account of the conditions under which an agent's moral beliefs are justified, so that the answer has to be that JoJo's moral beliefs are justified. On the other, we might wonder whether basic moral beliefs that an agent holds only because he is irrational could ever be justified. Since this seems to be so in JoJo's case, we might conclude that the answer has to be that his basic moral beliefs are not justified. I won't decide between these two ways in which we might answer the question in what follows. I will simply spell out the second way of answering the question in a little more detail.

Think again about the difference between the person who lacks peripheral vision and JoJo. In both cases, the world seems to them to be a certain way when it isn't that way. But in the case of the person who lacks peripheral vision, this fact about him doesn't suggest irrationality of any kind. The defect lies in his perceptual system, not in the capacities he possesses insofar as he is a reasoner. In JoJo's case, by contrast, the fact that the world seem to him to be a certain way when it isn't that way does suggest irrationality of some kind. It suggests irrationality because it entails a limitation on his abilities as a reasoner. There is, of course, an assumption that I'm making here, namely, that knowledge of basic moral truths is a priori accessible. But if this assumption is correct, as I think it is (Smith 1994, Chapters 5 and 6), then given that the fact that the world seems to JoJo to be a certain way in basic moral respects when it isn't that way is what explains his inability to know basic moral truths, it follows that that fact is also what explains why he isn't an ideal reasoner. An ideal reasoner is, after all, someone with the ability to know a priori truths.

If this way of thinking about the justification of JoJo's beliefs is correct, then it follows that we need to radically rethink the epistemic significance of the reflective equilibrium procedure (compare Scanlon 2002). Though JoJo succeeds in getting his beliefs into a wide reflective equilibrium, given that he achieves that wide reflective equilibrium only because he isn't an ideal reasoner, we should conclude that the beliefs he comes up with are not justified. To be justified, an agent's moral beliefs mustn't just be such that they would survive his attempts to get his beliefs into a wide reflective equilibrium. That wide reflective equilibrium itself mustn't be sustained

by the agent's inability to know certain a priori truths. We are therefore led to the conclusion, which may well come as a surprise to some, that whether or not we think that an agent's moral beliefs are justified will depend crucially on what we take the moral truth to be, as this will in turn affect which reasoning capacities we take the justification of an agent's moral beliefs to depend upon.[1]

References

Australian Government. 2005. "The Age of Criminal Responsibility, Australian Institute of Criminology." Accessed 21 January 2011. http://www.aic.gov.au/publications/current%20series/cfi/101-120/cfi106.aspx.

Christenfeld, N. 1995. "Choices from Identical Options." *Psychological Science* 6:50–55.

Davidson, Donald. 1963. "Actions, Reasons and Causes." Reprinted in *Essays on Actions and Events*, edited by Donal Davidson, 3–20.Oxford: Oxford University Press, 1980.

Davidson, Donald. 1970. "How Is Weakness of the Will Possible?" Reprinted in *Essays on Actions and Events*, edited by Donal Davidson, 21–42. Oxford: Oxford University Press, 1980.

Hall, Daniel. 2008. *Criminal Law and Procedure*. New York, NY: Cengage.

Hempel, Carl G. 1961. "Rational Action." Reprinted in *Readings in the Theory of Action*, edited by Norman S. Care and Charles Landesman, 285–86. Bloomington, IN: Indiana University Press, 1968.

Hume, David. 1740. *A Treatise of Human Nature*. Oxford: Clarendon Press, 1968.

Kant, Immanuel. 1786. *Groundwork of the Metaphysics of Morals*. London: Hutchinson and Company, 1948.

Lippman, Matthew. 2009. *Contemporary Criminal Law: Concepts, Cases, and Controversies*. Thousand Oaks, CA: Sage Publications.

Raz, Joseph. 1999. "Explaining Normativity: Reason and the Will." In *Engaging Reason: On the Theory of Value and Action*, 90–117. Oxford: Oxford University Press.

Scanlon, Thomas M. 2002. "Rawls on Justification." In *The Cambridge Companion to Rawls*, edited by Samuel Freeman, 139–67. New York, NY: Cambridge University Press.

Smith, Michael. 1994. *The Moral Problem*. Oxford: Wiley-Blackwell.

Smith, Michael. 2003. "Rational Capacities." In *Weakness of Will and Varieties of Practical Irrationality*, edited by Sarah Stroud and Christine Tappolet, 17–38. Oxford: Oxford University Press.

Smith, Michael. 2004. "The Structure of Orthonomy." In *Action and Agency*, edited by John Hyman and Helen Steward, 165–93. Cambridge: Cambridge University Press.

Smith, Michael. 2009. "The Explanatory Role of Being Rational." In *Reasons for Action*, edited by David Sobel and Steven Wall, 58–80. New York, NY: Cambridge University Press.

Wolf, Susan. 1987. "Sanity and the Metaphysics of Responsibility." In *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*, edited by Ferdinand Schoeman, 46–62. New York, NY: Cambridge University Press.

[1] Earlier versions of this paper were presented *at Moral Responsibility: Neuroscience, Organization, and Engineering*, a conference held at Technical University Delft in August 2009; at *Workshop on Reasons and Rational Choice* held at the London School of Economics in January 2011; and at the Philosophy Departments at Lingnan University and Monash University in March 2011. Thanks are due to the many people who gave me comments on these occasions. Special thanks are owed to Jay Wallace for the conversations we had while I was writing up the penultimate version and to Nicole Vincent for her written comments. Work on this paper was completed while I visiting the Humboldt University in Berlin, enjoying the benefits of a Forschungspreis from the Alexander von Humboldt Foundation.

Chapter 5
Blame, Reasons and Capacities

Rosemary Lowry

Abstract It is usually agreed that we must recognise that responsibility (in the sense of blameworthiness) comes in degrees if we are to accurately reflect the moral landscape of people's actions. In this paper I develop this view by constructing a framework which will allow us to determine the degree to which an agent is blameworthy for failing to act. This framework accommodates the close connection between an agent's blameworthiness and her reasons, which I argue should lead us to see reasons as coming in degrees. The view that reasons come in degrees is justified on the basis of two claims: first, reasons are constrained by what it is possible for the agent to do, and second, it may be possible *to some degree* for an agent to do something. I conclude the paper by demonstrating how this framework can be used to justify claims about the degree to which an agent has a reason, and the degree to which an agent can be blameworthy in a given case.

5.1 Introduction

While our legal system often requires us to find someone to be guilty or not guilty, it is usually recognised that in reality varying degrees of punishment and censure are appropriate.[1] In order to accommodate varying degrees of punishment and censure, we need an account of responsibility that comes in degrees. The term "responsibility" can be attached to different meanings. In this paper I will be concerned with the sense of responsibility that is associated with the notion of blame. I will be particularly concerned with the occasions when an agent is responsible for failing to act, in the sense of the agent being blameworthy for this failure.

Recognising that responsibility (in the sense of blameworthiness) comes in degrees is important if we are to accurately reflect the moral landscape of people's actions. This recognition helps us to avoid "a social and political picture which simplistically divides people into sheep and shepherds: those who just can't attain responsibility and those who have it absolutely." (Skorupski 2007:101). We must avoid such a picture, because in reality, responsibility is not an "all-or-nothing" concept; in Skorupski's words, "it is not a package deal".

[1] Honoré (1998) discusses this issue.

R. Lowry (✉)
Department of Philosophy, Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: r.j.lowry@tue.nl

N.A Vincent et al. (eds.), *Moral Responsibility*, Library of Ethics and Applied Philosophy 27, DOI 10.1007/978-94-007-1878-4_5,

My specific concern in this paper is to develop a framework which will allow us to determine the degree to which an agent is responsible for failing to act (in the sense of being blameworthy for this failure). If a framework for determining blameworthiness is to be adequate however, it must also accommodate the close connection between an agent's blameworthiness and her reasons. Heuer formulates this connection. In her words, "it is a necessary condition of justified blame that we can blame a person only for (not) doing something that she has a reason to do (or not to do)." (Heuer 2010:237) Following Heuer, I will call this the "reason condition of blame".[2]

The reason condition of blame implies that in order for an agent to be blameworthy for failing to perform some act (ϕ) in a given circumstance, it must be true that the agent has a reason to ϕ in this circumstance. Consequently, if we are to maintain the view that blameworthiness comes in degrees and also accommodate the reason condition of blame, it seems that our account should also allow reasons to come in degrees. That is, if an agent can only be blameworthy to some degree in a given case, then given the reason condition of blame, it is natural that this degree of blameworthiness should be explained by reference to the degree to which the agent has a reason in that case. (Note that I am not referring here to the commonly accepted idea that the *strength* of reasons may vary. Rather, I am suggesting that we reject the view that reasons are an "all or nothing" concept. Consequently, an agent may, for example, have a strong reason to ϕ, but only have this reason to a small degree.) The framework for determining blameworthiness should thus accommodate agents being blameworthy to *some degree*, and also having a reason *to some degree.*

What would justify the claim that reasons come in degrees? There might be different explanations (or arguments) that can be offered for this claim.[3] In this paper I will explore what I consider to be a promising explanation.[4] This explanation is based on two claims: first, reasons are constrained by what it is possible for the agent to do, and second, it may be possible *to some degree* for an agent to do something. The first claim is related to the well-known "ought implies can" principle, though this principle is usually taken to be solely concerned with an agent's capacities.

[2] The connection between reasons and blame can be explained by a more fundamental connection: If X has most reason to ϕ, then we tend to think (at least when X has the opportunity and access to information etc.) that X *ought to* ϕ (where "ought" is understood as an all-things-considered notion and a "reason" is understood as a pro tanto term). This connection between what we have reason to do and what we ought to do, offers a connection between reasons and blame. As Williams points out, blame operates in the mode of "ought to have" (1989:40). So if we can say, given certain external conditions such as opportunity and access to information, *X ought to* ϕ, we can also say, given these same conditions, *X will be blameworthy for failing to* ϕ.

[3] While they do not discuss the idea of reasons coming in degrees, Espinoza and Peterson (2011) offer some arguments for why we should think that moral duties and obligations come in degrees. These arguments might also support the idea that reasons come in degrees.

[4] I believe it is promising because of its explanatory power, its preservation of our moral conceptual world and its coherence with the plausible view that blameworthiness is closely related to our capacities. I will return to these virtues in Section 5.4.

There are however two senses in which A's possibilities constrain A's reasons. The first is related to A's capacities, the second is related to A's opportunities. Thus when assessing whether A has a reason to ϕ, we must investigate both her capacities and her opportunities. I will refer to the claim that in order for an agent to have a reason to ϕ, an agent must have certain capacities and opportunities, as the "CO condition on reasons". I will discuss this condition in detail in the next section.

Once we have accepted the CO condition on reasons, the claim that an agent may have a capacity or opportunity to some degree completes the explanation for why reasons come in degrees. That is, if an agent has the capacities or opportunities relevant to the CO condition *to some degree*, we should conclude that she may only have the associated reason to that degree. The claim that capacities come in degrees can be supported by Smith's account of rational capacities. According to Smith, the degree to which an agent has a (rational) capacity is determined by the degree of similarity between the actual world and certain possible worlds in which an agent systematically exercises this capacity. His view therefore provides a framework for understanding those cases where agents appear to have partial capacities. While Smith doesn't accept the CO condition on reasons[5] (and thus the view that reasons come in degrees), I extend his picture in order to produce a framework which allows us to assess whether, and to what degree, an agent has the capacities required for the satisfaction of the CO condition.

I also demonstrate how this same framework can be used to assess the extent of opportunity an agent has. As I shall explain, the distance between the actual world and certain worlds in which an agent systematically ϕ's, also informs us about the extent to which an agent has the opportunities required for the satisfaction of the CO condition. In this way, I provide a framework for understanding reasons as phenomena that come in degrees (and a related procedure for measuring this degree). Moreover, given the reason condition of blame, this framework also provides us with a non-arbitrary way of determining the degree to which an agent can be blameworthy in a given case.

This paper is devoted to developing the framework described above. To do this I begin by arguing for the CO condition on reasons. In Section 5.3 I outline Smith's discussion of rational capacities, and extend this discussion to account for the capacities and opportunities required for the satisfaction of the CO condition. In Section 5.4 I demonstrate how this framework can be used to justify claims about the degree to which an agent has a reason, and the degree to which an agent can be responsible (in the blameworthiness sense) in a given case.

[5] Smith at least denies the formulation of the CO condition that I will specify in this paper. He and Pettit claim that we should understand reasons in terms of what an idealised version of oneself (Å) would desire the unidealised self (A) to do. Though they claim that we should think of Å as an *advisor* rather than an *exemplar*, their view does not imply that reasons are sensitive to individual capacities. Rather, they claim that "it certainly seems possible that Å could desire that A ϕ in C and yet, due to A's own incapacities – remember, we are not supposing that A herself has and exercises all of the capacities that ensure that her desires conform to the principles of reason, only that Å has and exercises these capacities – A might actually be . . . incapable of coming to believe that this is so" (2006:150).

5.2 The CO Condition

Roughly speaking, the CO condition on reasons implies that in order for an agent to have a reason to perform some act (ϕ), it must be possible for the agent to ϕ. I claimed above that there are two relevant senses in which it must be possible for an agent to ϕ, if she is to have a reason to ϕ: the first relates to her capacities, and the second relates to her opportunities. It is a common claim that reasons are constrained by one's capacities. Streumer (2007), for example, asks us to consider Fred, who is completely and irreversibly paralysed. Fred hears someone screaming for help outside his house. As Streumer concludes, the idea that Fred has a reason to go outside and help this person is implausible because Fred *cannot* help the person.

The CO condition must be concerned with more than mere physical capacities if it is to be plausible however. If it merely implied that an agent must be physically capable of ϕ-ing in order to have a reason to ϕ, a bird physically capable of tapping on the window with its beak in order to warn me of danger, may have a reason to do so. A theory ascribing such reasons would be implausible however. In order for the CO condition to be plausible it should claim that in order for an agent to have a reason to ϕ, it must be possible that the agent ϕ *from the motive of her reason*. Specifying the CO condition in this way avoids the implausible implication that agents who can physically ϕ but cannot grasp the considerations that count in favour of ϕ-ing, have reasons to ϕ nonetheless. Interpreting the condition in this way also ensures that reasons are able to influence and guide those for whom they are reasons, and thus serve a *practical* purpose. And a practical purpose seems to be the appropriate purpose for a theory of reasons to serve.[6]

In addition to being capable of ϕ-ing from the motive of one's reason, one will need to have the opportunity to utilise this capacity, if the CO condition is to be satisfied. This is because if I have the capacity to apply first aid, but find myself stuck in a traffic jam, it is not in any real sense possible for me to apply first aid to the child who has fallen off his bike ten blocks away.[7] It is not possible because I have no opportunity to utilise my capacity to apply first aid. Consequently, the CO condition implies that I do not have a reason to apply first aid to this child. I must have the opportunity and capacity to ϕ from the motive of my putative reason[8] if it is to be genuinely possible that I ϕ from the motive of my reason and the CO condition is to be satisfied.[9]

One further point here is that if someone cannot currently ϕ, but has the opportunity and capacity to get themselves into a state where they can (and have the

[6] For the view that our moral concepts ought to serve a practical purpose, see Dennett (1984:155); Skorupski (2007:89–90).

[7] I am assuming here that by the time I could get there, the child would no longer need first aid.

[8] I say "putative reason" because if an agent lacks the relevant capacities or opportunities, she will, ipso facto, not have this "reason".

[9] Specifying that an agent must have these opportunities also helps to ensure that a theory of reasons serves a practical purpose (because an agent will only have reasons on which her opportunities allow her to act).

opportunity to) ϕ, then it seems that there *is* a real sense in which it is possible for them to ϕ. For instance, suppose a mother is currently incapacitated by depression, and so cannot presently care for her child. Suppose also however, that she is capable of taking medication (and that she has access to this medication, i.e. she has the opportunity to take it) and that this medication will remove her depression and reinstate her capacity to care for her child. It seems that in such a case, despite the woman's present incapacity, it is still possible for her to care for her child because if she would only take the medication (that she has, and is capable of taking) then she would be capable of caring for her child. We should thus interpret the CO condition in a way which allows agents who are currently incapable of ϕ-ing – but who have the opportunity and capacity to get into a state where they have the opportunity and capacity to ϕ – to have reasons to ϕ.

Given these clarifications, we are now in a position to formulate the CO condition with more precision. It is the claim that: if A is to have a reason to ϕ, then A must have the opportunity and capacity to ϕ from the motive of her putative reason, or she must have the opportunity and capacity to get herself into a state where she has the opportunity and capacity to ϕ from the motive of her putative reason.

An important aspect of the CO condition is that it can be satisfied to some degree, which will result in the agent having a reason to that degree. For an example where an agent's limited opportunity results in her having a reason to some degree, consider again the mother who is capable of taking medication which will remove her depression and reinstate her capacity to care for her child. If there is limited opportunity for the mother to get hold of the required medication, she will have a reason to care for her child to a smaller degree than if she could easily access the medication. For an example where an agent's limited capacity results in her having a reason to some degree, consider an acrophobe whose fear of heights cripples her in such a way that it is very difficult for her to think clearly or respond to important considerations when exposed to great heights. Suppose also that she is locked in a room on the tenth story of a building with her co-worker, who is contemplating suicide on the room's balcony. As the agent's acrophobia diminishes the degree to which she has the capacity to venture onto the balcony and dissuade her co-worker from jumping, she will only have a reason to help save her co-worker to *some degree*. Moreover, due to the reason condition of blame, blaming the agent for failing to act on this reason can be appropriate just to the degree that the agent has this reason. In the next section I will construct a framework for determining this degree. In Section 5.4 I will return to this last example in order to demonstrate how the framework I have constructed yields information about particular cases.

5.3 Capacities and Possible Worlds

Smith (2004) notes that a rational capacity manifests itself in a whole raft of possibilities, rather than single possibilities. This is because a capacity consists in intrinsic qualities which will cause the agent to exercise this capacity in a whole raft of circumstances very similar to the agent's actual circumstances. Consequently,

when determining whether an agent has a rational capacity to, for example, answer a certain scientific question in circumstances C, we must investigate whether she answers in a raft of circumstances very similar to C. However, the similar circumstances or possible worlds which we need to investigate are ones which share a similarity of a certain kind. In Smith's terminology, we need to "zero in" on what makes it true that an agent has the capacity to answer the scientific question. Consequently, the possible worlds we need to investigate are those where we hold constant the qualities that are both intrinsic to the agent and relevant to her answering the scientific question. For the sake of the example, let's suppose that these intrinsic qualities consist in a person's possession of a particular scientific knowledge base. If the agent answers in a raft of those possible worlds where we hold constant her knowledge base, this will establish that the agent has the *particular* scientific knowledge base which constitutes the capacity to answer the question.

Note that on the notion of capacity presented here, an agent may answer the question in the actual world, through luck, say, despite the absence of her capacity to do so. For instance, suppose I have no idea of the answer to the scientific question, but the answer somehow pops into my head shortly after I am asked. As my answering was a matter of luck, I do not have the particular scientific knowledge base which will cause me to systematically answer in those possible worlds where we hold constant my knowledge base. The fact that I do not systematically answer in these worlds is evidence that I lack the particular scientific knowledge base which constitutes the capacity to answer the question, regardless of whether I answer in the actual world. (While it may sound odd in ordinary English to claim that an agent who answered correctly lacked the capacity to answer correctly, I am using "capacity" in the technical sense outlined above, where the notion of a capacity is attached to a whole raft of possibilities, rather than single possibilities.)

We can extend Smith's method for determining whether an agent has certain rational capacities to those capacities required for the satisfaction of the CO condition on reasons. I argued above that in order for the CO condition to be satisfied, the agent must have both the opportunity and the capacity (i) to ϕ from the motive of her putative reason, or alternatively (ii) to get herself into a state where she has the opportunity and capacity to ϕ from the motive of her putative reason. The following three steps provide a procedure for determining whether an agent has the capacity specified by (i), i.e. the capacity to ϕ from the motive of her putative reason. For the remainder of this section I will refer to this capacity merely as "the capacity to ϕ", but note that this should be understood as: "the capacity to ϕ from the motive of her putative reason".

1. First identify which particular intrinsic qualities constitute the capacity to ϕ. This will be determined by examining the features of people who reliably ϕ.[10]

[10] The features I suggest here should by no means be taken as a definitive list. The features that appear intuitively to constitute a capacity may not be the features which do in fact constitute this capacity. In order to reach a more definitive conclusion, empirical research about the causal factors involved in motivated action would need to be consulted.

Intuitively, it appears that the possession of particular information and particular states of physical, motivational and rational functioning are causal factors involved in an agent's ϕ-ing.[11]

2. Then, hold constant those qualities that are intrinsic to the agent and relevant to the agent's capacity to ϕ. If the suggestion in the previous step is plausible, then we should hold constant the agent's possession of information and her states of physical, motivational and rational functioning. While holding constant the agent's possession of information and her states of physical, motivational and rational functioning, examine close possible worlds.
3. If an agent ϕ-s in a whole raft of these possible worlds, then this will demonstrate that she has the *particular* intrinsic qualities which constitute the capacity to ϕ (i.e. the particular information and particular states of physical, motivational and rational functioning).

These three steps spell out how we can determine whether an agent has the capacities required for the first way in which the CO condition can be satisfied. Note also that this framework allows for an agent to have a capacity to some degree. Step 2 specifies the possible worlds we must examine when determining whether an agent has a capacity to ϕ. While the qualities that are both intrinsic to the agent and relevant to her ϕ-ing are held constant in these possible worlds, other intrinsic features of the agent may be altered. If the agent only systematically ϕ's (and thus demonstrates

[11] We can demonstrate the plausibility of this claim by considering an agent who has a reason to save a drowning child (because this will save the child's life). In order for the agent to ϕ from the motive of her reason in such a case, the agent must possess certain information. More particularly, the agent must know about her reason. The agent could not rescue a child from the motive of her reason without the belief that the child is drowning (note here that such knowledge need not be knowledge that the agent could articulate. Instinctive knowledge about one's reasons may be sufficient for an agent to ϕ from the motive of her reason). Secondly, in order for an agent to ϕ from the motive of her reason, the agent must possess a certain state of physical functioning. More particularly, the agent must be physically able to ϕ. One cannot rescue a drowning child from the motive of her reason without the physical ability to swim or carry the child. Thirdly, in order for an agent to ϕ from the motive of her reason, the agent must have a certain state of motivational functioning. That is, the agent must be able to be motivated by this reason. One cannot rescue a child from the motive of one's reason, if the recognition of this reason cannot motivate the agent. Lastly, in order for the agent to ϕ from the motive of her reason, the agent must have a certain state of rational functioning. More particularly, the agent must recognise the relevant consideration *as* a reason. We should note here that there is a sense in which the agent could rescue the child *because* "the child is drowning", without the agent recognising that "the child is drowning" is, or provides, a reason to rescue the child. This may occur if the consideration "the child is drowning" connects with an agent's motivations through different means, e.g. through conditioning, or because it satisfies another of the agent's desires, such as a desire for praise. In such cases, the rescue mission may be caused by the agent's recognition of the consideration "the child is drowning". Such cases do not however constitute acting from the motive of one's *reason* (they at least do not constitute acting from the motive of the reason "the child is drowning"). To act from the motive of one's reason, the agent must be motivated by the consideration because of its believed status as a reason. Consequently, one cannot rescue a child from the motive of her reason without recognising that "the child is drowning" is, or provides, a reason to rescue the child.

her capacity to ϕ) in possible worlds where her other intrinsic features have been substantially altered, then she may only have the capacity to ϕ to a small degree. This is because the more the actual world has to be altered – in this first way – in order for the agent to systematically ϕ, the less the agent has a capacity to ϕ.

There is also a second way in which the actual world may have to be altered if the agent is to systematically ϕ. The more that *non-intrinsic* features of an agent must be altered in order for her to systematically ϕ, such as perhaps the experiences she has or the people she meets, the more distant the worlds where the agent systematically ϕ's become to the actual world. If these worlds are very distant in this second way to the actual world, then the agent will have only a small degree of *opportunity* to ϕ. The second way in which the actual world may need to be altered in order for the agent to systematically ϕ thus concerns the degree of opportunity that an agent has to ϕ. In summary, the degree to which we have to alter the actual world in order for the agent to systematically ϕ (while holding constant qualities that are both intrinsic to the agent and relevant to her ϕ–ing), reflects both the degree of capacity and the degree of opportunity an agent has to ϕ.

The degree of similarity shared by the closest worlds in which the agent systematically ϕ-s (i.e. closest to the actual world) and the actual world, will determine the degree to which the CO condition is satisfied, and thus the degree to which the agent has the associated reason. As discussed above, if these worlds are very dissimilar or distant to the actual world, the degree to which the agent has the capacity and/or opportunity to ϕ may be small. Moreover, this implies that the degree to which the CO condition is satisfied, and the degree to which the agent has a reason to ϕ, may also be small.

As was also stated above, there is a second way in which the CO condition can be satisfied: when an agent has the capacity and opportunity *to get herself into a state* where she has the opportunity and capacity to ϕ from the motive of her putative reason.[12] As I have already spelled out the three-step method for determining whether an agent has such a capacity and opportunity, I will not repeat these steps here. Instead I will return now to my earlier example in order to further explore what this framework tells us about an agent's reasons and blameworthiness in a given case.

5.4 An Example

If we reject the standard view of reasons as an "all or nothing" concept and accept the account outlined above, an agent will have a reason to ϕ to a degree which is proportional to the degree of her *opportunity* and *capacity* to ϕ from the motive

[12] Note that In order to assess whether the CO condition is satisfied in this second way, we will already have to know which epistemic, physical, motivational and rational states count as states in which an agent's ϕ-ing can be motivated by her putative reason. If we are unsure whether a given state (that an agent has the capacity to get herself into) is a state in which an agent's ϕ-ing can be motivated by her putative reason, then we will need to follow the method outlined above for assessing the first way in which the CO condition is satisfied, before concluding that the CO condition is (or is not) satisfied.

of her reason (or alternatively, to the degree to which she has the opportunity and capacity to get herself into a state where she has the opportunity and capacity to ϕ from the motive of her reason). That is, we now have a framework which allows us to determine whether, and to what degree, an agent has a reason. If an agent has a reason to ϕ, she must either (i) systematically ϕ from the motive of her putative reason in a whole raft of close possible worlds where we hold constant certain qualities, or alternatively (ii) there must be an array of close possible worlds (where we hold constant different qualities) in which the agent has got herself into a state where she has the opportunity and capacity to ϕ from the motive of her putative reason. The degree of similarity that the worlds specified in (i) or (ii) share with the actual world will determine the degree to which the agent has the reason.

To see how this account provides an adequate explanation of those cases where agents have partial capacities, I will return to an earlier example. Consider again the acrophobe who is locked in a room with her co-worker. Suppose that when we hold constant those features which are intrinsic to the agent and relevant to her helping her co-worker (from the motive of her reason), we discover that she only systematically helps her co-worker in those possible worlds where her attitudes and levels of fear have been substantially altered.[13] In this case, the closest worlds in which she systematically helps her co-worker (i.e. closest to the actual world) are very dissimilar to the actual world, and thus the degree to which she has a reason to help her co-worker will be small. This will also imply that the degree to which she can be blameworthy for failing to help her co-worker will be small. On the other hand, suppose that the closest worlds in which the acrophobe systematically helps her co-worker are worlds which are almost unchanged, but the acrophobe has practised her relaxation techniques or has taken the anti-anxiety medicine that she has in her purse. In this case, the degree to which she has a reason to help her co-worker will be higher. Moreover, the degree to which she can be blameworthy for failing to help her co-worker will also be higher.

As this example demonstrates, the framework spelt out in this paper allows us to properly explain cases where there seems to be a sense in which an agent has a reason, but due to the agent's partial incapacity there also seems to be a sense in which she does *not* have this reason. To say that such an agent has a reason to some degree seems preferable to stipulating some arbitrary level of capacity, above which the agent has a reason and below which she lacks this reason. This framework thus provides greater explanatory power than the view that reasons are an "all or nothing" concept.

Another virtue of this framework is that it supports an account of blameworthiness which can track individual capacities. This is a particularly important feature of the framework, given the close connection between blame and capacities. As Honoré explains:

[13] Where we assume that the acrophobe's attitudes and levels of fear are not part of the intrinsic properties which constitute the capacity to help her co-worker from the motive of her reason (and thus have to be held constant when determining if the acrobphobe has this capacity).

> Before imposing sanctions or attaching blame, law and morality require… that in the circumstances the agent had the capacity to reach a rational decision about what to do. When this capacity is present, blame for bad behaviour is appropriate and criminal liability may, depending on the state of the law, be imposed… different degrees of blame, punishment and censure correspond to the extent to which the agent's capacity is impaired. (Honoré 1998:138)

In addition to accommodating the connection between blame and capacities, and degrees of blameworthiness, the framework supports a plausible relationship between reasons and blame. In this way, the framework also preserves our "moral conceptual world".

5.5 Conclusion

In conclusion, I have offered a framework for determining blameworthiness which accommodates agents being blameworthy to *some degree*, and also having a reason *to some degree.* A crucial claim in the paper is that we should reject the standard way of understanding reasons as an "all or nothing" concept. The rejection of the standard view allows the framework to support a plausible relationship between reasons and blame, while also preserving a concept of blame that accommodates degrees of blameworthiness. Understanding reasons as coming in degrees also provides greater explanatory power than the standard view in cases where an agent's partial incapacity makes it unclear whether or not the agent has a reason. The framework outlined in this paper can explain such cases in a non-arbitrary way because it need not imply that an agent either has or lacks a reason. Rather, a partial incapacity may imply that an agent has a reason to some degree. As there is a reason condition on blame, the framework also informs us about the degree to which an agent can be blameworthy in a given case. Finally, the framework accounts for the complexity and variability of our capacities when attributing both reasons and blameworthiness to agents.

References

Dennett, Daniel. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge, MA: MIT Press.

Espinoza, Nicolas, and Martin Peterson. 2011. "Some Versions of the Number Problems Have No Solution." *Ethical Theory and Moral Practice* 13(4):439–51.

Heuer, Ulrike. 2010. "Reasons and Impossibility." *Philosophical Studies* 147(2):235–46.

Honoré, Tony. 1998. "Being Responsible and Being a Victim of Circumstance." Reprinted in 1999, *Responsibility and Fault*, 121–42. Oxford: Hart Publishing.

Pettit, Philip, and Michael Smith. 2006. "External Reasons." In *McDowell and His Critics*, edited by Cynthia Macdonald and Graham Macdonald, 140–68. Oxford: Blackwell.

Skorupski, John. 2007. "Internal Reasons and the Scope of Blame." In *Bernard Williams*, edited by Alan Thomas, 73–103. New York, NY: Cambridge University Press.

Smith, Michael. 2004. “Rational Capacities.” In *Ethics and the A Priori: Selected Essays on Moral Psychology and Meta-Ethics*, edited by Micheal Smith, 114–35. New York, NY: Cambridge University Press.

Streumer, Bart. 2007. “Reasons and Impossibility.” *Philosophical Studies* 136:351–84.

Williams, Bernard. 1989. “Internal Reasons and the Obscurity of Blame.” Reprinted in 1995, *Making Sense of Humanity: And Other Philosophical Papers*, edited by Bernard Williams, 35–45. Cambridge: Cambridge University Press.

Chapter 6
Please Drink Responsibly: Can the Responsibility of Intoxicated Offenders Be Justified by the Tracing Principle?

Susan Dimock

Abstract Normally, reduced mental capacities are thought to reduce responsibility. This is how, for example, the criminal defences of insanity/mental defect and automatism work. Persons who lack the capacity, due to a disease of the mind, to understand the nature of their actions or that they are wrong, and persons who lack the capacity for voluntary control of their bodily movements, are thought to lack essential capacities for criminal responsibility. These criminal law practices exemplify the intuition that responsibility tracks capacity, a view sometimes called "capacitarianism". An equally widely held view, however, operates to restrict the commitment to capacitarianism: the intuition that when a person is responsible for his own reduced mental capacities, the exculpatory value of those reduced capacities is discounted or even extinguished. The paradigmatic case of such reduced capacities involves persons who have reduced their capacities (of understanding, foresight, knowledge, advertence or self-control) through voluntary intoxication. This intuition is also reflected in the criminal law practices of most jurisdictions. Whether such practices can be justified is the topic of this paper. My conclusion will be that, even if we accept the capacitarian intuition and its limited application to those who are responsible for their own reduced capacities, the legal instantiation of them in criminal practice is unjustified. My argument is that our treatment of intoxicated offenders is not, in fact, supported by these commitments, contrary to what many theorists and jurists think.

6.1 Introduction

Normally, reduced mental capacities are thought to reduce responsibility. This is how, for example, the criminal defences of insanity/mental defect and automatism work. Persons who lack the capacity, due to a disease of the mind, to understand the nature of their actions or that they are wrong (the two pronged McNaughten rule of insanity), and persons who lack the capacity for voluntary control of their

S. Dimock (✉)
York University, Toronto, Canada, ON
e-mail: dimock@yorku.ca

N.A Vincent et al. (eds.), *Moral Responsibility*, Library of Ethics and Applied Philosophy 27, DOI 10.1007/978-94-007-1878-4_6,

bodily movements, whether because they are suffering from a disassociative state caused by a blow to the head, a severe emotional trauma, being swarmed by bees, or sleep-walking, and so who satisfy the legal definition of automatism, are thought to lack essential capacities for criminal responsibility. These criminal law practices exemplify the intuition that responsibility tracks capacity, a view sometimes called "capacitarianism".

An equally widely held view, however, operates to restrict the commitment to capacitarianism: the intuition that when a person is responsible for his own reduced mental capacities, the exculpatory value of those reduced capacities is discounted or even extinguished. The paradigmatic case of such reduced capacities involves persons who have reduced their capacities (of understanding, foresight, knowledge, advertence or self-control) through voluntary intoxication. Even if a person lacks the capacity to understand the nature or consequences of his actions, the capacity to know that his actions are wrong, or the capacity to exercise voluntary control over his bodily movements, if that lack of capacity can be traced to his own voluntary intoxication, the lack of capacity will not exonerate. This intuition is also reflected in the criminal law practices of most jurisdictions (common law as well as civil law systems). Whether such practices can be justified is the topic of this paper. My conclusion will be that, even if we accept the capacitarian intuition and its limited application to those who are responsible for their own reduced capacities, the legal instantiation of them in criminal practice is unjustified. I am not arguing that capacitarianism and its intuitive exception in cases involving the destruction of one's own capacities is itself unjustified: I think them quite correct. Rather, my argument is that our treatment of intoxicated offenders is not, in fact, supported by these commitments, contrary to what many theorists and jurists think. Although I will use Canadian legal practice as my example in what follows, the problems identified here should apply equally to other legal jurisdictions as well.

6.2 Components of Criminal Liability: Elements of a Crime

I am interested in the criminal responsibility of intoxicated offenders. In order to understand for what intoxicated offenders are responsible, let's remind ourselves of the basic elements of criminal offences. First, offences consist of an *actus reus*, a prohibited act, omission, or result. Persons are responsible only for committing or bringing about the *actus reus* of a crime. In order to commit the *actus reus* of an offence, the person's conduct must be voluntary and conscious. There must be an acting agent whose conduct is under her voluntary and conscious control. The voluntariness component of the *actus reus* is well summed up by Vertes J. in R. v. Brenton (R. v. Brenton 1999:para. 42).

> The concept of voluntariness... represents the fundamental principle of our criminal law that no act can be regarded as criminal unless it is a voluntary act: R. v. Stone (1999) 134 C.C.C.(3d) 353 (S.C.C.) (at 421). Thus it is an aspect of the *actus reus*. It is the minimal requirement that acts must be conscious acts. There must be a mind capable of exercising the will-power to do the physical act that represents the crime. There must be a state of

> awareness on the part of the actor that he or she is doing the act. One can phrase this principle in numerous ways but the point is that voluntariness is an aspect of all crimes since all crimes must have an *actus reus* [case references omitted].

Larry Alexander calls this The Voluntary Act Principle: "there can be no criminal liability in the absence of a voluntary act", and says that "it is the law in all Anglo-American jurisdictions that no person is guilty of a crime unless she commits a voluntary act" (Alexander 1990:85).[1] This requirement has been incorporated in the Model Penal Code (Model Penal Code §2.01).

Second, all criminal offences have as an essential component some mental element known as the *mens rea* of the offence. To commit a crime a person must have the required "guilty mind" or fault element. The mental element required for different offences varies. Sometimes a person must act with intention, knowledge, foresight, or wilful blindness to commit a particular crime. These are all subjective mental states, concerning what the actual defendant knew, foresaw, or intended at the time of committing the *actus reus*. When no particular *mens rea* condition is specified, recklessness typically suffices. Recklessness is a subjective awareness of or advertence to the risk that one's conduct might constitute or bring about the *actus reus* of the offence charged, and continuing despite that risk. Many criminal law theorists, as well as some judges, have insisted that all crimes must have a subjective *mens rea* requirement. Indeed, many theorists think that what distinguishes true crimes from regulatory offences or torts is just that the latter can be committed without subjective fault, whereas criminal conduct can only be committed with subjective awareness of the harm one does. But no criminal jurisdiction has operationalized such a blanket requirement, and all legal systems allow that some crimes might be committed with only the objective fault of (gross or criminal) negligence. Negligence does not concern what was in the mind of the accused at the time of action, but rather is determined by a standard of what a reasonable person in the same circumstances would have known, foreseen, intended or appreciated, and gross negligence is a marked departure from what a reasonable person would have done in the circumstances. A person can bring about a prohibited result without any subjective fault, being unaware of the risk that his conduct may produce that result, negligently; here fault is based on the fact that he did not know or attend to facts that a reasonable person in the same circumstances would have attended to, and criminal fault consists of precisely this negligent failure to attend to relevant features of his circumstances, where that leads to a prohibited outcome.

Voluntariness is also a component of *mens rea*. Again Justice Vertes sums up the law well.

> Voluntariness is also linked to the *mens rea* component. It is a principle of fundamental justice that every criminal offence punishable by imprisonment must have a *mens rea* component. There must be at least some minimal mental state as an essential element of the crime. . . . This requirement of *mens rea* may be satisfied in different ways. There may be a subjective or an objective approach. An offence could require proof of a state of mind

[1] Alexander (1990), quoting Dressler (1987:65).

> such as intent, recklessness or wilful blindness. The *mens rea* requirement could, on the other hand, be satisfied by evidence of negligent conduct (a "marked departure") measured against an objective standard (what should have been in the accused's mind had he or she proceeded reasonably) [case references omitted]. But there must be some *mens rea* element, as there must be an *actus reus*, since otherwise the offence would be one of absolute liability, something that in criminal law violates both s. 7 and s. 11(d) of the Charter: R. v. Hess, [1990] 2 S.C.R. 906. [Section 7 of the Canadian Charter of Rights and Freedoms guarantees the right to life, liberty and security of the person, and the right not to be deprived thereof except in accordance with principles of fundamental justice. Section 11(d) guarantees to those accused of criminal offences the presumption of innocence.] Voluntariness is the basic constituent element of the *mens rea* requirement. The conscious doing of an act (being the *actus reus*) encompasses the intention to do it and therefore constitutes the minimal *mens rea* for general intent offences.

Finally, there is a required connection between the *actus reus* and *mens rea* constituting a crime. This is typically referred to as the principle of contemporaniety. *Actus non facit reum, nisi mens sit rea*: the intent and the act must concur to constitute the crime. That is, the *mens rea* must be directed to the *actus reus* itself, and they must occur, if not simultaneously, at least in tight temporal relation. No act can be criminal without a mental element of fault.

These elements of criminal offences establish a set of basic requirements that criminal law systems must satisfy if they are to conform to principles of fundamental justice. First, criminal law typically requires subjective fault. Exceptions to this rule are just that: exceptional. Moreover, the Crown or prosecution bears the burden of proving every element of the offence, to the criminal law standard of proof beyond a reasonable doubt. This is required by the presumption of innocence. A person accused of a criminal offence has nothing to answer for unless and until the prosecution has proved that he committed the *actus reus*, with the required *mens rea*, duly related to the prohibited outcome. (Duff 2007; Duff 2009; Tadros 2005; Horder 2004; Gardner 2007).

6.3 Responsibility, Liability and Defences

Understanding the elements of a crime is important because it allows us to distinguish four different kinds of answers one might raise to an allegation of criminality: exemptions (youth, insanity, diplomatic immunity), justifications (self-defence, lesser evils), excuses (duress, necessity), and denials that one satisfied all of the elements of the offence. Exemptions and exculpatory defences of justification or excuse have been the subject of considerable philosophical and legal analysis, but they are not relevant to our purposes, because no one thinks voluntary intoxication exempts persons from meeting the demands of the criminal law, nor that it functions as a justification or excuse. If intoxication is relevant to responsibility and liability, it is so because it negates an essential element of the crime alleged.

Often persons answer a criminal charge by raising a *reasonable doubt as to some element of the offence*. Here they are not accepting responsibility for the conduct constituting a crime and raising considerations that block the normal transition from

responsibility to liability, as in the case of justifications and excuses. Nor are they denying that they are answerable to the criminal jurisdiction in question or that they are responsible agents fit to answer at all, as in the various kinds of exemptions. Rather, they are denying that they committed the crime at all. That a person cannot be criminally responsible for or liable to punishment on account of conduct not meeting all of the essential elements of the crime charged seems obvious. Suppose a person is charged with a homicide offence, murder or manslaughter. It is a complete answer to the charge that the person one is charged with killing did not die. Bringing about the death of another person is an essential component of the *actus reus* for homicide. It is every bit as much a full answer to deny that one's conduct satisfies the essential *mens rea* or fault elements of the offence charged, though not so intuitively obvious. If an offence requires a particular *mens rea*, that one do something knowingly, or with a specific intention, or being reckless with respect to the risk that one's conduct will bring about a prohibited result, and one does not have the required mental state, one is simply not guilty of the crime. To be guilty of an offence one must satisfy all of its essential elements, and the prosecution must be able to prove that one does satisfy all of those elements, beyond a reasonable doubt. Thus many defences take the form of raising a reasonable doubt as to a required element of the offence charged. Mistake and accident often function this way, defeating either the *mens rea* or *actus reus* elements of the crime.

Automatism is a defence in this same family, in that it functions to defeat an essential element of the crime. Automatism is a defence available to those who "act" without consciousness and voluntary control over their bodily movements. Paradigmatic cases involve persons who bring about a prohibited result while sleepwalking, while in a disassociative state caused by a blow to the head or a severe emotional trauma, or in the throes of a seizure or other condition that undermines their ability to control their bodily movements. If I strike a person, causing him injury, while in any of these states, I will not be held criminally responsible or made liable for that harm. Whether one is in a state of automatism is determined by the degree to which one is in conscious and voluntary control of one's bodily motions and actions. Conceptually, it does not depend upon the cause of one's automatism. Yet the law makes two distinctions within the class of automatistic conduct, based on the cause of the automatism: it distinguishes sane from insane automatism, depending on whether the cause is a disease of the mind as understood in our insanity or mental defect jurisprudence; and within the category of non-insane automatism, it treats differently automatism caused by voluntary intoxication. If automatism is a result of voluntary intoxication, it cannot function to raise a reasonable doubt as to an essential element of the crime charged, for a wide range of crimes. This is problematic, given that automatism seems to defeat both the minimal mental elements of the *actus reus* (voluntary and conscious control), as well as any subjective *mens rea* requirements the offence might have, since there is nothing the person knows, intends, foresees, or adverts to while in a state of automatism. Even if the fault condition is objective negligence, it seems unfair to hold a person responsible for failing to meet the standard of care a reasonable person would meet in circumstances where she is incapable of exercising any care at all. Thus it should raise a reasonable doubt

as to both *mens rea* and *actus reus*. As we shall see, in cases of automatism caused by voluntary intoxication, this is not the case, and the automatism defence is denied to intoxicated offenders, even if they would otherwise qualify for it.

6.4 Voluntary or Self-Induced Intoxication

Intoxication, if relevant to questions of criminal responsibility and liability, seems to be so because intoxication can affect a person's mental states, and more especially can diminish a person's capacities that are relevant to responsibility and liability. Indeed, it seems pretty plausible to think that intoxication might be relevant to the mental states of persons at the time they commit an offence, and so it may be used to raise a reasonable doubt as to whether the person had the required *mens rea* for the crime charged. If crimes require subjective *mens rea* – knowledge, intention, malice, planning, deliberation, foresight, awareness, advertent recklessness or wilful blindness – then intoxication should be relevant to assessments of guilt, because it is relevant to an essential element of crimes. An accused person should be able to use the fact of intoxication as an evidential basis for claiming that she lacked the *mens rea* of the offence and so to raise a reasonable doubt as to fault, whatever the fault elements may be. As the New Zealand Court of Appeal put it: "Drunkenness is not a defence of itself. Its true relevance by way of defence, so it seems to us, is that when a jury is deciding whether the accused has the intention or recklessness required by the charge, they must regard all the evidence, including evidence as to the accused's drunken state, drawing such inferences from the evidence as appears proper in the circumstances" (R. v. Kamipeli 1975:616). The success of such arguments will vary depending upon the degree of intoxication and the specific fault requirements of the offence charged. It might be more plausible to assert that a person lacked knowledge of a particular circumstance or foresight of a particular result because she was intoxicated, for example, than to argue that she was not aware of the nature of her act or that she did not act recklessly. But for any crimes requiring subjective fault, whether intent, purpose, knowledge, willful blindness, or advertent recklessness, it seems that intoxication ought to be considered in determining whether the Crown has proved all the elements of the crime beyond a reasonable doubt. (McCord 1990) Intoxication is relevant because it might prevent the formation of the fault element of some crimes. Not surprisingly, then, the intoxication defense began as a common law defense in recognition of the fact that an accused person may be sufficiently intoxicated not to have the subjective *mens rea* for the crime charged.

On the other hand, both policy considerations and general conditions underlying exculpation tend to lend support to a regime in which voluntary intoxication cannot be used to relieve persons of criminal responsibility and punishment. This tension has been played out in law in a number of ways.

The leading modern case is *Director of Public Prosecutions vs. Beard* (1920), a rape and felony murder case heard by the House of Lords. In Beard, the House of Lords developed the distinction between general intent and specific intent offences,

and they limited the defense of intoxication to only specific intent crimes. The distinction between the two rests, as the names suggest, upon a difference with respect to the *mens rea* element of the offences. The House of Lords began with the common sense relevance of intoxication to *mens rea*. "That evidence of drunkenness which renders the accused incapable of forming the specific intent essential to constitute the crime should be taken into consideration with the other facts proven in order to determine whether or not he had this intent" (D.P.P. v. Beard 1920: 501–02). This suggests that intoxication should function as the basis of a claim that the accused lacked the required *mens rea* to be guilty of the offence charged, and thus as an evidentiary consideration of relevance to proof of the essential elements of the crime.

But Lord Birkenhead's articulation in Beard of the principles to govern considerations of intoxication in criminal cases has been taken to require something quite different. It has been taken to mark a general distinction between crimes of general and specific intent. Although there is no canonical formulation of the distinction, the *mens rea* for general intent crimes is only a conscious performing of the prohibited act, whereas crimes of specific intent require a further purpose beyond the mere intention to perform the prohibited act, an ulterior purpose, or a fault element greater than recklessness (D.P.P. v. Majewski 1976). Examples of general intent crimes include all forms of assault, manslaughter, mischief, and breaking and entering. Examples of specific intent offences include robbery, breaking and entering with the intent to commit an indictable offence, assault to resist or prevent arrest, murder, theft, aiding and abetting a crime, attempted crimes, and being an accessory after the fact. Several scholars and jurists have challenged the dichotomy between general and specific intent offences as artificial, unprincipled, and indeterminate. (Quigley 1987a, b, c; Colvin 1981; Dickson J. in Leary v. The Queen 1978) While I agree with these critiques of the common law, they are not my purpose here.

The purpose of introducing the general/specific intent distinction was to limit the range of cases in which a "defense of intoxication" could be raised. However specific intent is understood, intoxication may raise a reasonable doubt over whether or not the accused had the specific fault element required. In this sense, intoxication may be a defense available to an accused person charged with a specific intent crime. But intoxication cannot be used to raise a reasonable doubt as to *mens rea* in the case of general intent offences (Leary v. The Queen 1978). Indeed, in many jurisdictions, intoxication can be substituted for the *mens rea* of every general intent crime, so that the prosecution may satisfy its burden of proving fault simply by proving intoxication (D.P.P. v. Majewski 1976; Leary v. The Queen 1978). This substitution rule has been widely debated, and my purpose is not to rehearse that debate or my reasons for thinking the substitution rule is unjust. (Dimock 2009) What matters here is that a person will be held responsible for the crime if it is proved that she committed the *actus reus*, even if she lacked the *mens rea* that would otherwise be required for quilt, if she committed the act while intoxicated. This will be so, moreover, even if the intoxication is extreme enough to raise a reasonable doubt as to the voluntariness of the conduct, even if it is so extreme, that is, as to produce a state of automatism. It is the justification of this practice that I now wish to examine.

6.5 The Fault of Intoxication

The rationale for the substitution rule and its subsequent limiting of the use of intoxication to raise a defense to a criminal charge is defended on a number of different grounds. But all, I suggest, are based upon a particular conception of the intoxicated offender, one which cannot be sustained given how voluntary or self-induced intoxication is understood by the courts. The image is familiar: a person drinks alcohol, consumes narcotics or prescription medication, mixes the two, either with the intent to become impaired, or at least being entirely reckless with respect to whether impairment results. He is blameworthy for his impairment, and so no wrong is done to him if he is held responsible for harms he then does as a result of incapacitating himself.

The willingness to substitute intoxication for the *mens rea* of every general intent offence or to find in intoxication an alternate basis of criminal liability stems from the conviction that individuals who become voluntarily intoxicated are morally blameworthy for doing so. When Canadian courts have addressed the constitutionality of the restrictions on the intoxication defense, many of the Justices have referred to the blameworthiness of becoming voluntarily intoxicated. Thus concerns about whether the restriction violates considerations of fundamental justice, which are centrally concerned with not punishing the morally innocent, are thought to be more easily met, because intoxicated offenders are not morally innocent, just in virtue of their self-induced intoxication. As Lamar C.J. said, intoxicated offenders are not "completely blameless" (R. v. Penno 1990). As he put it: "By voluntarily taking the first drink, an individual can reasonably be held to have assumed the risk that intoxication would make him or her do what he or she otherwise would not normally do with a clear mind."

It would seem, then, that the intoxication rule which allows voluntary intoxication to be substituted for the normal *mens rea* of general intent offences can be justified under a widely accepted principle of responsibility, namely, the "tracing" principle identified by John Martin Fischer and Mark Ravizza. (Fischer and Ravizza 1988) As is well-known, Fischer and Ravizza have argued that it suffices for responsibility that a person exercised guidance control over his action, which control just requires that the person's action issue from their own moderately reasons responsive mechanism. Applied to our case, the position would be that even if the intoxicated offender lacks the capacity to be reasons-responsive after becoming intoxicated, he may still be responsible for his conduct while intoxicated if becoming intoxicated was, at an earlier time, something over which he exercised guidance control. As Fischer and Ravizza put it, "When one acts from a reasons-responsive mechanism at *T1*, and one can be reasonably expected to know that so acting will (or may) lead to acting from an unresponsive mechanism at some later time *T2*, one can be held responsible for so acting at *T2*." (Fischer and Ravizza 1998:50) They describe their view as "a 'tracing' approach: when an agent is morally responsible for an action that issues from a mechanism that is not appropriately reasons-responsive, we must be able to trace back along the history of the action to a point (*suitably related to the action*) where there was indeed an appropriately reasons-responsive

mechanism." (Fischer and Ravizza 1988:50–51) The semicompatibilism of Fischer and Ravizza is not only widely adopted among responsibility theorists, but it seems just the kind of theory that the criminal law needs to undergird its practices. So if it has the resources to explain and justify the treatment of intoxicated offenders as responsible for their offences, even when lacking the *mens rea* typically required for them, by tracing back to a choice for which they were responsible (the choice to become intoxicated), then the tracing principle will justify the intuitions about capacities with which we began. The lack of capacity typically exculpates, but not when the incapacity can itself be traced back to a choice for which the incapacitated agent is responsible. Not surprisingly, Fischer and Ravizza use an intoxicated driver as their example to establish the intuitive force of the tracing principle.

Whether the tracing principle can justify holding intoxicated offenders responsible for their offences will depend, however, on whether the choice to become intoxicated is something for which they can be held responsible, and on whether that choice is "suitably related to the [criminal] action". As Fischer and Ravizza note, the tracing principle requires that the person to be held responsible for some act at *T2* must be reasonably expected to know that his conduct at *T1* will or might lead to action on the basis of an unresponsive mechanism at *T2*. He must be able to anticipate, that is, that his earlier choice to consume an intoxicant might lead to acting on unresponsive mechanisms in the future. As they say, "the degree of likelihood employed by the tracing approach would need to be context-relative". (Fischer and Ravizza 1998:50, fn 21) In the context of criminal liability, the degree of likelihood must be above the *de minimus* range if our practices are to satisfy the requirements of fundamental justice. I argue that they do not.

6.6 What Makes Intoxication Voluntary or Self-Induced?

The greatest injustices worked by our current intoxication rules actually stem from the way that voluntary or self-induced intoxication is defined. The problem lays with the responsibility conditions for voluntary intoxication. The image of the intoxicated offender that is relied upon is of a person who imbibes significant qualities of drugs, alcohol or both, over an extended period of time. Even if reaching a state of intoxication or impairment in not intended, any reasonable person engaging in such behaviour must anticipate that impairment might result from his actions. Indeed, it seems simply inconceivable that the person himself did not, at some point in the process of consuming the intoxicants, advert to the risk of impairment that his consumption might have. If this was the only type of person caught by our intoxication rules, they would not likely generate the controversy they have, and they could be justified by the tracing principle. (Although there might still be worries about those whose ingestion of intoxicants is the result of addiction, if addictive desires are not themselves moderately reasons-responsive.) But this is not the only type of person who is deemed to satisfy the conditions for voluntary intoxication.

The conditions on involuntary intoxication are stringent. A person cannot plead involuntary intoxication just because he did not intend to become intoxicated, or

if he did not know or even foresee that his conduct would produce intoxication. In order to be involuntary, a reasonable person could not have foreseen intoxication resulting from the person's conduct. If a person ingests or consumes anything which *he knows or ought to know is an intoxicant*, he cannot plead involuntary intoxication (R. v. King 1962). In Canada and elsewhere, there is a rebuttable presumption that impairment from alcohol or drugs is voluntary. Only if there is a reasonable doubt as to a defendant's ability to appreciate and know that he would or might become impaired, an inability for which he is completely without fault, can his subsequent intoxication exonerate. If a person voluntarily consumes alcohol or a drug which he knew or had any reasonable ground for believing might cause him to be impaired, then he cannot use his impairment to escape liability for a crime he then commits, even if he did not intend to become impaired. Among the authorities appealed to in support of this position is Justice O. W. Holmes, who wrote in *The Common Law* that: "As the purpose is to compel men to abstain from dangerous conduct, and not merely to restrain them from evil inclinations, the law requires them at their peril to know the teachings of common experience, just as it requires them to know the law." Applied to intoxication, it has been "taken as a matter of 'common experience' that the consumption of alcohol may produce intoxication and, therefore, 'impairment' . . ., and I think it is also to be similarly taken to be known that the use of narcotics may have the same effect" (R. v. King 1962).

The presumption that persons know that consumption of intoxicants is inherently dangerous and risks impairment is rarely overcome. This is so, even if the resulting intoxication is highly improbable, as long as it is the result of ingesting known intoxicants. Thus even if someone has a completely unpredictable reaction to a small amount of marijuana, for example, or someone else puts drugs into the person's alcoholic drink without his knowledge, his resulting intoxication is not involuntary because it is in part due to his ingesting substances that are known by reasonable people to be intoxicants (R. v. Brenton 1999; R. v. Talock 2003). The lack of fault for the offence due to involuntary intoxication can only exonerate if the intoxication itself was without fault, and fault for intoxication is in practice established merely by the consumption of anything reasonably known to be an intoxicant. Case after case demonstrates that the real test is proof of the voluntary consumption of intoxicants; once voluntary consumption is proved, persons are expected to have common knowledge about the dangers of their consumption, and so recklessness is simply inferred rather than proven. As Clackson J. summed up, "self-induced intoxication . . . means the accused voluntarily consumed a substance which he knew or *ought to have known* was an intoxicant and appreciated or should have appreciated that he risked becoming intoxicated" (R. v. Huppie 2008:para. 23; R. v. Chaulk 2007). As another Canadian judge put it, "the law in Canada requires that the Court find that the accused consumed the alcohol. A successful *mens rea* defence would involve evidence that the act of drinking was prompted by threats or mistake and thus not an act of volition. Examples that come to mind of this sort might be: the accused was forced against his will to drink alcohol; or a third party slipped alcohol into the drinks of an unknowing accused; or the alcohol, having been transferred

by a third party into an orange juice container, the accused was unaware he was drinking alcohol" (R. v. Thompson 1993:para. 31).

Put positively, if a person knows or ought to know that what he or she is voluntarily consuming is an intoxicant then any resulting state of intoxication is itself deemed to be voluntary. As Dyer J. put it, after reviewing the jurisprudence on voluntary intoxication, "a trial judge in dealing with voluntary consumption of drugs [must] consider whether an accused person knew or had any reasonable grounds for believing that such consumption might cause him to be impaired. In so doing, I do believe the Court should not permit negligence or carelessness on the part of an accused to become a defence. I think persons who take drugs or drink voluntarily are required to act reasonably in taking them and are to be taken to reasonably understand the likely results of taking them in most cases" (R. v. Kataria 2005: para. 102).

In practice, however, the possibility of taking drink or drugs responsibly seems to be ruled out from the start. The courts have ruled, for example, that a person cannot claim to know from experience how long a sleep-aid medication takes to work to escape liability for impaired operation of a motor vehicle, though such arguments would seem to suggest that the accused did not appreciate the risk that he would become impaired while he was in care and control of a vehicle. They have ruled that "It is not necessary for the Crown to show that the appellant knew the degree to which he would be affected. The Crown need only show knowledge that [the intoxicants in question] could affect him and that in fact they did so" (R. v. Jensen 1991:para. 25). Generally, the courts take it as a matter of common knowledge that drugs, whether illicit, prescription or unregulated such as cold medications or sleeping aids, should not be taken with alcohol or at a dosage higher than prescribed or recommended on the packaging, and such knowledge will suffice for proof that any resulting impairment is voluntary.

That voluntary intoxication can result from negligence and yet be the standard of criminal fault in any general intent offence involving crimes against the person has been codified in Canadian criminal law, which establishes gross negligence as a standard of penal fault, and creates an irrebuttable presumption that a person who becomes extremely intoxicated has that fault: he "departs markedly from the standard of reasonable care generally recognized in Canadian society and is thereby criminally at fault" (Criminal Code of Canada, s. 33.1 (2)). Thus if a person commits a general intent offense while voluntarily intoxicated, or a crime against a person even if so severely intoxicated as to be acting involuntarily, he will be deemed to have the *mens rea* necessary for conviction. Negligently becoming intoxicated suffices for criminal fault (R. v. Chaulk 2006).

The case of R. v. Brenton illustrates the problem with this approach. Mr. Brenton shared a marijuana cigarette with his landlady one evening after work. He had prior experience with the drug, though was not a habitual user, and had never had an unusual reaction to it. He smoked the joint hoping to relax so that he could sleep. Instead, he had an extreme and both statistically and subjectively unpredictable reaction to the drug, producing a state of automatism, in which he assaulted his landlady.

He was convicted at trial of the charges, even though the trial judge had a reasonable doubt as to the voluntariness of his conduct or his mental state at the time of the commission of the crimes. On appeal, Mr. Brenton argued that his conviction should be overturned because his intoxication was not voluntary. "The appellant argued, at trial and on appeal, that it cannot be said that he intended to become intoxicated or should have known that he would become intoxicated given the relatively small amount of marijuana he ingested. His purpose for smoking the marijuana was to relax so as to help him sleep. Therefore, it was argued, the result was an unintended and unexpected outcome and thus tantamount to non-voluntary intoxication." Justice Vertes rejected this argument: "I cannot agree with the appellant's submission. Generally speaking, if the ingestion of a drug (or alcohol) is voluntary and the risk of becoming intoxicated is within the contemplation or should be within the contemplation of the individual, then any resulting intoxication is self-induced. Involuntary intoxication is generally confined to cases where the accused did not know he or she was ingesting an intoxicating substance (such as where the accused's drink is spiked) or where the accused becomes intoxicated while taking prescription drugs and their effects are unknown to the accused. This is fairly basic law" (R. v. Benton 1999:paras. 30 and 31). Thus it is voluntary consumption of intoxicants, rather than any subjective appreciation that impairment might result, that is the fault of intoxication, fault that can be substituted for the *mens rea* of any general intent offence. Voluntary intoxication, then, can result from negligence without any subjective awareness of the risk of impairment or subsequent criminality.

That negligence is the fault criterion for voluntary intoxication is extremely important. Many cases involve the combination of alcohol and other drugs, whether banned substances, prescription medications or over-the-counter products. Many drugs in the latter two groups contain warnings against mixing them with alcohol, but the warnings actually suggest that sleepiness might result. While it might be negligent to drive an automobile or operate dangerous equipment in circumstances where one does or ought to anticipate extraordinary tiredness being experienced, it is not at all clear that such a warning suffices to establish that a reasonable person combining a small amount of alcohol with such drugs would or ought to anticipate that he might become violent and actually do or threaten harm to another. The fact pattern in Brenton, involving the consumption of at most half a marijuana joint, which produced a completely unexpected reaction, leading to a loss of voluntary control and violence, raises equal concern. Such an outcome was not subjectively foreseeable, nor, I would argue, even objectively foreseeable. Even a reasonable person would not have anticipated the resulting danger. Nonetheless, the trial judge felt compelled to find the accused guilty, even though entertaining a reasonable doubt as to the voluntariness of Mr. Brenton's conduct. (Criminal Code of Canada s. 33.1 (1)–(3)).

We can now see that the intoxication rules as applied are not actually supported by the tracing principle. The first problem, related to how voluntary intoxication is understood in the law, means that the substitution rule fails to meet the condition of the tracing principle that one can be reasonably expected to know that ingesting an intoxicant will (or may) lead to acting from an unresponsive mechanism

at some later time (impairment or loss of capacity). This is so because a person could satisfy the legal requirements for voluntary intoxication without satisfying the requirement that he can reasonably be expected to know that he will later become impaired to the point of being unresponsive, or becoming violent.

There is, moreover, a second way in which our intoxication rules fail to be supported by Fischer and Ravizza's tracing principle. According to their account, we must be able to trace back along the history of the subsequent criminal action to a point (*suitably related to the act of ingesting intoxicants*) where there was an appropriately reasons-responsive mechanism operating. But in fact the acts at *T1* producing intoxication are not "suitably related" to acts of violence that might be committed at *T2*. It is not reasonable to expect that persons engaged in the act of consuming alcohol or drugs can be reasonably expected to know that their conduct at *T1* creates a risk of criminality at *T2*. This is because the risk of criminality from intoxication is in fact so low, objectively speaking, that to say that a person ought to recognize that becoming intoxicated at *T1* constitutes a significant risk of a loss of capacity producing criminality at *T2* is to say that a person ought to believe what is objectively false.

Yet many people seem to think that such a risk is foreseeable from intoxication. Indeed, many judges and academic commentators suggest that becoming voluntarily intoxicated is necessarily reckless. The claim must be a necessity claim if the substitution rule is to be acceptable, because voluntary intoxication creates an irrebuttable presumption of criminal fault for general intent crimes. The problem with this line of argument should now be apparent. The law characterizes voluntary intoxication as intoxication resulting from the consumption of substances the person knew or *ought to have known* were intoxicants, and that he *knew or ought to have known* might cause impairment. Thus the law makes negligence sufficient for voluntariness, rather than the subjective standard of recklessness. But the issues here are too important to settle by semantics, so let's examine whether the ingestion of intoxicants to the point of impairment is necessarily reckless conduct. Only if ingestion of intoxicants really does create a foreseeable risk of criminality will the reasons-responsive mechanisms leading to the decision to consume intoxicants be suitably related to the subsequent criminal conduct (produced as it may be by non-reasons-responsive mechanisms at the time of its occurrence) so as to allow us to trace responsibility from the earlier time to the later. The answer to the question – is the ingestion of intoxicants to the point of impairment necessarily reckless? – is no.

It is surely problematic that a legally innocent action can be a conclusive and irrebuttable basis of criminal fault, indeed, fault for a vast range of crimes, including crimes the commission of which is punishable by life imprisonment. Many judges and legal theorists attempt to meet this concern by claiming that the fault element is the recklessness that necessarily attaches to the act of becoming intoxicated itself. Thus recklessness is the fault element, rather than intoxication per se, and it is satisfied by every voluntarily intoxicated offender. This was the tack taken in Majewski, and it has since been followed by many Canadian judges.

This line of thought has attracted many, in part because it would provide a principled way of distinguishing between general and specific intent crimes, something

that has otherwise seemed ad hoc, uncertain and unprincipled. The idea is that specific intent crimes have *mens rea* conditions beyond mere recklessness but general intent crimes require only recklessness. If that was true, and becoming voluntarily intoxicated is necessarily reckless, then the substitution rule would be acceptable. Proof of intoxication would suffice as proof of recklessness and so *mens rea*. But the argument trades on an ambiguity concerning "recklessness." While many general intent offences have recklessness as *mens rea*, recklessness as the fault element of crimes is more constrained than recklessness outside the law. To be guilty of a crime, a person must be reckless *with respect to the criminal act or result* specifically. It is not a crime to be reckless per se. Legal recklessness implies foresight of specific consequences or an awareness of or adverting to risks with respect to a prohibited act or result, and a decision to assume that risk. This presents a dilemma. On one horn, we must suppose that every person who becomes voluntarily intoxicated is reckless with respect to every prohibited act or result that falls within the bounds of general intent offences. This should function as a *reductio ad absurdum* of this way of understanding the argument; we cannot infer such foresight or advertence merely from the fact that a person became voluntarily intoxicated. On the other horn, we must admit that the recklessness evidenced by voluntary intoxication is not of the same kind as reckless in law, and therefore even if intoxicated offenders are reckless in some sense, it is not the sense required for criminal fault.

We should not, however, accept the general claim that becoming intoxicated is necessarily reckless or otherwise morally faulty, even understood in the non-legal sense of recklessness. Recklessness can be inferred from intoxication in some circumstances, but only given additional facts. A person who routinely becomes violent when he drinks alcohol, for example, could reasonably be expected to foresee the danger that he might assault someone if he drinks and so can be considered reckless with respect to that danger. But equally conceivably, a person could take all reasonable steps to avoid harming others while intoxicated.

In an earlier paper, I offered the following counter-example to the claim that voluntary intoxication is necessarily reckless.

> Mary has just achieved some very important personal goal. She has defended her Ph.D., secured a long-sought promotion, or earned tenure. She decides that a party is in order, at which friends, family, and colleagues will celebrate her achievement. But she is a cautious and responsible person, who knows the courts have been doing funny things with respect to host liability specifically and intoxication law in general. Since she also knows that she is likely to imbibe a lot of alcohol at the party, she takes all reasonable precautions. She arranges for the party to be catered by licensed professionals so that there will be no incidents involving food preparation and safety. She ensures that accommodations are made for people who must drive to the party at an inexpensive motel within walking distance from her home and has taxi chits available for local celebrants. Finally, she arranges to have a trusted friend shadow her. Her friend's job is to stay sober and ensure that she does not do anything untoward should she become intoxicated. All is going smoothly on the big night. Her friend is conscientiously performing his duty. Her guests are heartily enjoying themselves, and she raises her glass to every toast made in recognition of her achievement [Miss Manners notwithstanding]. Then, once she is clearly intoxicated, the caterer serves a rare seafood delicacy unknown to many of her guests. Her friend eats a morsel, has an allergic reaction, and dies. At that point, she is agitated and distressed, and intoxicated.

> There seems no reason to accept that she was reckless in becoming intoxicated or that she thereby demonstrated fault sufficient to establish *mens rea* for every general intent offence. Yet according to Canadian law, she has the *mens rea* for all general intent crimes involving violence or bodily interference. (Dimock 2009)

This example still seems to me to stand as a counter-example to the necessity claim, and to provide reason for rejecting the position that proof of voluntary intoxication can ground an irrebuttable presumption of recklessness.

Yet the claim that becoming intoxicated is necessarily reckless persists. It is claimed that it is common knowledge that intoxication is inherently dangerous. Such a claim is especially problematic in a country like Canada, where in most provinces the state itself sells the vast majority of the alcohol available to consumers. In the latest year for which there are statistics (April 1, 2007–March 31, 2008), beer and liquor stores in Canada sold $18.8 billion worth of alcoholic beverages, or 222.9 million litres. Provincial and territorial governments realized a net income of $5.2 billion from the sale of liquor and related products (e.g. liquor licenses). Our governments have not, then, told citizens not to consume alcohol or pharmaceuticals because of the risk of criminality, nor have they made general prior rules against such behavior. To the contrary, members of our society are inundated with advertisements extolling the pleasures of alcohol (including from government-owned liquor retailers) and promoting the ideal of better living through pharmaceuticals. Our governments certainly have not pointed, with a few notable exceptions such as impaired driving, to specific dangers that consuming intoxicants might produce. Instead, our government urges that we "drink responsibly".[2] If the very consumption of alcohol is necessarily criminally reckless, however, then at the very least the government is complicit in that fault, perhaps so much so that it has lost the right to hold citizens to account for it (Tadros 2009).

Chester Mitchell has argued that judicial or legislative treatment of voluntary intoxication as itself criminally negligent or reckless cannot be sustained on the scientific evidence. It is simply not true that the vast majority of people who ingest drugs or drink, even to the point of intoxication, are thereby reckless in doing so. As he says: "For almost all persons, the probability of their drug consumption causing a serious crime is too low to qualify as recklessness. Violent crimes are rarely compared to the common incidence of intoxication or drug-impairment. Furthermore, the relationship between intoxicants and crime is much more problematic, subtle, and indirect than is usually assumed" (Mitchell 1988:78). The claim that becoming intoxicated is itself criminal recklessness simply lacks empirical support.

> For ordinary intoxication, the evidence suggests a probability of resultant harm considerably lower than the level needed for criminal recklessness. The chances of drug users turning to violent or serious crime because of intoxication are at best remote. Most North Americans take alcohol and millions regularly become intoxicated without putting themselves or others at serious risk. Unfortunately, most crime-alcohol studies merely state the proportion of

[2] The Liquor Control Board of Ontario uses "Please drink responsibly" as a regular feature of it community messaging, including advertising on its bags. This suggests that the government thinks it is possible to drink responsibly.

> known criminality involving drug use. This may be as high as 50%. But to judge whether intoxication is reckless we require the opposite statistic, namely the portion of drunken events that involve serious criminal activity. This figure is certainly below 1%. (Mitchell 1988:88–89)

If these facts are correct, as they certainly are in Canada, the eventual criminal act is simply too remote and unforeseeable from the act of becoming intoxicated for intoxication to constitute *mens rea* for it, or for the two acts to be "suitably related" so as to bring them within the scope of the tracing principle. For most people who imbibe intoxicants, it is simply not true that criminal actions fall within the ambit of the act of becoming intoxicated (Gough 1996). The two acts are not "suitably related" such that a person should foresee criminality resulting from the use of intoxicants; not even a reasonable person can foresee connections that are statistically insignificant.

References

Alexander, Larry. 1990. "Reconsidering the Relationship Among Voluntary Acts, Strict Liability, and Negligence in Criminal Law." *Social Philosophy & Policy* 7:84–104.
Colvin, Eric. 1981. "A Theory of the Intoxication Defense." *Canadian Bar Review* 59:750–79.
Criminal Code of Canada, s. 33.1 (2).
Dimock, Susan. 2009. "The Responsibility of Intoxicated Offenders." *Journal of Value Inquiry* 43:339–68.
Dressler, Joshua. 2006. *Understanding Criminal Law*. 4th ed. New York: Lexis.
Duff, R.A. 2007. *Answering for Crime: Responsibility and Liability in the Criminal Law*. Oxford: Hart Publishing.
Duff, R.A. 2009. "Strict Responsibility, Moral and Criminal." *Journal of Value Inquiry* 43: 295–313.
Fischer, John Martin and Mark Ravizza. 1988. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.
Gardner, John. 2007. *Offences and Defenses*. Oxford: Oxford University Press.
Gough, Stephen. 1996. "Intoxication and Criminal Liability: The Law Commission's Proposed Reforms." *Law Quarterly Review* 112:335–55.
Horder, Jeremy. 2004. *Excusing Crime*. Oxford: Oxford University Press.
McCord, David. 1990. "The English and American History of Voluntary Intoxication to Negate Mens Rea." *The Journal of Legal History* 11:372–95.
Mitchell, Chester N. 1988. "The Intoxicated Offender — Refuting the Legal and Medical Myths." *International Journal of Law and Psychiatry* 11:77–103.
Quigley, Tim. 1987a. "Specific and General Nonsense?" *Dalhousie Law Journal* 11:75–125.
Quigley, Tim. 1987b. "A Shorn Beard." *Dalhousie Law Review* 12:167–94.
Quigley, Tim. 1987c. "Reform of the Intoxication Defense." *McGill Law Journal* 33:1–48.
Tadros, Victor. 2005. *Criminal Responsibility*. Oxford: Oxford University Press.
Tadros, Victor. 2009. "Poverty and Criminal Responsibility." *Journal of Value Inquiry* 43:391–413.

Cases

D.P.P. v. Beard [1920] 14 Cr. App. R. 159 (HL)
D.P.P. v. Majewski [1976] 2 All E.R. 142 (HL)
Leary v. The Queen [1978] 1 S.C.R. 29

R. v. Brenton [1999] 28 C.R. (5th) 308 at 320 (N.W.T.S.C.)
R. v. Brenton [1999] N.W.T.J. No. 113
R. v. Chaulk [2006] N.S.J. No 407
R. v. Chaulk [2007] N.S.J. No. 301, 2007 N.S.C.A. 84
R. v. Huppie [2008] A.J. No. 989
R. v. Jensen [1991] B.C.J. No. 4060
R. v. Kamipeli [1975] 2 N.Z.L.R. 610 at 616 (C.A.)
R. v. Kataria [2005] B.C.J. No. 2963
R. v. King [1962] 133 C.C.C. 1
R. v. Penno [1990] 2 S.C.R. 865
R. v. Talock [2003] 41 M.V.R. 269 (Sask. C.A.)
R. v. Thompson [1993] O.J. No. 3249

Chapter 7
The Moral Significance of Unintentional Omission: Comparing Will-Centered and Non-will-centered Accounts of Moral Responsibility

Jason Benchimol

Abstract It is reasonable to assume that much wrongdoing for which agents are generally thought blameworthy occurs by way of unintentional omission. In this paper, I explain why certain will-centered accounts of moral responsibility tend to struggle to provide convincing explanations of the theoretical basis for judgments of blameworthiness in cases of unintentional omission. To provide such explanations, these will-centered accounts typically rely upon a "tracing strategy", according to which an agent's blameworthiness for an unintentional omission necessarily presupposes that it is a casual result of some prior blameworthy intentional choice she apparently made. I argue that this sort of appeal to the tracing strategy, upon further inspection, produces distorting implications for the way we ordinarily think about the conditions of legitimate moral criticism in cases of unintentional omission. I conclude by identifying a peculiar assumption that defenders of these will-centered accounts of moral responsibility appear to adopt and that, once rejected, renders the volitionalist's appeal to the tracing strategy unnecessary for purposes of explaining the conditions of blameworthiness for unintentional omission. The upshot of my investigation is rather modest, but it does remain unclear just what advantage, if any, will-centered accounts of moral responsibility enjoy over their rival non-will-centered accounts.

7.1 Introduction

Because so much theorizing about moral responsibility has taken place against a background concern about whether the truth of determinism could be compatible with free will, we should not be surprised that many philosophers have had quite a lot to say about the conditions of moral responsibility for actions and for their consequences. This is presumably because the underlying issue is typically articulated as one that concerns whether the truth of determinism would entail that all of our

J. Benchimol (✉)
Department of Philosophy, University of Washington, Seattle, WA, USA
e-mail: jdbench@uw.edu

N.A Vincent et al. (eds.), *Moral Responsibility*, Library of Ethics and Applied Philosophy 27, DOI 10.1007/978-94-007-1878-4_7,

actions and their consequences are inevitable occurrences given certain antecedent states of affairs that obtained long before our births, and are thus not suitably "free" or "up to us" in the way typically thought necessary to serve as bases for the distinctively moral forms of evaluation involved in praise and blame. Given the prevalence of this particular way of orienting the issue, it is easy to understand why compatibilist and libertarian philosophers have been concerned primarily with explaining how our actions and their consequences can indeed still be "up to us" in the requisite sense to ground legitimate attributions of moral responsibility.

But perhaps this attentiveness to the question of how we can be morally responsible for what we do has inadvertently led to the construction of theories of moral responsibility that are too preoccupied with the conditions of responsible *action*, theories which in turn fail to provide convincing explanations of how agents can be morally responsible for some of the things they *don't* do. Very often, we take certain intuitively morally significant inactions to be legitimate bases for moral criticism of agents. Think of a husband (call him "Mikael") who, after seven years of marriage, completely forgets his own anniversary date. Or, after arranging his schedule to meet a struggling student outside of his normal office hours, a college instructor (call him "Per") subsequently fails to recall the appointment and forgets to show up. Another example: a lifeguard on duty (call her "Abby") falls asleep on the job, during which time an inexperienced swimmer is pulled away from the shore by a strong undertow, and Abby doesn't assist him as she should. Mikael, Per, and Abby are all intuitively blameworthy not, it would seem, on the basis of any particular actions they have performed, but on the basis of certain morally significant inactions. This intuition is supported by noting that judgments that each of these agents has exhibited a morally criticizable measure of ill will or disregard appear to be made appropriate by their inactions. Moreover, it is plausible to suppose that these examples do not constitute rare or anomalous instances of blameworthy inaction. It is, instead, reasonable to assume that much blameworthy wrongdoing occurs by way of this kind of inaction. Any satisfactory general theory of moral responsibility should be able to offer a compelling explanation why judgments of blameworthiness and the kinds of responses associated with them are intuitively appropriate in such cases.

My principal aim in this paper is to explain why certain will-centered approaches to moral responsibility – theories that regard intentional choice or decision as a precondition of legitimate moral criticism – struggle to provide a convincing account of the conditions of blameworthiness for the kind of inaction that is of moral concern in the cases I just introduced. This failure reflects, I think, a lack of attention in the literature on moral responsibility to the moral significance of inaction, and this paper is intended to open avenues for further reflection on this topic. In the next section, I will explain the conception of moral blameworthiness that will guide my inquiry, and I will clarify the precise nature of the kind of inaction that is my present concern. Then, I will briefly discuss how a particular variety of will-centered approaches to moral responsibility have attempted to explain judgments of blameworthiness for this kind of inaction by appealing to a "tracing strategy". I will then subject the way these theories use the tracing strategy to intense critical scrutiny, illuminating different ways I think such use produces distorting implications for

the way we ordinarily think about the conditions of legitimate moral criticism in the cases of inaction with which I am concerned. In the final section of the paper, I will expose a peculiar assumption that I think defenders of these will-centered approaches to moral responsibility adopt, and that pinpoints one possible source of their struggle to convincingly explain a substantial set of our intuitive judgments about the conditions of moral blameworthiness. Because non-will-centered theories of moral responsibility can easily reject this assumption, they may be better equipped to provide compelling philosophical explanations of judgments of moral blameworthiness in the cases of inaction that are my chief concern. The upshot of my investigation is rather modest, but it does draw attention to how unclear it is just what advantage, if any, certain will-centered approaches to moral responsibility enjoy over non-will-centered approaches.

7.2 Moral Blameworthiness and Unintentional Omission

My general interest in this paper is in the conditions of moral blameworthiness for certain morally significant inactions. Minimally, moral blameworthiness entails the propriety of a judgment that ill will or disregard has been shown (where proper regard is legitimately expected). In addition, an agent's moral blameworthiness depends crucially upon the propriety of certain distinctly moral kinds of response on behalf of both the offender (e.g., feelings of guilt, remorse, and a desire to apologize and attempt reconciliation) and those offended (e.g., feelings of resentment or indignation, attitudes of disapprobation, and requests for justification and acknowledgement of fault). For convenience, I will sometimes refer to the former kinds of response under the collective heading of "guilt" and the latter kinds of response under the collective heading of "resentment". It is reasonable to suppose that the propriety of a judgment that ill will or disregard has been shown, and the propriety of both guilt and resentment indicates that the conditions of moral blameworthiness are satisfied.

Some philosophers understand "moral blameworthiness" as what T. M. Scanlon has called a "desert-entailing notion".[1] On this view, a correct judgment of moral blameworthiness entails (among other things) that the blameworthy agent deserves to suffer some loss in virtue of what she has done. In a similar vein, others have thought that moral blameworthiness presupposes that it would be morally fair or just to respond to the agent who is blameworthy by applying "informal sanctions" that incorporate distinctly retributive sentiments.[2] My conception of moral

[1] Scanlon (1998:274–75).

[2] See Watson (2004). Because some philosophers have assumed that blame carries a characteristic force that requires justification in order for its application to be morally fair, questions about the conditions of moral blameworthiness have sometimes been framed as fundamentally involving the conditions that must be met in order for it to be morally "fair" or "just" for some moral judge to impose sanctions upon a wrongdoer. For an important criticism of the idea that the characteristic force of blame is to be located in these overt sanctioning behaviors, see Hieronymi (2004).

blameworthiness implies neither that an agent who is blameworthy deserves to suffer in virtue of this fact, nor that subjecting her to informal sanctioning behavior is morally fair. By the same token, this conception clearly does not imply the propriety of only what Gary Watson calls "aretaic appraisals".[3] An appraisal of an agent's character or, to use Watson's own term, her "evaluative capacities" as defective or poor in some respect need not involve a specific judgment that ill will or disregard was shown, and unless one has some legitimate stake in the appraisal, need not make guilt or resentment appropriate. So it should be noted that the conception of moral blameworthiness I adopt goes substantially beyond what might be called "negative aretaic appraisal", yet does not go so far as to incorporate any substantive conclusions about the fittingness of overt responses that incorporate retributive sentiments. While any understanding of moral blameworthiness will surely presuppose a corresponding conception of blame, the understanding of moral blameworthiness I have adopted here seems to me to capture features of blame that are most central to that notion as we commonly understand it.

Now I must clarify what I have in mind when I speak of "morally significant inactions". The cases of morally significant inaction in which I am interested are those whereby an agent seems to have violated a legitimate moral expectation not by way of anything she has done, but by way of something she hasn't done. In such cases, it is the fact that the inaction appears to involve the violation of a moral expectation that distinguishes it as "morally significant", thereby raising questions about whether the inaction can be taken as an appropriate basis for moral criticism. Clearly, not every inaction is morally significant in a way that raises such questions. But the example cases I introduced in the previous section are intuitive cases of morally significant inaction, for Mikael, Per, and Abby all appear to have violated legitimate moral expectations in virtue of what they haven't done.[4] I will refer to morally significant inactions involving such violations as "omissions", in order to distinguish them properly from inactions that do not involve such violations and that are thus morally insignificant.[5]

The term "omission", as I have proposed to use it here, is neutral regarding an agent's mental condition vis-à-vis the violation that makes a particular inaction an omission. Therefore, an omission to do s can obtain with or without an agent's awareness that he does not do s. When such awareness is absent, we may call it a case of unintentional omission. Patricia G. Smith has recently offered a compelling analysis of unintentional omission which highlights the fact that all that is necessary and sufficient for an agent's unintentionally omitting to do s is, roughly speaking,

---

[3] Watson, op. cit.

[4] I think this much can be granted, though I recognize that some will balk at the idea that forgetting an anniversary could involve the violation of a moral expectation. For this reason, I have included the other two example cases which, I take it, rather obviously do involve such violations.

[5] Here I am appealing, somewhat roughly, to Patricia G. Smith's account of omission as it is defended in her papers, Smith (1990), and Smith (2005b). Smith's account is an extension and development of Joel Feinberg's original remarks on the concept of omission. See Feinberg (1984:159–61).

(1) that the agent did not do s, and (2) that, by not doing s, the agent violated a norm of conduct that specifies required behavior for her.[6] Smith explains that:

> there is no positive act and no thought or intention by which to distinguish omission from inaction in the case of unintentional omission. There is only a standard. . . of human conduct, the [violation] of which turns [inaction] into omission by [connecting] it to a particular agent within a specific context. It is this connection that ultimately provides the needed link to agency and responsibility.[7]

While it may be granted that Mikael, Per, and Abby each violate a moral expectation by way of their inactions, this would not by itself entail that each is morally blameworthy, for moral blameworthiness also requires that ill will or disregard was in fact shown.[8] We may, however, assume for the sake of argument that each agent has (somehow) shown a certain degree of ill will or disregard. We can imagine that each would, in virtue of his or her respective unintentional omission, appropriately experience responses characteristic of guilt. Moreover, those who have been neglected, forgotten, or otherwise overlooked would also appropriately experience responses characteristic of resentment. Yet, it is worth noting that nothing in the description of their respective cases thus far suggests the presence of any positive act, thought, or intention either to violate the relevant moral expectations or to show what would readily be recognized by those wronged as ill will or disregard. This leads to a puzzle, I think, because we commonly tend to assume that some such mental act – choice, intention, decision, etc. – is required in order for some state of affairs to be genuinely reflective of ill will or disregard on an agent's behalf. If a positive mental act is necessary in order for it to be true that ill will or disregard was shown, then it is not immediately clear how Mikael, Per, and Abby could be morally blameworthy for their unintentional omissions.

A satisfactory solution to this puzzle will provide a compelling explanation of how an intuitively blameworthy unintentional omission can be reflective of ill will or disregard. Circumstances are plentiful in which agents forget, overlook or otherwise simply fail to act as they should, and we ordinarily take many of these unintentional omissions to constitute grounds for judging that ill will or disregard has been shown. Such circumstances are illustrated by the examples involving Mikael, Per, and Abby. Moreover, unintentional omissions frequently provide occasions both for the omitter to legitimately experience guilt on the one hand, and for those neglected, forgotten, or otherwise overlooked to legitimately experience resentment, on the other. In such cases, our confidence in judgments to the effect that these agents are open to moral criticism for their unintentional omissions depends crucially upon being able to explain how their morally significant inactions can reflect ill will or disregard in the way necessary to underwrite a legitimate attribution of moral blameworthiness.

---

[6] This is admittedly an oversimplification of Smith's account, but for my purposes here, the oversimplification is justifiable.

[7] Smith (2005a).

[8] The violation of a moral expectation entails wrongdoing, but wrongdoing does not by itself entail blameworthiness. This much seems to me to be common ground. For a contrary view, see Norman O. Dahl (1967).

7.3 Volitionalism

In this section, I will briefly explain an approach that certain will-centered accounts of moral responsibility tend to adopt in order to explain how phenomena like unintentional omissions can be reflective of ill will or disregard, despite the fact that no particular mental act involving a choice, decision, or intention appears to be essential to their obtaining. My concern is with volitionalist approaches to moral responsibility, which insist that a condition of moral blameworthiness for some action, omission, or attitude, is that it stem from a conscious mental act called a "volition". Michael J. Zimmerman, a prominent defender of volitionalism, characterizes the notion of a volition as "a decision or choice … that some event occur, a decision which is accompanied by an intention that it (the decision) be causally efficacious with respect to the event in question."[9] Other notable defenders of volitionalism include Neil Levy and R. Jay Wallace. Levy provides a wonderfully concise articulation of the volitionalist's basic theoretical commitment when he says that "an agent is [morally] responsible for something (an act, omission, attitude, and so on) just in case that agent has – directly or indirectly – chosen that thing".[10] Wallace stresses the fundamental role choice plays in grounding an agent's blameworthiness when he claims that "the primary target of moral assessment … is the quality of choice expressed in what we do … Indeed, the degree of our moral fault is determined essentially by the quality of the choices on which we act, regardless of whether we succeed in achieving the ends fixed by these volitional states."[11]

It is reasonable to regard the volitionalist as fundamentally committed to two basic claims about the conditions of moral responsibility and blameworthiness. First, the mental activity involved in intentional choice is the kind of agency that opens one up to moral appraisal in the way required to ground attributions of moral responsibility in its most basic sense.[12] Second, an agent's blameworthiness fundamentally concerns the quality of his will as it is reflected in volitional states (I shall henceforth just use the term "choice" to refer to such states). Choices are what agents are principally morally responsible for. Whether an agent is morally blameworthy, then, depends upon whether the quality of the agent's will reflected in a choice is defective from the moral point of view.[13]

---

[9] Zimmerman (1988).

[10] Levy (2005:2, emphasis in original).

[11] Wallace (1994:128).

[12] Here, I assume that it is uncontroversial that volitionalists must recognize constraints imposed by basic conditions of "moral attributability"; these are conditions that must be satisfied in order for an agent to be morally responsible for some action, omission, attitude, etc. in the first place. Of course, satisfaction of these conditions is not by itself sufficient for moral blameworthiness.

[13] Wallace (op. cit.:132) seems to think that intentionally violating a moral obligation others accept is what constitutes showing ill will or disregard. For Zimmerman, the crucial element of a choice that makes it one for which the agent is culpable, and hence blameworthy, is an occurrent belief that the choice is itself morally wrong. See Zimmerman (op. cit.: 40).

Given this account of the conditions of moral responsibility and blameworthiness, how might a volitionalist explain intuitions that the agents in my example cases are morally blameworthy for their unintentional omissions? Consider Abby's case. The volitionalist picture implies that Abby is morally responsible for unintentionally omitting to assist the swimmer only if her inaction is reflective of some choice she has made, for such a choice is the kind of mental act necessary to make her inaction attributable to her as a basis for further moral appraisal. Likewise, this picture suggests that she is morally blameworthy only if a defective quality of her will (from now on, I will simply take "ill will" or "disregard" to stand for "defective quality of will") is reflected in this choice.[14] So far, the volitionalist picture cannot provide the requisite explanation since, as I have already pointed out, it is clear that no mental act that could possibly be called a choice occurs while Abby dozes and the swimmer struggles for help.

Both Wallace and Zimmerman give a generally clear explanation of how volitionalism should attempt to accommodate cases like Abby's, albeit their explanations differ in certain respects. Their common strategy, however, is to attempt to trace Abby's blameworthiness to some prior episode of choice.[15] The idea is that Abby is blameworthy only if her unintentional omission is either an explicitly foreseen or reasonably foreseeable consequence of some prior choice that reflected ill will or disregard.[16] Because an unintentional omission inherently does not involve any kind of positive mental agency that could possibly be called a choice, and because an agent's blameworthiness is fixed entirely by the quality of her will as it is reflected in her choices, the volitionalist picture implies that the moral significance of Abby's unintentional omission, then, consists entirely in what it reveals about the quality of her will as it was reflected in some prior choice. In other words, Abby's blameworthiness in this case is essentially indirect: the intuition that she is blameworthy on the basis of her unintentional omission must be fully cashed out in a judgment that

[14] I take it that the notion of a choice reflecting a particular quality of will is commonly recognized. As I will elaborate in Section 7.5, choices necessarily implicate an agent's evaluations of reasons and other intentional mental states, and the quality of an agent's will seems to be constituted by the quality of just these evaluative mental states.

[15] See Wallace (op. cit.:138–39), and Zimmerman (op. cit.:93). Wallace claims that in cases involving negligence or forgetfulness – e.g., cases like Abby's – "one may have to trace the moral fault to an earlier episode of choice." A concise statement of Zimmerman's alignment with the tracing strategy can be found in his claim that a question of responsibility and blameworthiness "for an omission arises *only* where there is an initial volition of which the omission in question is itself a consequence."

[16] One question that faces the volitionalist, then, concerns just *which* choice it is reasonable to think an agent's moral responsibility and blameworthiness for an unintentional omission are traceable *to*. Volitionalists face a further question of how rigidly to construe the kind of cognitive connection between a choice and a consequence in order for consequence to be a basis for legitimate blame. Wallace (p. 138) seems to understand this cognitive connection as involving the consequence's *reasonable foreseeability* from the agent's perspective at the time of prior choice, while Zimmerman opts for *explicit foresight*. For a fascinating discussion of the problems that the foreseeability constraint poses for the tracing strategy, see Vargas (2005) and a reply by Fischer and Tognazzini (2009).

her unintentional omission is (minimally) a reasonably foreseeable consequence of some prior blameworthy choice.

Of course, it is possible that no prior choice occurred to which Abby's intuitive blameworthiness can suitably be traced. If this is so, then volitionalists face a further question of how to respond. It is of course open to the volitionalist to claim that an intuition that a judgment of blameworthiness is appropriate in Abby's case is flatly unjustifiable if the attempt to trace her blameworthiness to a prior choice fails. In the next section, I will motivate what I take to be a strong case for thinking that this claim ought to be resisted. Indeed, this claim would warrant our acceptance only if no compelling alternative explanation of how an unintentional omission can reflect ill will or disregard is available.

7.4 Problems with the Volitionalist's Use of the Tracing Strategy

In this section, I will attempt to motivate strong skepticism about the volitionalist's capacity to provide a compelling explanation of moral blameworthiness in at least some important cases of intuitively blameworthy unintentional omission by illustrating certain troubling implications that flow from the way volitionalists use the tracing strategy. Some of these concerns are by no means novel, but I will hereby give them novel expression. I will not argue that the notion of tracing is itself philosophically problematic. The following remarks are intended to highlight only what I take to be unsettling about certain implications of the volitionalist picture insofar as it tends to suggest that a very natural conception of the role certain unintentional omissions play in grounding moral criticism is substantially misguided.

Here are the claims I hope to motivate in this section: (1) even when an unintentional omission is not plausibly seen as a result of a prior blameworthy choice, this does not by itself seem to make a judgment of blameworthiness inappropriate and in some cases may even make such a judgment appear more appropriate; (2) the volitionalist's use of the tracing strategy sometimes entails, rather paradoxically, that we should have the least confidence in judgments of blameworthiness in what otherwise appear to be some of the most intuitively obvious cases of blameworthy negligence and forgetfulness involving unintentional omission; (3) the volitionalist's use of the tracing strategy substantially distorts our common sense picture of the kinds of response that blameworthy unintentional omissions intuitively make appropriate; and (4) the volitionalist picture entails that moral criticism of an agent on the basis of unintentional omission alone is always unjustifiable, such that episodes of unintentional omission can never in themselves make a difference to the kind or degree of moral criticism of an agent that is legitimate.

I'll begin by noting one respect in which the volitionalist picture implies a very strong result. It is one thing to say that, for a significant range of cases of intuitively blameworthy unintentional omissions involving neglect, disregard, or forgetfulness, judgments of blameworthiness are justified entirely in virtue of the fact that these unintentional omissions are indicative of a prior blameworthy choice. It is quite

another to say that this is so for every unintentional omission that appears blameworthy. Returning to our examples, what exactly must we suppose Mikael previously chose to do? Did he notice the date of the anniversary approaching yet nonetheless choose not to set a reminder? Did Per think, at the moment he agreed to meet with the student, about writing the special appointment down in his day planner but then choose not to do so? Did Abby really notice that she was getting drowsy, and then choose not to go get a cup of coffee? Whatever plausibility each of these stories has, it may be at least as plausible to suppose that no such prior choices in fact occurred. For the distinct possibility remains that at these putative prior moments of choice none of these agents realized that he or she was in a situation in which such choices were available.

This possibility leads to the motivation behind claim (1). One problem for the volitionalist enters when it turns out that an unintentional omission appears to be traceable to just such a failure of awareness rather than to some choice. If such a cognitive failure causally explains an unintentional omission, then a judgment that the agent has exhibited a degree of morally criticizable ill will or disregard is rather obviously not automatically rendered inappropriate. Instead, this prior cognitive failure might only provide additional grounds for judging that ill will or disregard has been shown. This can be so if the prior cognitive failure in question can be seen not merely as a descriptive failure, but as a normative failure that implies the violation of further moral expectations that these agents be attentive to the presence of circumstances in which they may need to take action to avert future wrongdoing.[17] Such episodes of inattentiveness may themselves constitute grounds for judging that ill will or disregard has been shown, and may be what explains a subsequent unintentional omission instead of a choice. Unless the volitionalist is willing, at this point, to very implausibly assert that the mere failure of an initial attempt to trace blameworthiness to a prior choice implies that a judgment of blameworthiness is inappropriate, he will need to engage in still further tracing.

Of course, there is nothing that guarantees that the volitionalist will be successful in this further tracing. Suppose that what explains Abby's falling asleep is not some prior choice she made to avoid taking precautions against dozing, but is instead her failure to be attentive to her situation in a way that would have inclined her to realize that she needs to take action to ensure she can discharge her duties. Now, the volitionalist could attempt to insist that Abby must have made some still prior blameworthy choice that resulted in her subsequent cognitive failure. But it is unclear both what kind of choice this could be, short of a choice to ingest some cognitively disabling substance, and why we should suppose that this is the only way to justify a judgment that she has shown ill will or disregard.

Or suppose that Per's forgetting the special office hours appointment is due to his prior failure to realize that he may need to take steps in order to ensure he remembers. Perhaps the volitionalist will have to claim that Per made some prior

[17] For a discussion of the kinds of normative requirements that can make such expectations reasonable, see Goodin (1986).

blameworthy choice that caused this cognitive failure, which in turn caused him to forget the appointment. As I have been trying to stress, it may very well be that no such choices actually took place, or worse, that all we find instead of such choices are still further intuitively blameworthy unintentional omissions to act in a way that would or might prevent a subsequent chain of negligence or forgetfulness. But what is crucial to note is that the mere fact that no such prior choices took place does not, by itself, seem to automatically imply that a judgment of blameworthiness is inappropriate. All this fact seems to imply, to my mind, is that the propriety of this judgment doesn't depend upon anything these agents have done or any of the choices they have made.

This leads to the motivation behind claim (2). There is nothing incoherent in the idea that an unintentional omission for which an agent is intuitively blameworthy might be explainable in terms of prior cognitive failures for which she is also intuitively blameworthy.[18] I have tried to suggest that it is at least as plausible to think that each of my example agent's unintentional omissions resulted from prior blameworthy cognitive failures as it is to think that they resulted from blameworthy choices that more proximally preceded them. The more intuitively blameworthy cognitive failures lie behind the unintentional omission, the more intuitively compelling a judgment of blameworthiness will appear to be, especially if this sequence of cognitive failures leading up to the unintentional omission involves several intuitive showings of ill will or disregard instead of just one. Yet, at the same time, the further this sequence stretches back in the agent's history, the more ad hoc it will be for the volitionalist who genuinely wants to explain the intuition of blameworthiness to insist that what ultimately sets this sequence of deep negligence or forgetfulness in motion is some choice that reflects enough ill will or disregard to which all the agent's intuitive blameworthiness can ultimately be traced.[19]

Now this produces what I think is a very puzzling result. I've suggested that an unintentional omission that stems from several prior blameworthy cognitive failures may in some cases only appear to increase our confidence that moral criticism is appropriate, insofar as this would indicate that the agent has routinely exhibited a deep and pervasive kind of blameworthy negligence or forgetfulness.[20] Yet it is just

[18] Indeed, this kind of case seems analogous to the kind of case that forms the subject matter of Steven Sverdlik's excellent discussion of what he calls "Pure Negligence". He claims that "there do seem to be cases of negligence where there is no deliberate prior abstaining from getting knowledge or a deliberate prior refraining from stopping a loss of knowledge. All that there is, in some cases, is an unwitting violation of a norm, preceded by an indefinitely long period in which it never occurs to the person to consider the relevant risks". See Sverdlik (1993:140–41).

[19] Interestingly, even if such a prior choice can be found, the more distant it is in the agent's history, the harder it will be for this choice to satisfy the foreseeability constraint on the tracing strategy (see note 16). Indeed, for an approach like Zimmerman's that requires explicit conscious foresight of all the consequences of a volition that can be legitimate grounds for moral appraisal of an agent, this constraint may be even less often satisfied.

[20] Indeed, it seems that something like just this kind of explanation applies in cases of persons who, for want of adequate moral reflection, become habitual unintentional wrongdoers who exhibit deep patterns of insensitivity to moral considerations.

this kind of case that forces a difficult choice upon the volitionalist. For he may either (very implausibly) flatly deny that the agent is blameworthy, or else bank on the rather incredible claim that the agent's degree of blameworthiness must be fixed entirely by the quality of her will reflected in some prior choice that supposedly set this chain of deep negligence in motion.[21] Opting for the latter would force the volitionalist to presuppose both that this prior choice reflected monumental ill will or disregard, and that the agent possessed incredible powers of foresight at this prior time of choice in order for any plausible version of the foreseeability constraint on the tracing strategy to be satisfied.[22] Here we see a striking implication of the volitionalist's claim that nothing but a volitional state can reflect an agent's quality of will in the way requisite to ground blameworthiness. For it could turn out that, paradoxically, some of our most intuitively compelling judgments of blameworthiness involving deep negligence or forgetfulness have the least stable theoretical backing on the volitionalist account. Such a result would pull at our intuitions in puzzling ways, for it wouldn't seem acceptable, given certain reasonable assumptions about the very blameworthy quality of will an agent's deep negligence or pervasive forgetfulness would reflect, to stake our confidence in a judgment of blameworthiness in the occurrence of some distant, and supposedly very blameworthy prior choice. But we might also feel at a loss to explain how anything besides a choice could reflect an agent's quality of will, thereby feeling pulled by the thought that our intuition that a judgment of blameworthiness is appropriate indicates that there must have been some prior choice to which all her blameworthiness can be traced.

Now I will attempt to motivate claim (3). This claim states that the volitionalist's use of the tracing strategy produces a substantial distortion of a very natural conception of the content of the kind of responses that blameworthy unintentional omissions seem to make appropriate. This distortion occurs because the volitionalist's use of the tracing strategy directs our attention away from what appears to be the object of principal moral concern in a case of blameworthy unintentional omission, and that consequently seems to be what the agent is open to criticism for. Intuitively, it is Mikael's forgetting the anniversary, Per's forgetting the special appointment, and Abby's failing to assist the swimmer that lead us to judge that each has shown ill will or disregard. This intuition is reinforced by considering the kinds of response that are made uniquely appropriate in virtue of just these episodes of forgetting or neglect. Mikael, Per, and Abby could understandably feel that these episodes by themselves make appropriate such responses as guilt, remorse, and overt

[21] It may seem unfair to suggest that the volitionalist must ground an agent's blameworthiness for a chain of deep negligence or forgetfulness in a single prior choice the agent made. Why couldn't the volitionalist insist that this chain is the result of more than one prior choice? My concern is not with such cases. Rather, my concern is with cases in which a chain of deep negligence or forgetfulness doesn't appear to be owing to *any* prior choices the agent has made. In such cases, I am not claiming that the chain of deep negligence is owing either to one ultimate choice or no choice at all.

[22] See notes 16 and 19 above.

attempts to demonstrate to those whose importance seems to be called into question that the offense, while granting that it was blameworthy, did not stem from any explicit choice to disregard them.[23]

But the volitionalist account suggests that such responses might substantially fail to address the proper object of moral concern in these cases. Instead, these kinds of response, if they were more clear-headed, would directly address not what wasn't chosen, but what was chosen. An adequate attempt at reconciliation on Mikael's behalf might even involve nothing more than expressing his remorse for whatever aspect of his putative prior choice was faulty, for this choice is after all, the volitionalist tells us, the principal object of moral concern in his case. It might actually be misguided for Mikael, Per, or Abby to think that there is any unique way that their respective unintentional omissions make guilt appropriate, or that an adequate attempt at reconciliation should involve acknowledgement that the relevant episode of forgetting or neglect was in and of itself morally problematic. These unintentional omissions need to be acknowledged only inasmuch as recognition of them serves as a path to acknowledgement that a prior choice expressed ill will or disregard, and if the latter acknowledgement occurs without the former, then it seems as though the volitionalist could in principle admit that nothing of moral significance is lost in an attempt at reconciliation. But this seems to be a rather blatant distortion of what we would normally take the principal objects of moral concern in such cases to be, insofar as we tend to think that what needs to be acknowledged when offering an apology or attempting reconciliation is the unintentional omission itself and what it seems to indicate about a wrongdoer's relations to others.

It also seems that appropriate blaming responses on behalf of those who were forgotten or neglected would principally concern the unintentional omission itself and what it seems to reveal about the wrongdoer's relationship to the one who has been forgotten or neglected. Mikael's forgetting would surely seem to make it appropriate for his wife to ask "How on earth could you have forgotten about our special day?!" Similarly, the struggling student might wonder how Per could have been so neglectful, given how much she was depending upon him for his assistance. But the volitionalist picture directs attention away from these episodes of forgetting and neglect, implying that their moral significance in these cases is wholly parasitic on the moral significance of some prior choice these wrongdoers have supposedly made. In all cases, then, the crucial question seems to be one that asks of an unintentional omitter, "How could you possibly have made that choice?!"

Finally, I need to motivate claim (4), which concerns the volitionalist implication that unintentional omissions can never, by themselves, constitute substantive

[23] While this may sound like an attempt at excuse, and so to *affirm* the idea that blameworthiness for the unintentional omission must be traceable from a prior choice, this appearance should be resisted. In such a case, we can imagine that each agent would find it important to point out that the disregard was not explicitly chosen. But this would not entail that each agent would be insinuating that no disregard was *shown*.

bases for moral criticism of an agent. Because an agent's blameworthiness for an unintentional omission is fixed entirely by the ill will or disregard reflected in a prior choice she made, the volitionalist must say that any additional substantive moral criticism of an agent on the basis of an unintentional omission amounts to an acceptance of moral outcome luck. This is because, as we have already seen Wallace claim, the only legitimate targets of moral assessment are qualities of will reflected in choices, and these qualities of will are what they are whether or not any particular unintentional omissions flow from these choices. It follows straightforwardly that, on the volitionalist account, there is never anything intrinsically significant about unintentional omission, from a moral point of view. Unintentional omissions are not grounds for moral criticism of agents over and above any moral criticism that applies to them in virtue of their choices. Assuming that Abby is in fact blameworthy for some choice she made prior to falling asleep, what justification could the volitionalist provide for denying that Abby would be just as blameworthy if the same choice obtained but she never fell asleep? What could possibly be the basis for insisting that Mikael is open to a milder degree of criticism if, after criticizably choosing not to take the necessary steps to ensure he would remember the anniversary, he nonetheless remembers?

At this point, I imagine that many will experience the standard conflicting intuitions typical of consideration of cases of moral outcome luck. Intuitively, it seems that Mikael would be more blameworthy if he both chose not to take precautions to remember his anniversary and forgot, than he would be if he made the same choice but nevertheless remembered. Similarly, a lifeguard who both chooses to brush aside her drowsiness and then falls asleep on the job is seemingly open to a more serious degree of moral criticism than one who chooses similarly, but never falls asleep. Of course, everything depends upon how these cases are described. If falling asleep on the job is an event that reflects additional ill will or disregard over and above whatever ill will or disregard was reflected in Abby's putative prior choice, then she will actually be more blameworthy than she would be if she'd chosen identically, but stayed awake. The challenge here consists in explaining how her dozing could reflect additional ill will or disregard without this being reflected in a choice. If we bring back into focus the volitionalist's claim that only a choice can reflect an agent's quality of will, then such an explanation might be thought to be impossible.

The worries I have tried to motivate in this section are aimed at highlighting some of the distorting implications of the way the volitionalist uses the tracing strategy to explain how certain unintentional omissions can ground legitimate judgments of blameworthiness. By no means have I tried to show, on the basis of the foregoing remarks, that volitionalism essentially fails to specify plausible conditions of blameworthiness. But I do hope that it is now relatively clear that taking volitionalism seriously could require a substantial revision to what I believe is our common sense picture of the way certain unintentional omissions ground judgments of blameworthiness. Next, I will try to cast doubt on the legitimacy of regarding choice as a basic precondition of blameworthiness, in hopes of showing why it is reasonable to think that volitionalism fares no better than rival non-will-centered accounts of moral responsibility, from a theoretical point of view.

7.5 Choosing Between Volitionalism and Non-will-centered Approaches

In this section, my principal aim is to expose a puzzling assumption that the volitionalist appears to adopt and that, if rejected, opens the door for solutions to the worries I motivated in the previous section. In order to accomplish this aim, I must first explain what I think is the most compelling rationale for the volitionalist to use the tracing strategy as he does. I believe that the volitionalist is motivated to use the tracing strategy as he does in order to reconcile his theoretical account of the conditions of blameworthiness with a common and conflicting conviction that ill will or disregard may be directly reflected not only in the choices an agent makes and the actions she performs, but also in some of the choices she doesn't make and some of the actions she doesn't perform. Let me explain why I think this conviction is one that is commonly accepted.

We should begin by noting that nothing in the volitionalist's account of the conditions of blameworthiness precludes his ability to agree that Abby's failure to assist the swimmer is explainable in terms of at least some of her evaluative mental states without its necessarily being a causal consequence of any prior choices she has made. We would quite naturally expect a lifeguard who sincerely believed (however implicitly) that the safety of the swimmers in her care is of utmost moral importance to be motivated by this belief to see to it that she is able to care for them when they need her most. Perhaps the fact that Abby doesn't see to this, then, provides direct evidence of certain evaluative beliefs and attitudes (henceforth, "EM-states") she holds (however implicitly) about the swimmers' safety. If we came to discover that Abby didn't believe their safety was very important, this could plausibly be taken to directly explain her susceptibility toward failing to care for the swimmers as she should.

So, it is at least possible that the fact that Abby does not monitor her situation more attentively indicates that she holds (however implicitly) at least some objectionable EM-states about the importance of the swimmers' safety, which, to my mind, suggests a relatively natural explanation of her inaction in terms of just these objectionable EM-states. It is important to note that the adequacy of this explanation does not depend upon whether any particular underlying EM-states are explanatorily relevant to her unintentional omission; rather, what matters is just that her inaction is capable of being explained in terms of some or other of her EM-states. There are two noteworthy features of this explanation: first, its compatibility with the volitionalist account of the conditions of moral blameworthiness, and second, its illustration of how the fact that an agent does not make certain choices, or does not do certain things, can be directly explainable in terms of certain of her underlying EM-states.

That we often take certain inactions to be directly explainable in terms of some of an agent's underlying EM-states is supported by noting how frequently such explanations are operative in some of our ordinary explanations. If you always fail to think to leave a tip on the table when we dine, I would seemingly have prima facie reason to directly infer that you believe there isn't anything worthwhile or important

or useful or desirable about tipping.[24] What is less clear, to my mind at any rate, is why I would have any stronger reason to infer that you explicitly chose to do something that has resulted in your routinely not thinking about whether you should leave a tip.[25] If I keep promising to bring that DVD you asked to borrow, but repeatedly forget to do so, then you could very reasonably take my forgetting as direct evidence of my belief that it isn't terribly important that I make good on my commitment. It isn't obvious why you would have any more compelling reason to think that my repeated forgetting in this circumstance provides direct evidence of any choices I have made.[26] I am not claiming that these particular EM-states are the only ones that can properly explain the relevant inactions. All I mean to point out is that there are at least some circumstances in which we very naturally think that an inaction can be directly explained in terms of certain EM-states themselves, and in which we generally do not think we have any stronger reason to explain such inactions in terms of any choices that an agent has made.

Now, back to Abby's case. Because the volitionalist is antecedently committed to the claim that only choices can reflect ill will and disregard in the way necessary to justify judgments of blameworthiness, he faces considerable pressure to attempt to locate all the relevant mental states that seem to constitute Abby's ill will or disregard and that intuitively explain her unintentional omission in some prior episode of choice. But, even if we assumed that Abby made some such choice, what would make this choice reflective of ill will or disregard is not the fact that it is merely some volitional state called a "choice", but is rather the fact that the evaluative beliefs and attitudes that formed the basis for this choice are morally objectionable. Indeed, to what extent could a choice be said to reflect ill will or disregard if it didn't provide some indication of the precise content and moral quality of the underlying EM-states that constitute its basis? Choices seem to be capable of reflecting ill will or disregard only because they necessarily implicate certain EM-states that, in turn, may be taken by others to be morally objectionable. We have reason to attach moral

[24] Consider a friend who is visiting from a country in which there is no established social practice of tipping. The fact that it does not occur to him to leave a gratuity when paying the check seems to be explainable in virtue of the fact that tipping is not an activity that he sees any reason to regard as important. My frustrated attempts to explain to him how important it is that it *does* occur to him to tip can be seen as my trying to convince him to believe that tipping is important, at least while he is a visitor in my country. What is *objectionable* in this scenario is not some prior choice he has made to ignore the social customs of my culture, but some underlying evaluative belief that directly explains his failure to think to leave a tip.

[25] I am not claiming that your failure to think to leave a tip is *only* explainable in virtue of the presence of these underlying beliefs and attitudes. I am only claiming that such an explanation is possible, and that there may be no reason to think that this kind of failure is explainable only in virtue of some prior choice you have made.

[26] But what about circumstances in which the forgetting is just a kind of "mental hiccup", and doesn't appear to be explainable in virtue of any of the agent's evaluative beliefs and attitudes? I am not saying that these circumstances are impossible. The fact that sometimes an episode of forgetting is explainable in terms of such mental hiccups does not jeopardize my claim that in some cases it might be appropriate to see an episode of forgetting as explainable in terms of an agent's evaluative beliefs and attitudes.

significance to just these underlying EM-states themselves, insofar as they concern our conceptions of ourselves in the world, who and what we value, and what we consider worth doing. Our choices are morally significant, then, just because they necessarily implicate certain of our EM-states. By contrast, it seems strange to say that we have reason to attach moral significance to choices simply because they are moments during which one's mental condition is in some volitional state.

If I am right so far, then we should accept the claim that a choice is reflective of ill will or disregard to the extent that the EM-states that form the basis for this choice are morally objectionable. One way to interpret the motivation behind volitionalism is that the principal target of moral appraisal is not merely a choice insofar as it is a choice, but the moral quality of the EM-states that form the basis for that choice.[27] For it is just these mental states that are capable of being evaluated as acceptable or objectionable from a moral point of view. This seems, to my mind, to be a compelling way to understand what is morally significant about an agent's quality of will, whereby what is centrally at issue in questions of this sort is the moral quality of the particular evaluative beliefs and attitudes that form the basis for the choices an agent makes.

But if we accept that quality of will concerns just the moral quality of these EM-states, then we should wonder why their failure to form the basis for an actual choice entails that they cannot justify a judgment that ill will or disregard has been shown if they are both objectionable and explanatorily relevant to an unintentional omission. At best, the fact that some of an agent's morally objectionable EM-states are reflected in a particular choice provides an explanation of how this particular choice can reflect ill will or disregard. The volitionalist can agree with this; as I have already noted, nothing in his account of the conditions of moral blameworthiness precludes his agreement that certain unintentional omissions are directly explainable in terms of an agent's morally objectionable EM-states. And if, as I have suggested, quality of will concerns just the moral quality of these EM-states, why should we think that these mental states cannot justify a judgment that ill will or disregard has been shown in cases where they are both objectionable and directly explanatorily relevant to her unintentional omission?

One possibility is that the volitionalist thinks that judgments of blameworthiness fundamentally concern an agent's choices because he assumes that an agent's EM-states can reflect ill will or disregard only insofar as they form the basis for some explicit choice she has made.[28] But, before we accept the volitionalist's claim to be

[27] Some might object that what is of interest is how the agent deliberates over these evaluative mental states, and how she selects which will be effective in moving her to action and which ones will not. I think this is a very implausible objection, insofar as it seems to present a picture of the agent as something that is capable of standing over and above all her evaluative mental states and distancing herself from them so significantly so as to be constituted purely by her rational will. This is why I do not think it is plausible for a volitionalist to maintain that it is the *fact* that a choice is a choice – a moment of pure self-determination – that is morally significant.

[28] One motivation behind this assumption is that these mental states cannot be identified as "the agent's own" except when they are somehow endorsed or are a basis for the agent's explicit identification with a certain evaluative perspective insofar as she chooses that they be effective in moving

offering the correct account of the conditions of moral blameworthiness, we can reasonably ask for an explicit defense of this assumption. For it is just this assumption that, I think, makes it very difficult for the volitionalist to explain how agents can be open to moral criticism on the basis of unintentional omissions that do not stem from any explicit choices they have made. If this assumption is rejected, then such explanations are much easier to provide, since they will not require one to advocate the kind of indirection and *ad hoc* stipulations associated with the volitionalist's use of the tracing strategy.

Let me provide a brief comment why one might think this assumption worthy of rejection. What seems fundamental for moral blameworthiness, I claim, is just the moral quality of the EM-states that are explanatorily relevant to an agent's conduct, and not the fact (if it is a fact) that these states happened to be implicated in any explicit choice. For whether these states happen to figure into a choice does not seem to change whether their content and moral quality can provide a reasonable basis for objections. Moreover, blaming responses (guilt, resentment, indignation, requests for apology and justification) seem to be responses to further judgments that the particular content and moral quality of an evaluative belief or attitude reflected in one's conduct is objectionable, and not responses to some further perception or belief that an objectionable EM-state is or was ever implicated in a choice. If I am right, then there is no reason to engage in the kind of tracing that the volitionalist endorses in order to explain an agent's blameworthiness for an unintentional omission. For what is important is not whether the unintentional omission can be traced to some prior episode of choice that reflects ill will or disregard, but instead whether the unintentional omission is explainable in terms of certain of an agent's EM-states whose precise content and moral quality would, if implicated in a choice, both entail that the choice reflects ill will or disregard and make various blaming responses appropriate. After all, this seems to be what the volitionalist is ultimately seeking, insofar as he recognizes that moral appraisal concerns the moral quality of just these EM-states. So, armed with this reason, we can reasonably ask the volitionalist to provide an explicit defense of the assumption that these EM-states must form the basis for an explicit choice in order for them to be capable of reflecting ill will or disregard.

P.F. Strawson pointed out that, as parties to interpersonal relationships of many kinds, it matters greatly what our intimates (and even what other strangers) believe to be worthwhile or significant.[29] But it is far from clear that these evaluative beliefs and attitudes matter to us only when they form the basis for a choice. If Mikael (however implicitly) doesn't place the same degree of evaluative significance upon celebrating important relationship milestones as his wife does, his evaluative attitude can, due to its content, be taken by her as a reasonable basis for objections within the context of their relationship, however mild in tone these objections may be. Per's

her to action. I will not consider this motivation here since there is much controversy surrounding it, and in any event this motivation is more reasonably seen as a *statement*, rather than as a *defense* of the volitionalist position. For further discussion, see Wallace (2002), Smith (2004), and Hieronymi (2008).

[29] Strawson (1962).

student would seem to have satisfactory grounds for (at least some) resentment if the student came to learn that Per didn't really care about the student's success, and this would be true no matter the particular choices Per has made. We might even feel slightly indignant if we learned that (perhaps only out of boredom and frustration with her job) Abby would rather catch a quick nap than safeguard the swimmers on the beach, and this feeling would be legitimate even if she generally manages to do a fine job staying alert and watching over them.[30] All of this suggests, to my mind at least, that an agent's quality of will is intimately associated with the precise content and moral quality of the EM-states that either potentially or actually form the basis for her actual choices. Here, my aim has been to illustrate how mere knowledge of another's EM-states, particularly those EM-states that concern oneself, can constitute grounds for legitimate (if only mild) objections within the context of an interpersonal relationship, but could also very appropriately give rise to feelings of guilt on behalf of those who recognize that they harbor such objectionable EM-states.[31] This is, I take it, further evidence that an agent's quality of will is a function of the precise content and moral quality of her EM-states, and that it is unclear why these EM-states must figure into any of an agent's explicit choices in order to ground attributions of blameworthiness.

Recent attempts to develop non-will-centered approaches to moral responsibility have elaborated how certain of an agent's EM-states can reflect ill will or disregard even when these states do not form the basis of any choices she has made. The way such explanations proceed is the subject of much attention in the current debate over the conditions of moral responsibility.[32] While I am confident that many of these explanations can be shown to be defensible, my aim in this paper has been, in part, to point out that one way of framing the debate between will-centered approaches to moral responsibility like volitionalism, and its rival non-will-centered approaches is as a debate over whether there is a defensible theoretical rationale for assuming, as the volitionalist does, that all moral blameworthiness for unintentional omission must be traceable to an agent's quality of will as it was reflected in some choice. If this assumption is indefensible, then we will have further reason to think it unclear why we should prefer volitionalism over non-will-centered approaches. Moreover, given the way the volitionalist's use of the tracing strategy tends to distort what may be regarded as a very commonsense picture of the moral significance of unintentional omission, it seems that a non-will-centered approach could more easily explain the important role EM-states play in grounding moral blameworthiness

[30] That indignation is only slightly felt does not indicate, by itself, that the absence of choice substantially mitigates to the point of near excuse. For it may be an implicit assumption that Abby's holding the underlying objectionable evaluative attitude is compatible with her making strong efforts to satisfy her duty. If this is so, then the overall quality of the reactive sentiment we experience will be conditioned by this further assumption, and the concomitant attitudes of approval this further assumption entails.

[31] Indeed, guilt seems appropriate when one realizes, as a result of one's conduct, that one holds an objectionable evaluative attitude or belief one previously hadn't recognized.

[32] See, for example, Arpaly (2003), Sher (2009), Moya (2007), and Smith (2005b).

without incurring the costs of the volitionalist's requirement that all tracing must culminate in a choice.

7.6 Conclusion

My aim in this paper has been to produce strong skepticism about the volitionalist's ability to provide a satisfactory account of the conditions of moral blameworthiness in cases of unintentional omission. I have tried to motivate a compelling case for rejecting the claim that the moral significance of unintentional omission consists entirely in what it indicates about the moral quality of an agent's prior choices and actions. The moral significance of both action and inaction, I suspect, consists in what each indicates about the precise content and moral quality of the underlying evaluative attitudes and beliefs that are explanatorily relevant to an agent's conduct. For this reason, I suspect that theories that have articulated the conditions of responsible agency primarily in actional terms fall short of offering a compelling theoretical explanation of our practices of moral criticism.

Acknowledgments Thanks to Angela M. Smith, Janice Moskalik, and the participants of the University of Delft Conference on Moral Responsibility, Neuroscience, Organization, and Engineering for helpful comments on both written and presented versions of this paper.

References

Arpaly, Nomy. 2003. *Unprincipled Virtue: An Inquiry into Moral Agency*. Oxford: Oxford University Press.

Dahl, Norman O. 1967. "Ought and Blameworthiness." *The Journal of Philosophy* 64:418–28.

Feinberg, Joel. 1984. *Harm to Others*. Oxford: Oxford University Press.

Fischer, John M., and Neal Tognazzini. 2009. "The Truth About Tracing." *Nous* 43:531–56.

Goodin, Robert. 1986. "Responsibilities." *The Philosophical Quarterly* 36:50–56.

Hieronymi, Pamela. 2004. "The Force and Fairness of Moral Blame." *Philosophical Perspectives* 18:115–48.

Hieronymi, Pamela. 2008. "Controlling Attitudes." *Synthese* 161:357–73.

Levy, Neil. 2005. "The Good, the Bad and the Blameworthy." *Journal of Ethics and Social Philosophy* 1:1–16.

Moya, Carlos J. 2007. "Belief and Moral Responsibility." In *Intentionality, Deliberation, and Autonomy: The Action-Theoretic Basis of Practical Philosophy*, edited by C. Lumer and S. Nannini. Burlington, VT: Ashgate pp. 273–287.

Scanlon, Thomas. 1998. *What We Owe to Each Other*. Cambridge, MA: Belknap Press.

Sher, George. 2009. *Who Knew? Responsibility Without Awareness*. Oxford: Oxford University Press.

Smith, Angela M. 2004. "Conflicting Attitudes, Moral Agency, and Conceptions of the Self." *Philosophical Topics* 32:331–52.

Smith, Angela M. 2005a. "Responsibility for Attitudes: Activity and Passivity in Mental Life." *Ethics* 115:236–71.

Smith, Patricia G. 1990. "Contemplating Failure: The Significance of Unconscious Omission." *Philosophical Studies* 59:159–76.

Smith, Patricia G. 2005b. "Feinberg and the Failure to Act." *Legal Theory* 11:237–50.

Strawson, P.F. 1962. "Freedom and Resentment." *Proceedings of the British Academy* 48:1–25.

Sverdlik, Steven. 1993. “Pure Negligence.” *American Philosophical Quarterly* 30:137–49.
Vargas, Manuel. 2005. “The Trouble with Tracing.” *Midwest Studies in Philosophy* 29:269–91.
Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.
Wallace, R. Jay. 2002. “Scanlon’s Contractualism.” *Ethics* 112:429–70.
Watson, Gary. 2004. “Two Faces of Responsibility.” In *Agency and Answerability*, edited by Gary Watson, 260–88. Oxford: Oxford University Press.
Zimmerman, Michael J. 1998. *An Essay on Moral Responsibility*. Totowa, NJ: Rowman & Littlefield.

Chapter 8
Desert, Responsibility and Luck Egalitarianism

Diana Abad

Abstract Desert and responsibility are key concepts in political philosophy, most notably in discussions on justice. It is just that people get what they deserve, and what they deserve seems to have something to do with what they are responsible for. This tenet is as close to a fundamental constant as one can get in practical philosophy, so that even some egalitarians, luck egalitarians, make room for exceptions dictated by it: only differences people are not responsible for should be equalized, differences people are responsible for are not unjust, because they are deserved. In this paper I shall contest the second part of this tenet that what people deserve is somehow linked to what they are responsible for. To this end, I shall give a detailed account of the concept of desert in the first half of this paper. In the second half, I shall consider the implications of this for luck egalitarianism, and conclude that while luck egalitarianians can counter some criticisms that are grounded on a wrong understanding of the concept of desert, they cannot rest content in relying on the purely formal notions of responsibility and desert, but need to provide substantial arguments to support their conclusions.

8.1 Desert and Responsibility

Regarding desert, astonishingly many people think that only persons can deserve something, and only in virtue of actions for which those persons are responsible. This is astonishing, to my mind, as in everyday language all sorts of things are said to be deserving in virtue of all sorts of properties. For example, we do say that criminals deserve punishment, and that good pupils deserve to get good marks, but we also say that the most beautiful contestant deserves first prize at the beauty contest, that great paintings deserve to be admired, and that unique landscapes deserve to be preserved.

In this paper I shall examine the conceptual connection between desert and responsibility, and I shall argue that there is none. That is, I shall argue that those who claim that it is a conceptual truth that only persons can deserve something solely for actions for which they are responsible are mistaken. I have absolutely no

D. Abad (✉)
Department of Philosophy, University of Potsdam, Potsdam, Germany
e-mail: diana.abad@uni-potsdam.de

N.A Vincent et al. (eds.), *Moral Responsibility*, Library of Ethics and Applied Philosophy 27, DOI 10.1007/978-94-007-1878-4_8,

idea whether or not the claim that only responsible action can ground desert is true substantially. Personally, I am inclined to doubt it, but that is neither here nor there, as that is not the question I will pursue in this paper. I shall not argue for the truth or falsehood of any given desert judgment. My aim is just to clear up a conceptual confusion.

Since it is the concept of desert I aim to elucidate, and not the concept of responsibility, I shall have a lot to say about the former and rather little about the latter. I will take a lot for granted as far as responsibility is concerned although I realize that things are far from clear in that regard. Most importantly, I will take for granted that we all know what it is to be responsible for an action. Basically, this is already all I have to say about responsibility.

On to desert then.

8.1.1 Desert: The Basics

The first thing to be said about desert is that it is a normative notion, that is, the fact that somebody deserves something implies that she ought to get it, although not unconditionally, but only *pro tanto*.[1] So, the fact that somebody deserves something is always a reason for giving it to her, although not always a conclusive reason. Something more important could count against it. The best team deserve to get the cup, but if the best team do not actually win, they ought not to get it.

The next thing to be said is that desert is a three-place relation, "x deserves y in virtue of z". Let us call x the desert subject, y the desert object, and z the desert base.[2] There are many desert judgments that do not explicitly state a desert base.

Sometimes, we simply say, for example: "The team deserve the cup." This is elliptical, though. One cannot deserve something for no reason at all, but only in virtue of something, the desert base, as Joel Feinberg has shown in his seminal paper "Justice and personal desert", giving the first and most influential analysis of the concept of desert. Moreover, the desert base must be attributable to the desert subject.[3]

Hence, desert judgments like "The team deserve the cup." need to be understood as abbreviated and as always implying a desert base which needs to be attributable to the desert subject, the team in this case. So, we have to supplement the statement with a desert base, for instance like this: "The team deserve the cup because they have played so well.", because having played well is attributable to the team, whereas we cannot supplement the statement like this: "The team deserve the cup because of water's boiling point.", because water's boiling point is not an attribute of the team's.

Here the trouble begins, because it is not quite clear what is "attributable" to a desert subject. Me, I do not mean to say anything more with this than that the desert

[1] Cf. Feinberg (1963:60).

[2] Cf. McLeod (1999:61–2).

[3] Cf. Feinberg (1963:58ff.).

base has to be an attribute of the desert subject's, be it an action or a property. It just has to be theirs, nothing more, nothing less.[4] However, there are others who think that only actions for which the desert subject is responsible are "attributable" to them in the requisite sense.[5] This also means, of course, that only persons can be desert subjects, because other sorts of things cannot perform such actions.

This is the view I shall contest in this part of the paper. I will call it the responsibility view. Before doing so, however, let me explain why this view is so curiously widespread, even though the concept of desert clearly works differently in everyday language.

8.1.2 Feinberg and Rawls

I think the seed for the responsibility view was already laid in Feinberg's analysis of desert. In it, Feinberg is concerned with a particular sort of desert objects: prizes, grades, rewards and punishment, praise and blame, compensations, in short: he is concerned with certain forms of treatment as desert objects only. And the only desert subjects he is concerned with are persons.[6] He exclusively looks at persons who receive these desert objects of prizes, grades, rewards, etc.

Concentrating as he does on persons as desert subjects suggests that only actions for which the desert subjects are responsible can be desert bases, because if only persons can be desert subjects this must be due to something that is peculiar to persons, something like actions for which they are responsible which no other animal or object can lay a claim on.

This is corroborated by the particular set of desert objects Feinberg considers. Only persons deserve these sorts of things, and only in virtue of actions for which they are responsible. It just would be nonsensical to assign punishment, say, to any other sort of desert subject than persons, and it would be unfair to do so on any other basis than an action for which they are responsible.

After having sprung from this seed, the responsibility view throve and prospered further with the publication of Rawls's *A theory of justice* a few years after Feinberg's paper which enormously influenced the subsequent treatment of the concept of desert in philosophical discussion.[7] Rawls as well considers only actions for which persons are responsible as desert bases. As he says in *A theory of justice*:

> The precept which seems intuitively to come closest to rewarding moral desert is that of distribution according to effort, or perhaps better, conscientious effort.[8]

---

[4] Cf. Cupit (1996:92ff.); Feldman (1995:186–7).

[5] Cf. e.g. Sadurski (1985:117); Rachels (1978:157); Rachels (1986:143); Sher (1987:37ff.); Smilansky (1996).

[6] Feinberg (1963:62,55).

[7] Abad (2007:part 1, chap. II).

[8] Rawls (1971:311).

Evidently, a conscientious effort consists in an action for which a person is responsible.

So, it seems as if Feinberg and Rawls subscribe to the responsibility view, and, influential as their theories have been, as if Feinberg and Rawls really are responsible for the responsibility view being so widespread. I think the latter of these claims is true. I do think that Feinberg's and Rawls's analyses of desert are responsible for many people holding the responsibility view.

The former claim, however, is false. Neither Feinberg nor Rawls really held the responsibility view. It is true that they limit their discussion of desert to cases in which persons deserve something in virtue of actions for which they are responsible, but this is due to the fact that they are concerned with examining the particular connections between desert and justice. Still, this restriction does not mean that other things than persons cannot be desert subjects, nor that other things than responsible actions cannot be desert bases. Neither Feinberg nor Rawls excludes the possibility of a painting's deserving admiration. The thing is just that the question whether or not a painting deserves admiration is not a question pertaining to justice.[9]

8.1.3 Against the Responsibility View

Even so, many philosophers hold the responsibility view. So, what is wrong with it?

What is wrong with it is that it puts the cart before the horse. Instead of examining the concept of desert and then deriving from that which individuals may permissibly replace the variables x, y, and z, they try to derive the concept from the individuals which they already know must be the only permissible ones. This is methodologically unsound, and it begs the question whether those individuals really are conceptually the only permissible ones.

Consider, by way of analogy, the predicate "x is a flightless bird", and consider a zoologist, a newbie on the field of ornithology, who is convinced that penguins are the only flightless birds there are. Now one day our zoologist is confronted with an ostrich. There are two ways of reacting open to our newbie-zoologist: first, he can correct his belief and say that penguins evidently are not the only flightless birds there are; or second, he can flatly refuse to recognize an ostrich as a flightless bird and qualify it as something else instead on the grounds that, obviously, it is not a penguin. Clearly, the first way of reacting is the adequate one. Unfortunately, regarding desert, the proponents of the responsibility view take on the equivalent of the second way.

As Wojciech Sadurski, one of the many champions of the responsibility view, puts it representatively:

[9] Cf. Feinberg (1963:55); Pogge (1989:63ff.); Abad (2007:14, 21–9) for a fuller discussion of this point.

> When we are pronouncing judgments of desert, we are inevitably making judgments about persons whom we hold responsible for their actions. It makes no sense to attribute desert. . . to persons for actions or facts over which they have no control.[10]

This is rather a curious statement about desert judgments. What exactly does Sadurski mean by saying that in talking about desert we "inevitably" talk about persons responsible for their actions, and that talking about other things in connection with desert "makes no sense"? It is obviously not inevitable in the sense that we cannot but, or do not, talk about desert in any other way, because we do so all the time. Let us say, for example, that a man may deserve something good in virtue of his noble birth, as the Ancient Greeks believed. In what sense does it "make no sense" to say this?

Sadurski leaves it at that and does not elaborate what he means. More is the pity, since it needs elaborating, because clearly, that judgment does not "make no sense" such that we could not possibly understand what it means. We do. We might be inclined to disagree, but we understand it alright. There are several other possibilities, though: Sadurski might mean that it is analytically false to say this, or that there is some kind of Strawsonian truth-value gap, or a category mistake, or quite literally, that this sentence is meaningless.

However, as of yet, we have not been offered a concept of desert such that the very meaning of the word excludes such desert judgments or makes them meaningless, or which implies a truth-value gap or a category mistake in such cases. Hence, to say that the desert judgment that a man deserves something good in virtue of his noble birth is analytically false, or meaningless, or that there is a truth-value gap regarding this judgment, or that there is a category mistake, would be to beg the question in a zoologist-reaction-number-2-kind of way as just described.

The desert judgment in question may very well be false, and moreover, it may very well be that only those desert judgments are true that have persons as desert subjects and actions for which they are responsible as desert bases. My point here is purely conceptual. These desert judgments may be false, but not analytically so, and in no sense are they meaningless. If they are false, they are not so because the concept of desert does not allow for these kinds of desert judgments. It does, as I shall go on to show. If they are false indeed, they must be so because of substantial reasons rather than conceptual ones.[11]

The proponents of the responsibility view might mean something else entirely, though. They might say that only desert judgments about persons and the actions they are responsible for are "real" desert judgments, whereas any other desert judgments are merely metaphorical, or derived, just loose talking, or manners of speech. We understand them well enough, but they are not to be taken seriously since they are not to be taken literally. Alternatively, they might want to distinguish different concepts of desert, say, one that pertains to questions of justice and that deals only with persons and responsible actions, as Feinberg and Rawls have it, and other

[10] Sadurski (1985:117, cf. fn. 5 above).

[11] Cf. Lamont (1994) for a similar line of argument.

concepts which may be useful in other contexts, all of them "real" concepts of desert and to be taken seriously.

The trouble with this proposal is that it presents us with a fragmented view of desert. On the face of it, the concept of desert works the same way in any of the many varied contexts in which it may be used, it is only the individuals whose names replace the variables that are vastly heterogeneous. So, an analysis of desert as a single, unified concept which covers all cases would be, on grounds of simplicity, superior to an analysis that chopped desert up into different concepts according to context. After all, an analysis of a concept is supposed to take seriously the different ways the concept is actually used and make sense of them. An analysis which resulted in saying that there is no way to make sense of the different uses, that they have nothing in common even though the same word is used, or an analysis that said that most people use a concept in a metaphorical way only, or indeed one that came to the conclusion that people do not know what they are saying really, would be a poor analysis if there was another one available.

So, it is really three problems the responsibility view has: First, it begs the question regarding whether there can be other desert subjects than persons and other desert bases than actions for which those persons are responsible. Secondly, it has to treat other desert judgments as meaningless, where at worst they are false. Thirdly, it fragments the concept of desert.

8.1.4 The Concept of Desert

So, what is the concept of desert?[12] What does it mean to deserve something? To deserve something means that it is appropriate to get it. And what does this in turn mean, that it is appropriate? Propriety, as I will introduce the term, consists of two components: the one is a relation I shall call fittingness, and the second is a certain normative element I shall call requirement.

Fittingness is a pretty straightforward thing. Puzzle pieces, for example, fit each other. But also states of affairs may be said to fit each other. Another way of putting this would be to say that what is fitting is a "response" to what it is fitting to. Returning the ball is a fitting response to having been served in tennis; having been asked what time it is, it is fitting to tell; going for a walk is fitting to the weather's being good; and so on, and so forth. Obviously, unlike particular cases like puzzle pieces, there may be more than one fitting response to given states of affairs. Hence, instead of going for a walk when the weather is good, it may also be fitting to go for a swim, or to hang out your laundry to dry, or to get inside if you are prone to get sunburnt. What is fitting to what really depends on the particular case and its circumstances.[13] To come back to the responsibility view for a second, though: it should be evident that it is not only actions for which persons are responsible that

[12] For a full discussion, cf. Abad (2007).

[13] Cf. Bittner (2001:chap. 4). Some of the examples mentioned are his.

can have fitting counterparts. Puzzle pieces and the weather's being good are cases in point.

In the case of desert, the desert object needs to be fitting to the desert base. This is not all there is to it, though, because as I have pointed out at the beginning, desert is a normative notion. So, the second component I have mentioned needs to come in, the normative element of requirement. Desert is a case of not only fittingness, but of propriety, and by that I mean that it is not only a fittingness relation, but one where the fitting counterpart is required.

Clearly, not everything that is fitting is also required, that is, not everything that is fitting is also appropriate. Two puzzle pieces may be put together, but they just as well may not. They do not require being put together as if that were their natural state they belonged in. Just so, the weather's being good does not require that I take a walk. It would just be a good idea, a fitting thing to do. I may as well not. Nothing goes wrong if I do not. By contrast, something does go wrong if what is appropriate does not occur.

Thus, to deserve something means that it is appropriate to get it, and this in turn means that it is fitting and required that the desert subject get the desert object in virtue of the desert base.

However, this still is not all there is to it, because there are cases of propriety which are not cases of desert. That is to say, there are cases where something is fitting and required, but not deserved. Consider for example a major scale. Anyone who has ever played seven notes of any major scale on any musical instrument will know that the eighth note is not only fitting, but required too. That is so because a major scale's seventh note is a leading note that leads on to the eighth note. If the eighth note does not follow the leading note leads to nothing, and that is just not right. Something goes wrong if a major scale's octave is not completed. This is evidenced by the fact that most listeners can hardly bear to hear it so that they add the eighth note either mentally or by singing it. This unbearableness for listeners is explained by the impropriety of the eighth note missing. Hence, it is not only fitting, but appropriate to add the eighth note after having played the first seven of a major scale. Yet, it is not the case that the first seven notes of a major scale deserve that the eighth be added.

To return again to the responsibility view: as requirement is the only difference between fittingness and propriety, and since not only actions for which people are responsible have fitting counterparts, these are also not the only things that require a fitting counterpart, as this last example of the scales shows.

The question now, though, is: as not all cases of propriety are cases of desert, which ones are? Here it would be very easy for the proponents of the responsibility view to jump in and say: only those cases of propriety are also cases of desert where the subject is a person and the base is an action for which she is responsible. To which I respond, as I have already at length in the last section, that this leaves out too many judgments of desert to be taken seriously as a good analysis of the concept.

Instead, I propose that only those cases of propriety are also cases of desert which are based on something fitting or unfitting, or in other words: desert is that propriety that is appropriate in virtue of something "fittingness-affecting". In still other

words, desert is something like "second-order fittingness": deserving something means, utlimately, that it is fitting (and required) to get it because one has done or is something (un-)fitting.

I realize of course that at this point at the latest things are getting just a tiny bit intricate, so let me elucidate this by way of some examples. I will use some of those I have mentioned at the beginning:

1. Criminals deserve punishment in virtue of their crimes. This means that punishing criminals is appropriate, and this in turn means that it is fitting and required to punish criminals. What makes it so? Committing crimes, the desert base, is itself something unfitting; let us say breaking and entering is unfitting to the concept and right of property. And this is what makes it fitting and required, that is appropriate, that they be punished. So, punishing criminals is fitting (and required), because their committing crimes is unfitting. Hence, criminals deserve punishment. If committing crimes was not itself fitting or unfitting to anything, if punishing criminals was only fitting (and required), because of their committing crimes period, their punishment would not be deserved but appropriate. The normative force is the same, the punishment is required in just the same sense both times, it is just that in the first case we can say that it is "deserved" whereas in the second it is "appropriate".
2. The most beautiful contestant deserves first prize in the beauty contest. In this case, being the most beautiful is the desert base, so this is what must be fitting to something. Well, the context is a contest and this contest is about beauty, so presumably, being the most beautiful is fitting to the concepts of contest and beauty. If this is so, this is what makes it appropriate that the most beautiful contestant get first prize, and hence she deserves it. Again, if things are not so, that is, if being most beautiful is not fitting to anythings, it is just appropriate period that the most beautiful contestant get first prize.
3. Great paintings deserve to be admired. What makes great paintings great, let us say, is that they exemplify to an extraordinarily high level what art is about. To be honest, I am making this up, I do not know the first thing about art and what makes paintings great. But remember that I am not concerned with the truth or falsity of this desert judgment, but with what it means. So, whoever says that great paintings deserve to be admired says something like this: their greatness is the paintings' desert base; their greatness is fitting to something, let us say to the idea of what art is about; so this is what makes it appropriate to admire them, and therefore they deserve it.
4. Certain landscapes deserve to be preserved. For instance, the UNESCO talks about Natural Heritage Sites like this. The idea is that these landscapes have particular features that are these landscapes' desert bases, and so that these features are fitting to something. Let us say that, whatever those features are, they are unique, and their uniqueness is fitting to Earth's marvelous variety of landscapes. (I am speculating again.) Hence, it is the fittingness of those

unique features that make it appropriate to preserve those landscapes, and so they deserve preservation.[14]

Obviously, not all propriety relations are based on something fitting or unfitting. The example of the major scale was a case in point. Hence, even though it is appropriate to add the eighth note to the first seven, it is not deserved, as playing seven notes of a major scale is not fitting to anything. But if it were, if there was some obscure story in which we could say that, then we would have to say that the seven notes deserve to have the eighth note added.

8.1.5 Conclusion

To wrap it all up: desert is that propriety that is based on something fitting or unfitting. So, propriety is a subset of fittingness; those fittingness relations are also propriety relations in which the fitting counterpart is not only fitting, but also required. In turn, desert is a subset of propriety; those propriety relations are also desert relations which are based on something fitting or unfitting.

This analysis makes clear that actions for which persons are responsible are not the only things there are that have fitting counterparts, or that require the fitting counterparts as appropriate, or in virtue of which one can deserve something. To think so is wrong for the three reasons I have given: it begs the question, it renders too many desert judgments meaningless, and it fragments the concept of desert.

So, the responsibility view of desert is wrong. We have to understand desert as I have explained, and there is nothing in that account that precludes other things than persons to be desert subjects nor other things than actions for which they are responsible to be desert bases.

8.2 Desert and Luck Egalitarianism

Now that the concept of desert and the role responsibility does, or rather, does not, play in relation to it are clear, it can be used to untangle misunderstandings in discussions in political philosophy in which both the concepts of responsibility and of desert figure crucially. One such discussion is the one on luck egalitarianism. It is outside the scope of this paper to engage in a fully fledged analysis of this debate, examine in detail how any given authors use the concepts of desert and responsibility, show how they go wrong on the basis of the foregoing considerations, and set them on the right track. However, I shall give a very rough and exemplary outline of how applying the correct concept of desert can help further the discussion substantially: I shall discuss two critics of luck egalitarianism, Serena Olsaretti and Nicole

[14] It is because of examples of kinds 3 and 4 that Smilansky's defense of the connection between desert and responsibility fails: paintings and landscapes are not ever "positively responsible" for anything, nor can they ever be "negatively responsible", cf. Smilansky (1996:160).

A Vincent, and show, first, how, with the proper account of desert expounded in this paper, luck egalitarianism can counter their criticisms, but also, secondly, that relying on this account of desert is still not enough to vindicate luck egalitarianism.

Luck egalitarianism, to quote Olsaretti, is the position that holds that those inequalities between people are unjust that are traceable to "circumstances that individuals could not reasonably foresee and avoid. By contrast, individuals are justly held responsible for, that is, they are liable to pick up the costs and reap the benefits of, events they could reasonably foresee and avoid".[15] This position is motivated by a powerful intuition Vincent calls the responsibility-tracking intuition, "i.e. the intuition that people should *take* responsibility for those things for which they *were* responsible, and that no one is entitled to expect others to take this responsibility for them".[16] So, the idea is if you make your own free choices and act accordingly, you have to suffer the consequences for better or worse. Hence, if you recklessly drive your motorcycle at high speed without wearing a helmet and have an accident, or freely gamble away all your savings, it is your own fault and you cannot expect your health insurance to pay your hospital bills and social security or anybody else to help you out.[17]

As common and as powerful as the responsibility-tracking intuition is, it is not quite clear what it means. Vincent argues that there are actually two distinct notions of responsibility at play here, one she calls outcome responsibility which involves attributing a particular state of affairs to a particular person, and one she calls liability responsibility which concerns the question "who should now do what" in consequence of that state of affairs being attributable to that person.[18] So, the responsibility-tracking intuition, precisely understood, should really read like this: you have to take liability responsibility for what you are outcome responsible for.[19]

At this point, two questions arise, and though Vincent and Olsaretti both consider both questions they each specially focus on one of them: first, does liability responsibility really follow from outcome responsibility, and secondly, if so, how do we determine what consequences exactly one is liable for given one's outcome responsibility.

8.2.1 How to Determine the Consequences One Is Liable For

To start with the second question, Olsaretti shows that, even assuming that liability responsibility does follow from outcome responsibility it is not as easy to determine what consequences exactly one is liability responsible for given one's outcome

[15] Olsaretti (2009:165–6).

[16] Vincent (2009:41).

[17] These are the examples Olsaretti discusses in her paper.

[18] Vincent (2009:45).

[19] Cf. Vincent (2009:46).

responsibility as luck egalitarians relying on the responsibility-tracking intuition would have it:[20]

> consider Bert's [the reckless motorcyclist's driving at high speed without a helmet and subsequently injured in an accident] case more closely. On reflection, it appears that these consequences are not so self-evident after all. For example, do these consequences include being left to the side of the road? Even if not, is the strength of the obligation on passers-by an obligation to take him to a hospital conditional on the gravity of Bert's condition and/or on the costs, to them, of taking him to a hospital? Should Bert pay for treatment only of those injuries that resulted from the accident itself, or also for medical conditions that resulted from the unforeseeable effect of the accident on certain hitherto unknown predispositions to illnesses? Or even for any medical treatment he will need henceforth? And at what price should the treatment be charged, so that that price may also be deemed "a consequence of his action"? (May a hospital have a policy of charging imprudent motorcyclists more than others?) Are the consequences of Bert's action also that passers-by may appropriate his motorbike from the side of the road? May he lose his job if, once he has recovered from his accident, his limpness makes him a less attractive employee? May he be denied life insurance henceforth?
>
> The list of questions could go on.[21]

So, the responsibility-tracking intuition by itself does not determine what consequences should follow from outcome responsibility. All sorts of consequences issue from a given outcome, and we need to rule out those that are "unduly harsh towards those who end up in dire straits through their own choices"[22] as well as those that "are of the wrong, because irrelevant, sort".[23] Only if we can do this, Olsaretti maintains, are the inequalities that result from people's choices justified on luck egalitarian terms. Hence, in addition to the responsibility-tracking intuition luck egalitarians need what Olsaretti calls a "principle of stakes" which does just that.

The trouble, though, is, according to Olsaretti, that no account of a principle of stakes works, and that, therefore, luck egalitarianism should be rejected.

I am sure Olsaretti is right about most of the candidates for a principle of stakes she considers and that they do not serve to supplement the responsibility-tracking intuition in the requisite way. However, she also considers and rejects desert as a principle of stakes, and here, naturally, I beg to differ from her. Though she clearly sees that there is an advantage to the desert view, namely a "proportionality constraint" which rules out unduly harsh consequences, this is to "deliver too little . . . for too high a price",[24] because desert by itself cannot determine what consequences outcome responsibility might justifiably have. Rather, outside considerations are needed, so that this commits us to a view of responsibility as "[uniquely entailing] one's own favoured account of stakes".[25] In the case of desert as a principle

[20] Olsaretti (2009:167, 169).
[21] Olsaretti (2009:172).
[22] Olsaretti (2009:166).
[23] Olsaretti (2009:183).
[24] Olsaretti (2009:185).
[25] Olsaretti (2009:186).

of stakes, this means that, since, "as even defenders of desert have been willing to grant", what someone deserves "is settled by the institutional context in which desert claims are made, rather than by the notion of desert itself".[26] Hence, adopting desert as luck egalitarianism's principle of stakes would mean understanding responsibility as "entailing" institutions that settle desert and, by extension, responsibility, which is said "too high a price", therefore we should not adopt it, and so luck egalitarianism fails.[27]

There are several problems with this line of argument, though. First of all, here is a defender of desert who is not willing to grant that what someone deserves is in every case settled within an institutional context. Sometimes this is the case, but more often than not, it is not. And even the institutional cases of desert are not intelligible without understanding desert preinstitutionally. So, no, desert is not essentially institutional.[28]

Secondly, as I have explained in the foregoing sections, desert involves far more than "proportionality", by which I take Olsaretti to mean what I call "fittingness". While this is a central component, as should be evident by now, it is not all there is to it. Desert is not only fittingness, but required fittingness, that is propriety. Hence, to adopt desert as the principle of stakes for liability responsibility allows us not only to find out which consequences are fitting that a subject bear for her outcome responsibility, but also requires her bearing them. So, the link between outcome responsibility and liability responsibility provided by desert also comes with the requisite normative force. Hence, adopting desert as the principle of stakes does not provide "too little", but just the thing required to serve luck egalitarianism's purposes.

Finally, Olsaretti has a problem with desert as a principle of stakes, because she mistakenly believes that what someone deserves is settled by institutions "rather than by the notion of desert itself", which leads her to conclude that one needs to understand responsibility as entailing institutions. Presumably, if what someone deserved was settled by "the notion of desert itself" the problem would not arise. However, as should be clear from the preceding discussion of desert, "the notion of desert itself" is a formally normative relation that by itself does not determine anything. What desert object a desert subject deserves is in every case, institutional or not, settled by the desert base and nothing else. To say that the subject deserves it is not to settle anything, it is just to say that it is appropriate that she get it. Hence, to demand that what someone deserves to be settled "by the notion of desert itself", because otherwise it cannot serve as a principle of stakes for luck egalitarianism, is unreasonable.

Incidentally, as Vincent shows, the same is true of the notion of responsibility: it "only provides a formal structure within which ... other normative considerations determine how people may be treated, but contrary to what most people seem to

[26] Olsaretti (2009:185).

[27] Cf. Olsaretti (2009:186).

[28] For a full discussion, cf. Abad (2007:16–9).

think responsibility does not generate practical demands of its own".[29] So treating the question of the principle of stakes as "a question of what *responsibility itself* requires",[30] as Olsaretti does, is equally unreasonable.

Thus, that other considerations settle what someone deserves, and by extension what they are liability responsible for, is not to understand responsibility as entailing those other things. This, indeed, would be "too high a price" and absurd too, but since both responsibility and desert are just formal relations that provide normative links between different things, the problem does not arise. No relation conceptually entails the individuals it relates to each other.

These considerations certainly do commit luck egalitarianism to "a particular view about the principle identifying the grounds of responsibility", namely the desert view.[31] I fail to see, though, how this constitutes "too high a price", because this is just what a principle of stakes is supposed to do on Olsaretti's own terms. A problem arises only if we do not keep in mind what sort of responsibility we are talking about here. Obviously, desert cannot ground outcome responsibility. To say that I am only outcome responsible for what I deserve to bring about is bizarre. However, we are not talking about outcome responsibility here, but about liability responsibility. So, in looking for a way to normatively link consequences that are neither unduly harsh nor irrelevant to certain states of affairs someone is outcome responsible for, we are indeed looking for a principle of stakes that identifies the grounds of liability responsibility, and desert can fulfill this need. Far from being too high a price then, this is just what Olsaretti said luck egalitarianism needed.

So, if luck egalitarianism adopts desert as the principle of stakes, then, *contra* Olsaretti, it can rule out consequences that are unduly harsh or irrelevant as consequences someone should be liability responsible for given her outcome responsibility. However, luck egalitarianism still needs to establish that liability responsibility really does follow from outcome responsibility in the first place.

8.2.2 How to Derive Liability Responsibility from Outcome Responsibility

As I have already said, Vincent shows that outcome responsibility and liability responsibility are two different responsibility concepts. Not only do they refer to different objects, but they also are differently orientated in time: while outcome responsibility looks backwards, liability responsibility looks forward. Hence, outcome responsibility refers to a state of affairs one has brought about in the past, whereas liability responsibility refers to consequences one will bear in the future.

Since outcome responsibility and liability responsibility are two quite different concepts of responsibility, Vincent argues that the one does not follow from the other

[29] Vincent (2009:49).

[30] Olsaretti (2009:186).

[31] Olsaretti (2009:185).

"automatically", that is it is not logically entailed, as the responsibility-tracking intuition would have it. Rather, it would need additional normative premises to "bridge the inference gap", for example something like: "those who are outcome responsible for X should take liability responsibility in manner Y". Vincent calls such additional normative premises "reactive norms, since they are norms that govern our reactions to outcome responsible parties".[32] However, luck egalitarians do not offer any reactive norms as bridging premises, but simply assume that liability responsibility automatically follows from outcome responsibility. Since this is wrong, Vincent concludes that the responsibility-tracking intuition must be rejected.[33]

This conclusion is surprising in its abruptness, because, first, we could just have luck egalitarians read Vincent's analysis and surely they would immediately recognize the need to offer reactive norms and do so. Secondly, Vincent herself thinks that there are reactive norms that bridge the inference gap since there are "normative considerations that ... play a key role in validating the transition from claims about a person's outcome responsibility to conclusions about their liability responsibility".[34] Thirdly, the responsibility-tracking intuition is a very powerful intuition many people share, and not just luck egalitarians. We should be wary of discarding it just like that, but rather see whether there is a way of retaining it without running into the problems Vincent points out.

Regarding the first point that luck egalitarians might just agree with Vincent and belatedly offer reactive norms to supplement their theory, Vincent might reply that this essentially is to give up luck egalitarianism, because, to borrow from Olsaretti, the idea behind luck egalitarianism presumably is to see only those inequalities as justified that can be derived from a liberal concept of freedom of choice and the concept of responsibility which are compatible with luck egalitarianism.[35] Thus, the idea seems to be to start out from concepts as thin as can be so as to get as widespread approval as possible and justify inequalities from there. Suggesting to simply add some reactive norms really amounts to abandoning this underlying idea, because all of a sudden the starting points do not seem to be so thin and universally approvable anymore.

Well, if luck egalitarians really are as inflexible as all that, so much the worse for them, of course. However, I am not convinced that there might not be a more yielding kind of luck egalitarian who would not mind a spot of extra justification for the odd reactive norm or two supplementing the responsibility-tracking intuition. Still, even if there is not, luck egalitarians are not the only people who want to derive liability responsibility from outcome responsibility. In fact, and this is the second point, Vincent herself seems to want to do just that, and she introduces reactive norms to do the very trick.

[32] Vincent (2009:47).

[33] Vincent (2009:46–8).

[34] Vincent (2009:49).

[35] Cf. Olsaretti (2009:179).

And to be sure, reactive norms will serve. Vincent correctly argues that we need some additional premises to bridge the inference gap from outcome responsibility to liability responsibility in the responsibility-tracking intuition, and her reactive norms do that. But she also sees that "this now raises the question of where such reactive norms might come from". So, she recognizes the need to justify those norms, because they may allow treating people harshly, and this is where normative considerations of, say, justice, utility, caring, beneficence and so on come in that inform those norms.[36]

However, this will not do. The problem with Vincent's account of reactive norms as bridging premises is that they are arbitrary. Reactive norms would not only provide a link between outcome responsibility and liability responsibility but also spell out the consequences one is liability responsible for. However, as we have seen in the last section, it is important that this be a link such that it connects only the right sort of consequence to outcome responsibility.

Informing reactive norms by normative considerations like utility or caring does not help in this regard. Say we care about our fellow citizens' safety so much that we really want them to wear helmets when riding their motorbikes so as to spare them the ghastly consequences of possible head injuries. Moreover, those head injuries are a real strain on our health insurance system. In order to achieve this we institute a reactive norm that motorcyclists who have an accident while riding without a helmet be ordered to scrape chewing gum from underneath school desks as soon as they recover. This is both caring and not at all too harsh, as almost every motorcyclist is bullied into wearing a helmet by the prospect of this distasteful task resulting in a significant drop in those specific head injuries, and those who are not at least do something useful to pay something back to society. The trouble is that this consequence of gum-scraping is not covered by the responsibility-tracking intuition. What is powerful about the intuition, even though it employs two different responsibility concepts, is that we intuitively see that there is something to the idea that people who are in a situation through their own fault should deal with it themselves, that they should suffer the consequences. However, in the example it is all too natural to ask, "Why gum-scraping?" So, the intuition does not cover any consequences that may arise out of situations people are in through their own fault, but only those that are linked to them in some special way. Linking them through a reactive norm does not work, as there is a reactive norm in place in the example, but we still do not intuitively see that a reckless motorcyclist really should scrape gums. This might be covered by the reactive norm, but it is not covered by the responsibility-tracking intuition.

Therefore, while reactive norms can serve as normative premises bridging the inference gap from outcome responsibility to liability responsibility, they can do so only in an arbitrary way, and thus fail as an explanation of the normative force of the responsibility-tracking intuition.

[36] Vincent (2009:47).

In view of this, Vincent might be tempted to reject the responsibility-tracking intuition altogether: we cannot derive liability responsibility from outcome responsibility automatically, we need reactive norms as additional bridging premises. Reactive norms fail to do the trick, so let us chuck the responsibility-tracking intuition and face it – liability responsibility cannot be derived from outcome responsibility at all.[37] But this would be too easy a dismissal.

Whether or not Vincent believes in the responsibility-tracking intuition, many, if not most, people do. It is such a powerful intuition, and this is the third point mentioned at the outset of this discussion, that we should be loath to go to such lengths as rejecting it if not necessary. And it is not, for we can tell which specific consequences are responses to certain states of affairs they issue from, and which are not. In the first case, the consequences are fitting to what happened before, in the latter, they are not. And the force of the responsibility-tracking intuition, its being so widespread and powerful, indicates that they are not only fitting, but required, hence appropriate. So, as easy as that, we can derive liability responsibility from outcome responsibility: a person should take liability responsibility for a state of affairs she is outcome responsible for, because that is the appropriate thing for her to do.

So, there is a certain link between reckless, helmetless motorcyclists getting into accidents and their paying their hospital bills themselves, namely the latter being a response to the former, while there is no such link – other things being equal – between their getting helmetless into accidents and scraping gum from underneath school desks. Hence, with the help of the notion of a "response" we get what consequences are fitting to what states of affairs. And the responsibility-tracking is evidence for our believing that they ought to be thus linked, that they are not only fitting but appropriate.[38]

We can probably even say that, given the dangers of motorcycling and the fragility of the human body, motorcyclists do something unfitting when they drive without helmets, and therefore, that it is not only appropriate but that they deserve to pay their own hospital bills. Obviously, though, whether or not we can say this, whether the person outcome responsible deserves to take liability reponsibility, depends on the responsibility-tracking example. At the least, however, it is appropriate for her to do so.

Some might worry that this is nothing but an exercise in terminology, that I have invented a new name for the responsibility-tracking intuition, but have not justified it. The last bit of this is true, I have not justified the responsibility-tracking intuition, but that is because that is not my business here. Rather, I have tried to make sense of how exactly to derive liability responsibility from outcome responsibility. In order to do this, I have simply taken for granted that liability responsibility should

[37] I do not get this impression from her paper, but in private correspondence Vincent leans that way.

[38] I do not mean to say that this link between outcome responsibility and liability responsibility holds necessarily. As I have argued in the first part of this paper, to say this would need substantial arguments for each situation in which such a link is said to hold. The present discussion, though, is not on this point but on luck egalitarianism, and for the sake of this discussion, I will conveniently assume that we are talking about a situation in which this link does hold.

follow from outcome responsibility. I have done so on evidence of the responsibility-tracking intuition's plausibility and force. The trouble was, as Vincent showed, that the intuition is not enough to derive liability responsibility from outcome responsibility, it needs to be supplemented. Vincent tries to do this with reactive norms, but that will not work. The foregoing considerations show that, rather, the responsibility-tracking intuition needs to be supplemented by the notion of propriety. This is not to give just a new name to the game. It is not to say that persons outcome responsible ought to take liability responsibility because they ought to, it is to say that they ought to because it is appropriate. And that this is not the same is what I have shown in the first part of this paper.

8.2.3 Two Questions or One?

Now it seems as though in the last two sections I have given two rather similar answers to the two quite distinct questions Olsaretti and Vincent focus on, that is on the one hand the question how the transition from outcome responsibility to liability responsibility works, and on the other hand the further question, assuming the transition does work, what consequences one should be liable for given one's outcome responsibility. In answer to both questions I have argued that desert will do the respective trick. So, does it not seem as though I have become confused about what the difference between the two questions is somewhere along the way? Well, no. Rather, the concept of desert is so great that it is capable of doing both tricks at once. Let us see how this works.

Remember that the concept of desert is made up of two components, fittingness and a normative element which makes what is fitting appropriate, that is which requires that the desert subject, characterized by the desert base, get the fitting desert object. The second question, what consequences one should be liable for given one's outcome responsibility, is the question what desert object is a response to a particular desert base a particular desert subject has set. The first question, how liability responsibility follows from outcome responsibility, is the question of the normative force of propriety in this case. I have not specially argued that there is such normative force there in this case, but have simply taken the power of the responsibility-tracking intuition as sufficient evidence for its holding here.

In other words, in the desert relation "x deserves y in virtue of z", the second question concerns the fittingness between y as liability responsibility and z as outcome responsibility of subject x, whereas the first question concerns the transition marked by "deserves".[39]

So, rather than having confused the two questions by employing the concept of desert in the discussion of luck egalitarianism, I have killed two birds with one stone.

[39] I have already said in the last section that whether liability responsibility given some outcome responsibility is a case of desert or of propriety is a matter of the particular constitution of the outcome responsibility a.k.a. desert base, and also that nothing hangs normatively on this conceptual distinction.

8.2.4 Luck Egalitarianism

Now, what does all of this mean for luck egalitarianism? Vincent and Olsaretti reject luck egalitarianism, or the responsibility-tracking intuition it rests on the, grounds that luck egalitarians cannot provide answers to the two questions mentioned. I have shown that luck egalitarianism cannot be rejected for that reason, because, if luck egalitarianism is supplemented by the concept of desert as explained, it can provide such answers after all. However, this does not mean that desert is all in the luck egalitarians' court. Luck egalitarians may employ this concept in order to draw the conclusions they want to draw, but that does not mean that desert may not equally well be employed by luck egalitarianism's opponents.

This is so because, as I have now said repeatedly, the *concept* of desert by itself does not determine who deserves what for what, and because, as I have emphasized all along, the concept of desert does not presuppose the concept of responsibility.

So, while it is reasonable to suppose that an unemployed gambler who has lost all his savings in a casino (who is outcome responsible for this) deserves to be barred from welfare aid (ought to take liability responsibility in this manner) it is equally reasonable to suppose that the gambler's situation now being thus that he has no adequate access to food, medicine, and shelter, he deserves to get welfare aid. This being equally reasonable means that in both cases the consequences are equally good responses to the antecedent states of affairs, so that we can say that the gambler both deserves to get welfare aid in virtue of his situation and deserves to be barred from welfare aid in virtue of his outcome responsibility.

Luck egalitarians, therefore, cannot simply point to the gambler's desert in order to draw the conclusion that he ought to be barred from welfare aid, because there are other considerations of desert to be taken into account, and since all of these are equally considerations of desert none of them are more weighty than the others simply on the grounds of their being deserved. So, in this regard the concept of desert will not help the luck egalitarian.

Neither can luck egalitarians point out that it is the gambler's own fault, the gambler is liability responsible and so he should be barred from welfare aid no matter what other desert considerations count against that, because those other desert considerations are not based on responsibility. The answer to that would be: so what? What on Earth is so very special about responsibility that it should break the tie between different considerations of desert, the normative force of which I have already shown does not rest in the slightest on responsibility? We would need additional arguments for that.

Nor can luck egalitarians point to the concept of justice and say that the gambler's deserving to be barred from welfare aid is a matter of justice whereas the gambler's deserving welfare aid is a matter of charity, and justice outweighs charity. Again, this would need some substantial arguments, that justice is only concerned with what people are responsible for, which have as of yet not been given, and that would be a counterintuitively thin notion of justice anyway.[40] After all, it is not

[40] Cf. Vincent (2009:49).

implausible to suppose that giving people what they deserve is always a matter of justice. Moreover, even if the one really was a matter of justice and the other a matter of charity – whoever said that justice was more important?

Well, the luck egalitarian might say, all of this is well and good, and if this was an ideal world in which there was enough for everyone no problem would arise. Alas, our resources are scarce, there is not enough for everyone, not even enough for everyone deserving, so that unfortunately we need to limit distribution in some way. Surely, it is preferable to do so in a just way. And distributing resources according to desert based on outcome responsibility is such a just way after all. So, given the scarcity of resources, such distribution is justified since at least I have granted that it is just.

Certainly, but this would be just to settle the matter by pragmatic considerations. This does not make picking out desert based on outcome responsibility of all desert considerations any more warranted in terms of justice, and it does not settle the matter regarding other possible just ways to distribute scarce resources, that is, this consideration by itself does not show that from among various just ways to distribute scarce resources we should pick the luck egalitarian option.

8.3 Conclusion

So, there is no conceptually necessary link between desert and responsibility. The responsibility view of desert is wrong. Nevertheless, if luck egalitarians avail themselves of the concept of desert, they thereby have an instrument with the help of which they can both explain how to derive liability responsibility from outcome responsibility and what specific consequences one is liability responsible for given one's outcome responsibility. However, it is important for them to note that this is only an instrument with the help of which they can explain these things. The concept of desert is not there to do all their argumentative work for them. They still need to provide substantial arguments to support their conclusions, and so far, these arguments have not been forthcoming.

Acknowledgments This is a revised version of a paper I presented at the conference "Moral responsibility: Neuroscience, organization & engineering" on August 24–27, 2009 in Delft. I thank Rüdiger Bittner, Logi Gunnarsson, Martina Herrmann, Ute Kruse-Ebeling, Nicole Vincent, and the audience at the conference for their help.

References

Abad, Diana. 2007. *Keeping Balance. On Desert and Propriety*. Heusenstamm: Ontos.
Bittner, Rüdiger. 2001. *Doing Things for Reasons*. Oxford: Oxford University Press.
Cupit, Geoffrey. 1996. "Desert and Responsibility." *Canadian Journal of Philosophy* 26:83–99.
Feinberg, Joel. 1963. "Justice and Personal Desert." First published in *Nomos VI: Justice*, edited by Carl J. Friedrich and John W. Chapman, 69–97. New York, NY: Atherton Press. Quoted from *Doing and Deserving. Essays in the Theory of Responsibility*, Joel Feinberg, 55–94. (Princeton, NJ: Princeton University Press, 1970).

Feldman, Fred. 1995. “Desert: Reconsideration of Some Received Wisdom.” First published in *Mind* 104: 63–77. Quoted from *Utilitarianism, Hedonism, and Desert. Essays in Moral Philosophy*, Fred Feldman, 177–92. (Cambridge: Cambridge University Press, 1997).

Lamont, Julian. 1994. “The Concept of Desert in Distributive Justice.” *The Philosophical Quarterly* 44:45–64.

McLeod, Owen. 1999. “Contemporary Interpretations of Desert. Introduction.” In *What Do We Deserve?* edited by Louis P. Pojman and Owen McLeod, 61–9. Oxford: Oxford University Press.

Olsaretti, Serena. 2009. “Responsibility and the Consequences of Choice.” *Proceedings of the Aristotelian Society* 109:165–88.

Pogge, Thomas W. 1989. *Realizing Rawls*. Ithaca, NY: Cornell University Press.

Rachels, James. 1978. What People Deserve. In *Justice and Economic Distribution*, edited by John Arthur and William H. Shaw, 167–96. Englewood Cliffs, NJ: Prentice-Hall.

Rachels, James. 1986. *The Elements of Moral Philosophy*. New York, NY: Random House.

Rawls, John. 1971. *A Theory of Justice*. Cambridge: Belknap Press.

Sadurski, Wojciech. 1985. *Giving Desert Its Due: Social Justice and Legal Theory*. Dordrecht: Reidel.

Sher, George. 1987. *Desert*. Princeton, NJ: Princeton University Press.

Smilansky, Saul. 1996. “Responsibility and Desert: Defending the Connection.” *Mind* 105:157–63.

Vincent, Nicole A. 2009. “What Do You Mean I Should Take Responsibility for My Own Ill Health?” *Journal of Applied Ethics and Philosophy* 1:39–51.

Chapter 9
Communicative Revisionism

Lene Bomann-Larsen

Abstract Causal determinism may appear more salient when informed by empirical science. Thus even if neuroscience brings nothing substantially new to the debate on free will and human agency, it may enforce a call for revision of moral and legal practices. If our moral agency is a result of causal luck, desert-based moral practices seem unwarranted. A current trend in compatibilism agrees with the Strawsonian approach to moral responsibility in rejecting radical revisionism, but supplies the Strawsonian approach with a contractualist normative foundation. This paper argues that the contractual justification for desert-based moral practices fits well with a communicative theory of punishment. It is argued, however, that while a contractual communicative view on punishment may require some forms of hard treatment, it is unlikely that it will warrant extremely severe treatment. Hence determinism calls for a weak to moderate revision of our punitive practices insofar as these at present impose extreme suffering.

9.1 Introduction

There is some dispute over whether neuroscience brings anything substantially new to the compatibilist/incompatibilist debate on the significance of determinism for moral responsibility, and, as a corollary, whether neuroscience should have any impact on criminal law. Some argue that its impact will be substantial, because hard science has the weight and the prerogative to convince both decision-makers and the general public about what smart philosophers have known for a long time, namely that since there is no free will, and retributive punishment requires freedom of the will, we have to revise our justification for punishment (Vargas 2005; Greene and Cohen 2004), or eliminate punishment altogether (Pereboom 2007). Others maintain that the law makes no assumptions about free will, merely about intent and competence. Intent and competence, to be sure, do not hinge on free will; hence there is no reason why determinism either in its pure philosophical or empirical form should have any revisionary potential with regard to punishment (Morse 2004).

L. Bomann-Larsen (✉)
Department of Philosophy, Classics, History of Art and Ideas, University of Oslo, Oslo, Norway
e-mail: lene.bomann-larsen@hf.uio.no

N.A Vincent et al. (eds.), *Moral Responsibility*, Library of Ethics and Applied Philosophy 27, DOI 10.1007/978-94-007-1878-4_9,

However, even granted that the law makes no assumptions about free will and that the penal institution as such is not threatened by neuroscience, some concerns arise in the wake of scientifically informed determinism that are worthy of consideration. For one, the legitimacy of the penal institution rests partly on its correspondence with a "common sense of justice", and this common sense of justice is arguably informed by folk intuitions about free will (Greene and Cohen 2004). These intuitions may be altered, or deemed illusionary when the truth about human mind and agency revealed by neuroscience hits the public (ibid.). Though we cannot help but act as if our actions are "up to us", in our reflective moments we will realize that up-to-usness is illusory (Smilansky 2000), and hence cannot serve as the basis for penal practices. Second, since punishment by definition involves deliberately imposing suffering on a person, justifying it is a difficult task even in the absence of deterministic challenges. Science enforcing causal determinism does not make the task easier. Even if determining who is "guilty" and who is not in legal terms does not make any explicit references to free will, punishing the guilty appears to make such reference, insofar as its justification at least partly relies on "desert", which seems inextricably connected to free will. This interconnection is described by Galen Strawson as follows:

> What is it to be capable of being truly deserving of praise and blame for one's actions? . . . Given that an agent is a moral agent, it is capable of being truly deserving of praise and blame for its choices and actions when and only when it is capable of free choice and free action. (1986:1)

To be deserving of praise and blame, that is, to be morally responsible in "the full sense", is described by Derk Pereboom[1] thus:

> For an agent to be morally responsible for an action is for it to belong to her in such a way that she would deserve blame if she understood that it was morally wrong, and she would deserve credit or praise if she understood that it was morally exemplary. The desert at issue here is basic in the sense that the agent, to be morally responsible, would deserve the blame or credit just because she has performed the action . . . and not by virtue of consequentialist considerations. (2007:86)

Desert-entailing responsibility then, is a backward-looking form of responsibility; "Desert is concerned with what the person deserves to get for what she has done" (Smilansky 2000:2). It is thus closely connected to the person; we do not simply judge the action A as bad and point to X as the causal agent of A, but claim that X is bad for doing A, where A belongs to X in some intimate way.

Compatibilists about determinism and moral responsibility may recognize the force of the argument that if our actions are not really "up to us", we cannot be truly deserving of praise and blame, and yet attempt to salvage some concept of desert that does not require ultimate up-to-usness; but which still suffices to warrant our retributive attitudes and the practices that express them. One such (quasi-compatibilist)

[1] Pereboom himself takes it that we cannot have this kind of responsibility under determinism. To him, full moral responsibility requires ultimate control.

attempt consists in giving a contractualist,[2] or interest-based, argument for the legitimacy of our moral practices, which warrants maintaining them largely, if not wholly, unrevised. Considering the kind of creatures we are and what we tend to value – not least our moral agency and the ownership of our own actions – and what we stand to lose, were we to jettison our moral agency, it seems plausible that it is in our interest to leave our moral practices largely intact. Within such a contractualist scheme, the attribution of desert-entailing responsibility is justified with reference to the interest people take in being treated as responsible moral agents whose actions are up to them; that is, in being treated as moral agents for those types of actions for which they are willing to assume responsibility.

However, even if the contractualist argument succeeds in justifying, and hence leaves largely unrevised, our informal moral practices, the notion of desert it engenders may not be sufficiently robust to leave unrevised our *punitive* practices. With a substantive notion of "true desert",[3] which implies that the agent was fully free to choose to do otherwise[4] and still decided to do wrong, it is easier to see hard treatment as something intrinsically *good*. A contractualist notion of desert, on the other hand, requires that the justification of hard treatment must also find its basis in some form of agreement on what is a reasonable mode and scope of punishing those of us who end up as criminals, given the deterministic presumption that our actions are ultimately a matter of luck, in spite of how much we value perceiving of our actions as "up to us". Imagining a veil of ignorance behind which we do not know whether we will be causally lucky or unlucky in terms of moral agency is a useful thought-experiment.[5] My thesis is that under such conditions, we would choose a scheme where punishment is expressed as a form of moral censure, that is, a backward-looking form of punishment connected to "desert", but one in which punishment does not impose severe or extreme suffering (e.g. life-sentences, capital punishment).

This paper argues in support of three claims: (1) the contractualist argument establishes a basis for a publicly instituted censure of wrongdoing, aiming at addressing the offender as a moral agent and offering him an opportunity to repent and reconcile himself with his victim and society; (2) this censure must be expressed in terms of hard treatment; but (3) severe treatment, i.e. imposing extreme suffering, will not be warranted.

---

[2] The "contractualism" I am referring to here is of a Scanlonian kind (Scanlon 2003). It is thus constrained by some normative notions such as fairness and reasonableness. It is not a rationality-based contractualism like Rawls' Original Position, according to which contractors are guided by enlightened self-interest in ignorance of their position in society.

[3] As expounded in the quote from Galen Strawson above.

[4] Not necessarily in terms of the Principle of Alternative Possibilities, but at least in terms of having ultimate control over one's actions by having chosen to be the person who acts in the way one does. Cf. Galen Strawson's Basic Argument (2003).

[5] When I introduce the metaphor of the Veil of Ignorance, this is not intended to be a shift to a Rawlsian situation of choice, but aims merely to illustrate the radical character of the requirement that we should place ourselves at the receiving end of the principles we hold, in order to determine whether they are acceptable.

In Section 9.2, I give a brief outline of a contractualist theory of desert. In Section 9.3 I offer an argument for why contractualist desert is fitted to a communicative theory of punishment (Duff 2001), where the public expression of condemnation (Feinberg 1970) is combined with addressing the offender as a moral agent. In Section 9.4 I sketch the conditions for determining an appropriate mode and scope of punishment. I suggest that under conditions where we are ignorant of how we fare as moral agents, we will choose a form of censure which (1) goes beyond mere symbolic censure and thus involves some form of hard treatment, but which (2) does not impose extreme hardships or irreparable suffering. I also consider, and reject, some objections to both these claims. Section 9.5 concludes the paper.[6]

9.2 Justifying Desert in Contractualist Terms

There is a trend in contemporary compatibilism to provide a Strawsonian conservative approach (P. Strawson 2003) to moral responsibility attribution with a contractualist, or interest-based (Lenman 2006; Bomann-Larsen 2010), normative foundation. The contractualist approach recognizes the force of our reactive attitudes, but also recognizes that there is a problematic inference from the fact that we have and value these attitudes, to allowing these attitudes to justify our moral and punitive practices. By invoking our interest in maintaining our practices, their normative basis is secured because reference to our interests gives the opportunity to fend off accusations of unfairness; it is not unfair to people to treat them as morally responsible agents if it is in their interest to be so treated.

To the contrary, it would be unfair *not* to, because that would amount to an unjustifiable dismissal of the claim of persons to have their legitimate interests taken into account. The alternative to treating a person on the basis of *intrinsic* features of herself or her act (Smilansky 2000:1) is to treat her on account of *extrinsic* features, such as concern for the general good.[7] The wrongdoer will be objectivised, regarded

[6] I assume that everything argued in this paper is compatible with even the strongest possible threat from neuroscientifically informed determinism; what may be roughly labeled as the No-Action-Thesis (Morse 2004). Should it turn out to be true that conscious willing is in fact an illusion; i.e. that we merely *feel* like agents (Wegner 2002), we can still want to be treated as the agents we feel like being. Illusionism is fully compatible with a contract-based justification for desert and a communicative theory of punishment; insofar as it is unavoidable to maintain a pragmatically meaningful notion of "intent", and insofar as our agency – illusory or not – *matters* to us. Of course, the argument requires that we are reason-responsive, but even illusionists about conscious willing have to admit that we are reason-responsive at some level, if they are to argue that punishment could work as a deterrent. (Indeed if they are to argue at all without performative contradiction).

[7] It could be objected that the kind of determinism which is presupposed here undermines the distinction between intrinsic and extrinsic features, because if true, all features are in the relevant sense extrinsic. But we do not need to do away with the distinction. Consider for example Neil Levy's argument that even if we do not make conscious decisions, "the mechanisms that make the decisions are . . . ours; they have our values, our beliefs, our goals (we have them by them having

as an item to be manipulated or a problem to be managed, or treated in a paternalistic way, as a misbehaving child. And since it is presumably not in our interest to be subjected to rules treating us in this way, the contractualist argument concludes that this form of treatment is wrong.

The contractualist argument provides a justification for treating each other as morally responsible agents in *the full sense*, even on the condition that we are not, in the metaphysical sense, free. Desert-entailing responsibility may be burdensome; as its corollary is blame, resentment and sanctions which are unpleasant to experience, and particularly so when they reflect on the *person*, not simply on the action viewed in agent-neutral terms. Yet, even if it may sometimes be contrary to our interest at a given time t_0 to be treated as full moral agents, it is in general, our lives seen as a whole, in our interest to be treated as such. The reason is that desert-entailing responsibility has another side as well; moral praise and credit. And we cannot have the one without the other; we are rationally committed to hold that we either *are* the kind of beings to whom moral appraisal applies, or we are not. The burdensome upshot of being treated as a moral agent to whom praise and credit applies when one has done something worthy of these responses, is to be subject to resentment, blame and perhaps even punishment when one has done something that is worthy of those responses. But it is a price we are willing to pay, because being treated as moral agents is extremely important to us. As Dennett notes:

> Blame is the price we pay for credit, and we pay it gladly under most circumstances. We pay dearly, accepting punishment and public humiliation for a chance to get back in the game after we have been caught in some transgression. (2003:292)

Moreover, we may even challenge the assumption that we only value negative responses because they are corollaries of positive responses: it is not given that simply because being subjected to blame and resentment is unpleasant, we do not value being subjected to negative moral sanctions when we know that we have done something wrong. In cases where we strongly feel remorse, we may appreciate having our self-reactive attitudes asserted by others, and to be given the opportunity to set things right by sincerely repenting and paying back what we owe.

But how far can the contractualist argument get us in terms of moral responsibility? It can be argued that it does not get us any form of *desert*; it merely suggests a *derivative* notion of moral responsibility (Smart 2003). A hard incompatibilist like Pereboom (2007) accepts that we can attribute responsibility for morally wrongful actions in such a way that we can express moral criticism; that is, in such as way that it makes sense for X to regret that she did A, and opt for improvement in the future (but not in such as way as to assert that X deserves reproach for doing A). Such an approach leaves room for backward-looking apology and forward-looking reform, but not for blame and punitive sanctions. However, as soon as this much is admitted, it seems we already have the raw materials we need to establish a weak

---

them) and when they decide, *we* decide." (Levy 2007:243). Those causal forces that are me, then, are intrinsic features that may be distinguished from causal forces operating outside of me (our outside of my consciousness) and which cannot be attributed to me as a person.

notion of desert. Firstly, it does not make sense to ask X to apologise *sincerely* for a past action without asserting that the action belongs to her in some intimate way. Secondly, it does not make sense to expect moral reform from X with regard to future actions without X herself sincerely taking her past actions to belong to her in some intimate way (Bomann-Larsen 2010).

Respecting a moral agent as such requires some form of correspondence between first-person and third-person responses; that is, between responses to one's own actions versus responses to other people's actions. First-person responses come in (at least) two forms: agent-remorse and agent-regret (Williams 1981). Agent-remorse is the form of self-blame that is the upshot of understanding that one has culpably (intentionally or by a form of culpable negligence or recklessness) done wrong. The appropriate third-person correlate to agent-remorse is blame.[8] Agent-regret, on the other hand, responds to the thought "How much better if it had been otherwise" (ibid.:27), and so does spectator-regret; the spectator, too, will join in and say: "How much better indeed." Were we to reject desert-entailing responsibility, we would be left with only one third-person response; spectator-regret, and ideally with only one form of first-person response; agent-regret. But as long as agents are likely to insist on agent-remorse, the consequence of jettisoning desert-based moral practice is a radical asymmetry between how the agent views her own action as something intrinsically belonging to her, and the third-person response of spectator-regret treating her action as some kind of unfortunate event. As response-types, neither agent-regret nor spectator-regret are specifically directed towards persons (self or other) as *moral* agents; they conflate moral and causal agency and thus conflate intentional actions and mere accidents. Regret-responses therefore do not sufficiently express respect for persons as the moral agents they take an interest in being treated as by others.

The contractualist view gives a justification for leaving unrevised those elements of our moral practices that relate to desert. Let us now turn to how a contractualist justification of desert can inform a theory of punishment, which, arguably, is harder to justify because it involves a stronger imposition of suffering on the agent than does informal moral sanctioning.

9.3 Determinism and Theories of Punishment

Criminal law and the institutionalisation of criminal punishment may be viewed as the public expression of blame. Certain types of moral wrongs are a public concern. I am here relying on Antony Duff's account of crimes as public wrongs (Duff 2007), and in the present context I will only be concerned with punishment for those crimes that are properly speaking also moral wrongs, and which are crimes for that reason; that is, which are condemned *because* they are morally wrongful (as opposed, for instance, to violations of administrative regulations, though it may also be considered morally wrong to violate these because it is wrong to violate legitimate laws

[8] Whether actually expressing blame is appropriate will depend on other factors as well.

(see ibid.:81, 89–93). In short, I am concerned with the type of crimes that can be classified as *mala in se*; wrongful independently of their formal status as crimes.[9] It is mainly for such actions that moral agents may be asked to repent, out of respect for their own moral agency, as well as out of respect for their victims and the moral community as a whole.

If we agree that there is a reason, in contractualist terms, for people to accept desert-entailing moral responsibility, then we have established the *foundations* of a backward-looking theory of punishment that ties the offender to his punishment in terms of desert; in short, the foundations of a retributive theory. That is, we have established the basis for a theory of punishment which addresses past wrongdoing, justified in terms of what we owe to the offender and to the victim of wrongdoing, according to which properly addressing wrongdoing involves treating wrongdoers as *deserving* of censure. This fits well with an expressive or communicative theory of punishment, where, according to Feinberg:

> Punishment is a conventional device for the expression of attitudes of resentment and indignation, and of judgments of disapproval and reprobation, on the part either of the punishing authority or of those "in whose name" the punishment is inflicted. (Feinberg 1970, quoted in Duff 2001:27)

Duff further adds that the communicative theory of punishment

> . . .communicate(s) the censure that offenders deserve by portraying punishment as a species of *secular penance*. That account is retributivist: it justifies punishment as the communication of deserved censure (Duff 2001:30).

However, communicative punishment also contains three forward-looking elements in aiming at persuading the offender to *repent*, and hopefully *reform* himself and be *reconciled* with the community (ibid.).

In arguing that a communicative theory of punishment is suited to the contractualist argument for desert, it must be shown why it is better suited than alternative theories. The contractualist argument justifies choosing one scheme in terms of its desirability compared to alternative schemes. Successfully arguing in favour of a communicative theory of punishment therefore partly depends on the lesser desirability of conceivable alternatives.

Traditionally, there are two main contenders to the retributivist justification of punishment solely in terms of what "the guilty deserve to suffer" (Duff 2001)[10]:

---

[9] Though I remain, at present, neutral on the question of how to identify which wrongs are public wrongs, i.e. whether these are to be picked out by a singular principle, such as the harm-principle, or by plural principles, or in terms of communal values (Duff 2001, 2007).

[10] Note that nothing I say in favour of a communicative justification of punishment with reference to desert is intended to set aside the need for protecting society from dangerous offenders. Should the offender be dangerous, he might have to be detained, but the justification for that could be of a different kind – protection of community, not punishment – and should, if so, only be used to the extent necessary for this purpose, and under as humane conditions as possible. For example, we already detain dangerous psychotics, but we do not punish them. The need for protection may come in addition to the retributive element in an expressive or communicative theory, but with a different and not necessarily punitive justification. The question is how much of the material component of

pure consequentialism and abolitionism; whereof, to be precise, the latter is not really a justification for punishment at all, but for its abolition, in whole or part. Abolitionism takes many forms; from the abolition of punishment altogether, to the abolition of particular forms of punishment, e.g. imprisonment as a mode of punishment (Duff 2001:31). Within the free-will debate, the equivalent of (total) abolitionism is often referred to as *eliminativism* about punishment (Pereboom 2007; see also Vargas 2005). Here, it is recommended that the very *concept* of punishment should be eliminated along with the concepts of desert and guilt. In the following I will focus on the strong revisionary claim that punishment as such – as a concept and an institution – ought to be eliminated from our moral practices.

While radically different in crucial respects, both (pure) consequentialist justifications of punishment and various claims to the elimination of punishment have one important feature in common; they see no justification for a backward-looking form of punishment. The perpetrator is not to be treated as deserving of suffering for what she has done in the past, but should be dealt with in a different way and/or for different reasons. Such reasons may be deterrent, protective, and/or corrective/reformatory. While both sets of theories recognise the need for protecting society against harmful actions, they are sceptical towards a backward-looking kind of punishment, simply because it entails no benefit for anyone, apart, perhaps, from the satisfaction of a presumably unjustifiable thirst for revenge. To be sure, both consequentialist and abolitionist theories are proposed independently of the free-will debate, for independent moral reasons, but they are also typically revisionist theories within the determinism/free-will debate, where they seem to acquire an additional edge against retributivism on the basis that if no one is free, everyone is fundamentally innocent on the basis of causal luck, and on this score retributivism cannot satisfy its own basic and absolute prohibition against punishing the innocent. Hence, the issue arising from determinism is not only that retributive punishment does not benefit anyone, but more fundamentally, that such punishment is profoundly unfair. So why bother rescuing retributivism from the threat of determinism? Why not simply bite the bullet and abandon retributivism along with its presuppositions? Why should we not, as Greene and Cohen (2004) suggest, choose – in our reflective moments – a purely forward-looking justification for punishment?

Forward-looking theories offer different answers to the luck problem. Consequentialists are all happy about punishing the innocent (i.e. the unlucky) for the sake of deterrence and protection of the community, that is, to impose suffering on someone for entirely person-extrinsic reasons. Now, it is a familiar argument against consequentialism that punishing the innocent in order to further some social aim amounts to using people as mere means, and granting the truth of determinism reinforces this objection: not only will consequentialist punishment under determinism *open* the door to punishing the innocent (which arguably could be avoided by

imposing suffering can be put into the punishment itself, and how much is left to be justified in terms of protection.

side-constraining consequentialism); it systematically *requires* it, because everyone is by definition innocent.

However, the "use as means" objection begs the question against the consequentialist, who does not believe that there is anything wrong in using people as mere means. The contractualist argument offers a better response, namely that it is hardly justifiable to a person to treat her as a mere means.[11] Hence a purely consequentialist theory of punishment is unlikely to be accepted, given the contractualist argument as a basis for our moral practices.[12] Now, it may be objected that nothing bars contractors behind a veil of ignorance from settling for consequentialist principles which then, by being chosen, are justifiable *ex hypothesi* to those subjected to them. Contractors may well choose principles that treat them as means but give a better prospect of avoiding harm. And admittedly, contractors may conceivably settle for purely consequentialist principles of punishment, but all my argument needs is to show that it is equally or perhaps more plausible that they will not. Although there may be exceptions in some cases, it is not implausible to assume that people in general prefer to be treated as ends, even though this might occasionally harm rather than benefit them, in light of the alternative of giving up their autonomy *tout court* to avoid all kinds of harms.[13]

For those who believe that it is wrong to treat people as mere means, no matter which reason they have for holding the belief, it may seem that the only logical and morally acceptable upshot of determinism is to eliminate the concept of punishment altogether, and replace the punitive institution with an institution that serves the interests of society and individuals in a non-punitive way. Eliminating punishment does not entail that we cannot protect society from dangerous individuals (cf. fn. 9). When necessary, we can justify detaining such individuals in the same way as we justify detaining carriers of highly contagious and dangerous viruses (Pereboom 2007:116). But eliminativists argue that we are not justified in making detainees suffer beyond what is the inevitable consequence of the means of protection. Moreover, detention, or other means for controlling risky behaviour, should be combined with measures for rehabilitation, which can be[14] a way of addressing the person rather than using him as a mere means for some extrinsic end (Pereboom 2007). Yet arguably, something paternalistic seems to be going on once punishment

[11] Here both Scanlon and Rawls differ from a classical contractarian like Hobbes. To Hobbes, the value above all values is security, whereas to contemporary contractualists, e.g. justice, freedom and autonomy are rated higher, and thus different justifications are yielded.

[12] Confer Rawls (1999a).

[13] Further, as an anonymous referee has pointed out, an accusation of begging the question could be made against baking into the contract the values that are supposed to be justified by the contract. But it is not question-begging if the assumption about people's reluctance to being treated as means is a psychological rather than a moral one. Contractors choose moral principles on the basis of some commonalities in human psychology. What these commonalities are is of course ultimately an empirical question which cannot be settled here, but again, that it matters to most of us to be treated as somehow "end's in ourselves" is not an implausible assumption.

[14] I say "can be" because arguably, there are forms of rehabilitation that may entail using people as mere means. Cf. Stanley Kubrick's film "A Clockwork Orange".

is replaced with mere rehabilitation, even if rehabilitation is in the best interest of the person. In addition to bereaving persons of the opportunity to truly repent for their past crimes, the reformatory approach conflates classes of offenders who are normally taken to differ in terms of their capacity for exercising moral agency; i.e. it makes no principled distinction between competent adults and the mentally ill, or between adults and children (Bomann-Larsen 2010). In that sense it appears to fall short of paying due respect to persons who take an interest in being treated as competent moral agents, by treating them on a par with non-competent agents. From a contractualist point of view which assumes that most people take an interest in being treated as moral agents, the alternative seems to leave little to be desired.

The contractualist argument states that it is unfair to treat persons contrary to their interest in being treated as morally responsible agents; i.e. to treat them as objects to be manipulated, as children, as anything less than competent adult moral agents. Purely forward-looking theories of punishment fail to take into account the fundamental interests of persons to be treated as if their actions are really up to them. In Antony Duff's words:

> To use censure simply as a useful technique for modifying conduct is to treat the person censured as a not responsible and autonomous subject, but as an object to be manipulated by whatever techniques we can find (Duff 2001:28).

Only some form of backward-looking, that is, retributive, theory seems to sufficiently accommodate the claim of the offender to be treated as a moral agent. What I have referred to as strong (or "positive") retributivism is the view that it is a moral duty to punish the guilty, simply because they are guilty and deserve to be punished. However, it is hard to see how this justification of punishment can be reconciled with the contractualist argument which, after all, takes as given that our moral agency is ultimately a matter of luck. On the strong retributivist account, justice *demands* the punishment of the guilty; hence guilt is a *sufficient* condition for punishment (Duff 2001:19). But this thought is challenged by the rejection of "up to usness" because even though it is in the interest of persons to be treated as moral agents in the desert-sense (say, to be blamed), it does not follow that it is in their interest to be *punished*. Hence, on this scheme guilt is not *sufficient* for punishment as imposed suffering; we need to fulfil an additional condition of *acceptability* of the principles of punishment.

Communicative punishment seems to preserve those elements of retributivism that are desirable to preserve. It retains the idea that punishment is a form of public censure which ties the agent to her past action, but it need not presuppose anything about ultimate "up to usness". It suffices to presuppose the importance to us to be treated as the agents we conceive of ourselves to be, and our acceptance of such treatment in terms of that importance.

9.4 Finding a Reasonable Standard for Determining the Mode and Scope of Punishment as Communication

My first claim was that the contractualist argument for moral responsibility establishes a basis for a publicly instituted censure of wrongdoing, aiming at addressing the offender as a moral agent and offering him an opportunity to repent past actions

and reconcile himself with the moral community. Granted that this claim is accepted by the reader as plausible, it is time to turn to the second and third claims, namely that (2) censure must sometimes be expressed in terms of hard treatment, but that (3) severe/extreme treatment is unwarranted.

To the second claim, it can be objected that the contractualist argument merely warrants symbolic punishment, in the form of public moral criticism and demand for public apology, and that this should suffice for reconciling the offender, the victim and the community. This is a rather strong revisionist view, which implies a radical transformation of the expressions of our reactive attitudes to wrongdoing. Still, the view is not implausible. Arguably, what choosers decide on behind a veil of ignorance determines the proper terms of communication. Hence, if the parties agree to a revisionary scheme where repentance is expressed in merely symbolic terms, it can not be objected that symbolic punishment fails to appropriately express condemnation.[15] There are several ways of responding to this objection, however. It is true that whatever the parties behind the veil agree on constitutes the proper terms of communication. But it seems the parties cannot set the terms of appropriate communication at whim. As soon as a backward-looking principle of punishment is accepted, that is, a principle according to which punishment aims to address the perpetrator as a moral agent by communicating society's views on the blameworthiness of her conduct, we are also committed to a proportionality criterion; punitive sanctions must be arrayed according to the degree of blameworthiness or seriousness of the conduct (von Hirsch 1994:125). Proportionality is related to fairness and thus something owed to offenders and victims alike. To punish a minor crime in the same way, or harsher, than a major crime, would be unreasonable and unfair to the offender, but also to the victim of the major crime. Now, the proportionality requirement does not necessarily entail that the treatment must be hard, but as I will argue, there are good proportionality-related reasons for hard treatment. I will return to this shortly.

To the third claim, it can be objected – in an opposite sentiment – that if we take the contractualist argument seriously as establishing full moral responsibility even in a deterministic world, there is no limit at the outset to choosing any degree of punishment that we see fit. Thus either (a) no revision of our penal practices is really needed, and in consequence, causal determinism is bracketed out as irrelevant to theories of punishment; or (b) the contractualist justification for punishment may in fact yield draconian punishment for rather minor offences, if choosers see it fit. I believe this last objection too can be addressed with reference to the proportionality criterion, which is essential to all forms of backward-looking punishment where the sanctions aim to reflect the severity of the wrongdoing. But before returning to the discussion of proportionality, I will elaborate somewhat on the objection (a) that determinism is irrelevant to our theory of punishment and hence that if the contractualist argument succeeds in establishing desert and warranting censure-based punishment, then no revision of our penal practices is needed.

I will argue that in searching for a reasonable standard for determining the acceptable mode and scope of punishment, we must take seriously *both* the fact that our

[15] I owe this point to Kasper Lippert-Rasmussen.

actions are not really up to us *and* the fact that it is in our interest to be treated as if our actions are up to us, in a way that appropriately addresses wrongdoing as something for which the agent is responsible. In other words: (1) we must be able to justify the terms of punishment to those who will (as a matter of causal luck) suffer from it; that is, to potential offenders, and (2) if we are to express censure properly, the censure must communicate society's view on the gravity of the wrong, both in absolute terms and relative to other wrongs; i.e. the punishment must "fit the crime". To let a wrong go by without appropriately addressing it signals that we do not really regard it as wrong (Duff 2001:28). And yet excessive punishment from the point of view of the offender could be rejected by a chooser in the contract situation, who does not know whether he will (as a matter of causal luck) end up as an offender or not. Contractualist communicative punishment thus consists in reaching equilibrium between "too little" and "too much" according to the above-mentioned constraints, and requires that the interests of the perpetrator and the interests of the moral community are balanced.

I have suggested that we employ the metaphor of the veil of ignorance to determine the appropriate measure of communicative contractualist punishment. What is of the essence in this context is that the parties "do not know how they have fared in the natural lottery" (Rawls 1999b:113) – i.e. "that they do not know their own place in the distribution of natural talents and abilities" (Rawls 1999c:178) – including their moral talents and abilities. It is also fundamental that the parties recognise one another as someone who has legitimate claims; that is, as someone to whom justification is owed.

Not knowing how one has fared in the natural lottery implies that one does not know whether one will be lucky or unlucky in terms of the factors that facilitate either moral or criminal behaviour. This reflects the underlying intuition of the fairness objection against retributive punishment in a deterministic world, and of the contract-based response to the objection. The intuition is the same as that driving luck egalitarianism. People are not to be made to carry the burden of (bad) luck, only the burden of their choices.[16] But under determinism, causal luck encompasses

[16] Given this idea, we could arguably be compelled to choose the principles of criminal justice on the same basis as we choose our principles for distributive justice, which entails choosing a scheme for criminal justice which distributes burdens according to the principles that would be most beneficial to those who are the worse off in virtue of bad causal luck. On this interpretation and in the present context, the worse off are those who end up as criminals. However, there is an important difference between criminal justice and distributive justice which does not warrant treating them on the same scale, even if we accept the intuition underlying luck egalitarianism as generally sound. Criminal justice is about *creating* (new) burdens, not distributing already existing burdens. The burdens created by criminal justice do not exist prior to the institutions which create them, and they need not be created at all. So from the idea that when we distribute existing burdens people should only be made to carry the burdens which result from their voluntary choices, nothing follows with regard to how we should distribute the burdens we create; or to whether we should create them at all. The rationale for criminal justice is thus different from the rationale for distributive justice. (I owe this point to Jakob Elster). I hold on to this thought even if it could be argued, as pointed out by an anonymous referee, that criminals do create burdens (directly and indirectly) and therefore, punishing them is a way of redistributing the burdens they have created. I read this

everything, including our choices. The contractual approach responds that it is still in our interest to maintain a distinction between chance and choice which corresponds with the distinction between factors over which we exercise control and factors over which we do not, even if we never exercise *ultimate* control.

Yet, in taking the fact of causal luck seriously, I have to acknowledge that my moral persona is contingent on luck factors and therefore that I should not judge, at the outset, that criminals are some mysterious "others" who deserve to suffer simply for being who they are. We easily grasp the thought that "it could have been me" when witnessing someone being the *victim* of crime. We sympathise with the victim precisely because we can imagine ourselves being in her shoes, and we demand the punishment of the perpetrator as if he had wronged ourselves. It is much harder to imagine oneself as a different moral persona, say, as a murderer, or a rapist. But given that we are not to know whether we will be the lucky or unlucky ones, we are obliged to place ourselves in the perpetrator's shoes as well, and acknowledge, with Scanlon, that "There but for the Grace of God go I" (Scanlon 2000:294). And since one of the things we must take into account when we decide on the severity of punishment is that we might end up as receivers of the punishment, we have reason in the contract situation to not allow extremely severe punishments. Thus our current punitive practices are not unaffected by determinism, and some revision of the mode and scope of punishment is called for.

However, as suggested above, there seems to be something unsatisfactory to merely saying: "I am sorry (for killing your child)" – in particular when the killing has been deliberate. Mere apology does not seem to bring us beyond agent-regret; beyond the: "How much better if it had been otherwise". In order to properly own up to what one has done, then, one must accept the imposition of some form of suffering beyond the mere unpleasantness of public apology. Similarly, society must show how seriously it takes the crime by imposing a fitting amount of suffering. "Censure cannot be expressed adequately in purely verbal or symbolic terms; the hard treatment is necessary to show that the disapprobation is meant seriously" (von Hirsch 1994:121).

To let a wrong go by insufficiently noticed is to silently endorse it, that is, to express that it does not really matter; that it is not important. Take rape as an example. Say we know that the rape is a one-time event. There is no possibility that the offender will ever do it again. Add that no one but the involved parties learns about the rape, so punishing the rapist has no deterrent effects on other potential rapists. In this case there are no forward-looking reasons either for punishing or detaining the rapist. Yet merely to ask the offender to apologise to his victim communicates

objection as a version of the "Removal of Unfair Advantage" justification of punishment. While the theory bears with it some interesting challenges, it has been vastly criticised – in my view correctly – for giving a distorted picture of what makes crimes wrongs. "The criminal wrongfulness of rape, for instance, in virtue of which it merits punishment, does not consist in taking an unfair advantage over all those who obey the law" (Duff 2001:23). True, there may be types of crimes where unfair advantage is an appropriate explanation of their wrongness, but these will typically not be the kinds that are *mala in se*, with which I am mainly concerned here. In any event, creating a new burden (punishment) does not nullify the original burden (the crime), it adds another.

a lot: it communicates to the offender that he did not do anything really wrong.[17] It communicates to his victim that she was not really wronged, and, hence, that rape is not really a wrong.

Von Hirsch points out that "if censure conveys blame, its amount should reflect the blameworthiness of the conduct . . . the severity of a sanction expresses the stringency of the blame . . . Hence punitive sanctions should be arrayed" (ibid.:125). That is, punitive sanctions must be constrained by *ordinal* proportionality. (ibid.:128) Ordinal proportionality compares different crimes in terms of their relative seriousness, and requires that differentiation of punishment should display the relative seriousness of crimes.

Now, symbolic punishments may meet the criterion of ordinal proportionality, that is, there may be different symbolic expressions for different offences. Just like awarded medals which have the same material composition may signify different merits, so too may symbolic punishments signify different types and degrees of wrongdoing.[18] Consider for instance the "Hester Prynne"-sanction; Hester Prynne had an "A" for "adulteress" painted on her forehead. We could imagine a system where criminals were branded, say, "R" for rapist or "M" for murderer. Such branding would amount to differentiated symbolic punishment because the degree of social shaming would reflect the degree of wrongness. However, it seems that for symbolic punishment to work as *punishment*, it would have to be pretty harsh – as in the Hester Prynne-case – in the sense of being truly socially shaming.[19] Otherwise it would fail to communicate the wrongness of the action. Hence even if ordinal proportionality is compatible with symbolic punishment because symbolic punishment may be arrayed, the material content of the punishment (here, in terms of the consequential social stigma and resultant isolation, mistrust and fear which would appear to be its corollaries) seems to be just as harsh, if not even harsher than imprisonment, and *ipso facto* amount to a form of hard treatment.[20]

Let us now turn to the objection to my third claim above that the contractualist argument will not warrant punishment in the form of *severe* or *extreme* suffering. Consider the argument for draconian punishment for drunken driving. Thomas Pogge has argued that a contractualist approach to punishment may in fact yield draconian punishment even for strict liability offences (Halvorsen 2002:fn. 16, 180–81). Drunk driving illustrates this well: Why should we not accept punishment by death for drunken driving, if, say, by executing 10 people we could prevent 1000 deaths caused by drunken drivers?

---

[17] Of course, he will have violated a legal rule, which provides a reason to punish him, but we do not primarily want to communicate that he has violated a regulation; we want to communicate that he has violated his victim. The legal prohibition against rape is after all there because rape is wrong, and it is wrong independent of the law.

[18] I owe this point to an anonymous referee.

[19] If the norms were completely internalized, it could of course be the case that symbolic punishment worked without social shaming. But in one important sense, this would not be punishment at all, but rather some form of self-reproach.

[20] That said, this form of punishment is not ruled out by anything I say here. All I am saying is that hard treatment is a necessary component for punishment to be justified, not that the form of punishment needs to be imprisonment.

Perhaps we would, if the risk of being killed by drunken drivers was severely reduced by the deterrent effect. It is questionable whether it would deter, however; drunken driving is a strict liability crime because the *mens rea* component is absent. People who drive drunk do so because they are drunk and their judgment is impaired, hence it is questionable whether a threat of capital punishment would have a deterrent effect. Of course, it might deter people from drinking, but that is not in itself crime, or it might create an incentive to lock down the car keys, which would be a good thing. But there are other concerns as well. For one, the kind of contractualism I am expounding here is constrained by fairness considerations. Punishing drunken driving *killings* with death leaves a lot of space for moral luck; those who end up as killers are morally unlucky, and it seems unfair to punish someone on the basis of moral luck.[21] Hence, if we accept capital punishment for drunken driving killings, it seems we should also accept if for drunken driving which, accidentally, does not result in someone's death. And even if we do accept capital punishment for killings caused by drunken driving, it is unlikely that we would accept capital punishment for drunken driving which does not result in someone's death. Given global causal luck, the risk of ending up a drunken driver is much greater than the risk of ending up the victim of a drunken driver. After all, most drunken drivers never kill anyone. Knowing this it is not evident that it would be rational to choose the death penalty for drunken driving.

Secondly, I am presupposing that people take an interest in being treated as moral agents – that is the basis for the contractual justification of punishment. This interest can only be served by differentiating between punishments according to degrees of blameworthiness. Drunken driving might result from horrible recklessness, but it is not as bad as murder. Consider capital punishment for drunken driving from the point of view of proportionality. It can of course be argued that since drunk drivers willingly accept the risk of killing others, punishing them by death is not entirely unfitting. And still, drunk driving is not murder, even when it results in unjustified killing. Taking moral agency seriously, we cannot conflate risky behaviour which might cause death – even when in fact it does cause death – with intentionally and deliberately killing another human being. And in terms of ordinal proportionality, the most severe punishments should be spared for the most severe wrongdoings; the ultimate punishment must be spared for the ultimate crime. Hence, if we assign punishment by death for drunken driving, we have exhausted our opportunity to mete out the most severe punishment for the most severe crimes; and arguably, murder is a greater wrong than killing someone by drunken driving.[22]

[21] I am assuming that it matters to people to be regarded as agents, even if they are not to take into account what kinds of agents they are, and thus that the principles guiding society should reflect this interest in being treated as agents. This calls for a distinction between mens rea crimes and strict liability crimes, where the latter does not involve any agency control but more, and external, luck.

[22] It could be objected that if the punishment for drunk driving was death, the punishment for murder could be torture + death (I owe this point to Kasper Lippert-Rasmussen). But this seems to violate the proportionality principle as well. i.e. "A head for an eye" seems disproportionate in *cardinal* terms, as discussed below.

However, ordinal proportionality considerations only tell us to array punishments relative to the blameworthiness of different kinds of conduct. It cannot give us an answer to where to set the upper and lower limits for punishment, for instance, what kind of punishment should be given for the most severe crimes. Ordinal proportionality must therefore be paired with *cardinal* proportionality. Cardinal proportionality is non-relative (von Hirsch 1994:129) and sets the limit on justifiable intrusions on the liberty of persons. In addition to the considerations of ordinal proportionality in relation to strict liability offences versus *mens rea* offences above, cardinal proportionality too explains why the former should not be subject to draconian punishments. To live in a society where unintended harm is draconianly punished would be unbearable and intolerably intrusive on people's liberty. Thus, while ordinal proportionality in principle could allow for rather harsh punishments for minor crimes, provided only that the punishments given for graver crimes are even harsher, cardinal proportionality prevents us from harshly punishing lesser wrongs. We could imagine a society which punished throwing gum on the sidewalk with imprisonment.[23] But this seems too intrusive of liberty to be acceptable – not least because humans are fallible, and a society which was not willing to cut people some slack would be extremely difficult to live in.[24] Moreover, we know that in so far as people are allowed to drink, and to own cars, some of us will occasionally get drunk and lose self-control, and jeopardise people's lives by getting behind the wheel. Society accepts this risk, while it could be avoided by banning alcohol or cars. But we take prohibiting either to be overly intrusive and illiberal, and if the prohibition is incompatible with liberal principles then so, it seems, is punishing its accidental consequences by severe means.

Nothing said so far rules out capital punishment for capital crimes, and life-time sentences of imprisonment for major crimes short of murder. Yet I have suggested that the contractualist argument will not allow for extreme punishment. But why is that? Or put differently: where do we set the *maximal* limit on punishment; i.e. the standard in relation to which punitive sanctions are to be arrayed?

Given the contractualist argument, the justifiability of the punishment itself depends on its acceptability to those who are to be punished. What we need to ask, then, is which modes and scopes of punishment it is reasonable to expect those who will suffer the punishment to accept; and we need to ask this in ignorance of whether we will be in the perpetrator's shoes.[25]

---

[23] There are of course many countries that do practice severe punishments for minor crimes; e.g. Singapore, where people suffer imprisonment for littering the streets. I owe this point to an anonymous referee.

[24] I owe this point to Jakob Elster.

[25] It can be objected, as an anonymous referee has, that if contractors who behind the veil do accept a crime-preventing system of hard punishment, it is reasonable to expect them to accept such punishment if they turn out to be offenders. But again, there are certain constraints on the contract: not every conceivable agreement is reasonable. Acceptability is not entailed by agreement – it is entailed by *reasonable* agreement. There are some independent standards determining what the contractors may agree on. Unreasonable agreement may display a failure to take into account the

First, I believe that from that point of view it would not be regarded as acceptable to impose a scheme of punishment that required one to pay with one's life, either in terms of a life-time imprisonment or in terms of capital punishment. Second, even if imprisonment is acceptable as a mode of punishment, extreme conditions in prison such as isolation, brutal treatment and similar hardships are hardly likely to be acceptable. Third, long sentences short of life may be unacceptable. Say a woman becomes a murderer at 20. A sentence of a length that would make it impossible for her to ever have children could be regarded as imposing an unacceptable burden. Basically, while it is plausible that most offenders take an interest in being treated as moral agents and therefore to accept the imposition of some form of suffering as a "vehicle for the expression of condemnation" (von Hirsch 1994:125), it is equally plausible to assume that it is in their interest to be given an opportunity to reconcile themselves with the community. Punitive principles that impose unbearable sacrifices on the part of offenders by closing off all possibility of reconciliation, or by inflicting irreparable harm, may be rejected as unacceptable on the terms that they are not likely to be accepted by those suffering under the principles.

I have argued that in the contract situation we would accept proportionality-constrained backward-looking hard punishment – but not extremely severe punishment – as a means of communicating moral censure. Let us now consider two final objections to this contractualist communicative theory of punishment. First, the hardship of punishment, and hence the message it conveys, is relative; a life-time sentence could arguably be perceived of as a less harsh message to a 90-year old than, say, ten years imprisonment to a 17-year old. Not to mention the message that being sentenced to death conveys to the religious fanatic terrorist aiming to become a martyr. I have argued that the punishment should aim to communicate the degree of blameworthiness of the act, i.e. that the punishment fit the crime. This implies that if two crimes are of identical seriousness, the punishment should be identical too. But it seems identical punishment may communicate different messages to different offenders.[26]

Second, it seems the contract-based argument cannot justify punishing those who for some reason or other do not take an interest in being treated as moral agents at all, either because they lack the capacity to respond to moral reasons, or because they do not want to be members of the moral community.

---

point of view of those subjected to the rules. What I am suggesting is that if the people behind the veil truly take into account that it is ultimately just a matter of luck who ends up as criminals, they might consider that if they were in those unlucky shoes, they would want to be given a second chance. Combined with an interest in being treated as agents in relation to those actions over which they do exercise some control (and not in relation to what merely happens, i.e. outcomes of moral luck), this starting point sustains the distinction between *mens rea* crimes and strict liability offences, while at the same time it should also yield a more forgiving attitude towards the former. Of course, if the contractors are Hobbesian in the sense that their only concern is security, then the picture may look very different, but Hobbesian rationality arguably does not give a plausible account of human psychology.

[26] I owe this point to Kasper Lippert-Rasmussen.

The former objection does not pose a problem that is peculiar to communicative punishment, nor to a contractualist one. Insofar as it poses a problem at all,[27] it does so to *all* theories of punishment, determinism aside. Deterrence-based theories must deal with the problem that a threatened punishment may deter some but fail to deter others, and while there is in principle no reason why deterrence-based theories should not differentiate punishment according to what effectively deters different individuals, doing so would be practically impossible and raise all sorts of epistemic problems. Some 90-year olds may value every day they have left of their lives, some 17-year olds may not care about their lives at all. But the law must strike a uniform compromise on the basis of the presumed efficiency of the threat for all subjects. Similarly, all desert-based theories face the same problem. According to retributivism, the guilty deserve to suffer, but the extent to which the guilty actually *do* suffer will vary widely across different individuals; yet again, the law must be uniform and address all subjects as one.

The latter objection may be warded off in similar way. Any theory of punishment must deal with those who are not to be regarded as full members of the moral community, but who pose a danger to others. While they may be unable to repent, society still needs to protect other citizens from them. "Atypical perpetrators", like psychopaths, pose a problem to any theory of punishment. Only eliminativist theories may avoid this problem, because they do not treat anyone as full moral agents, and reject the very notion of punishment. Hence they need not distinguish between punishable and non-punishable persons. According to this view, we are all in one sense "atypical perpetrators" insofar as we are perpetrators. Yet as suggested in Sect. 9.2, eliminativist theories must carry the justificatory burden of showing why normal, morally competent adults ought to be treated in the same way as those who are usually regarded as incompetent.

Purely consequentialist theories may on their part distinguish between those who are punishable, i.e. able to be deterred by the law, and those who are not (Greene and Cohen 2004). But they, too, need to deal with the latter group, though the former group – the punishable – may include more individuals than it would under a desert-based jurisprudence (i.e. many psychopaths are able to be deterred by the law, but they do not understand the concept of moral wrong and can thus not be *deserving* of punishment). There will then, on any theory of punishment, always be a class of offenders which the theory cannot account for, and which still needs to be dealt with. That is a problem for liberal society, but not one which is peculiar to the theory proposed here. The contractualist communicative theory I have expounded is only idiosyncratic on one score: namely that not only those who fall outside of the moral community for lack of capacity, but also those who choose not to be part of it (because they do not value being treated as moral agents), are unjustifiably punished. That is a bullet the theory should be willing to bite, just as these people must be willing to bite the bullet of being treated as incapable, and be detained should they pose a risk to others, as the upshot of their voluntary alienation.

[27] I say "if" because there is still considerable leeway for judges and juries to exercise judgment in individual cases.

9.5 Conclusion

I have argued that even if determinism is true and our actions are not ultimately up to us, we can still justify desert-based moral practices with reference to our interest in being treated as full moral agents. Such a contractual notion of desert gives us the basis for a backward-looking form of punishment as a vehicle for publicly communicating censure of wrongful actions. This contractual communicative theory of punishment is revisionary because it takes seriously the fact that our agency is owed to causal luck, hence the principles of punishment must be justifiable also to those who end up at the receiving end of punishment. It is plausible that parties to a contract, who do not know how they will fare in terms of luck, will accept that some form of hard treatment is required in order to appropriately communicate censure, but that they have reason to reject extreme punishment, that is, punishment which either makes reconciliation with community impossible, or which imposes irreparable damage.

Acknowledgments I want in particular to thank Jakob Elster, Vidar Halvorsen, Kasper Lippert-Rasmussen and an anonymous referee for helpful discussion and useful criticisms during the process of writing this paper. Thanks also to participants at the conference "Moral Responsibility: Neuroscience, Organization & Engineering" at TU Delft, August 24–27 2009, and to Nicole Vincent for opportunity and support.

References

Bomann-Larsen, Lene. 2010. "Revisionism and Desert." *Criminal Law and Philosophy* 4:1–16. doi: 10.1007/s11572-009-9081-x.

Dennett, Daniel. 2003. *Freedom Evolves*. New York, NY: Viking Press.

Duff, R. Antony. 2001. *Punishment, Communication and Community*. New York, NY: Oxford University Press.

Duff, R. Antony. 2007. *Answering for Crime. Responsibility and Liability in the Criminal Law*. Legal Theory Today. Portland, OR: Hart Pub.

Feinberg, Joel. 1970. *Doing and Deserving*. Princeton, NJ: Princeton University Press.

Greene, Joshua, and Jonathan Cohen. 2004. "For the Law, Neuroscience Changes Nothing and Everything." *Philosophical Transactions of the Royal Society of London* 359:1775–785.

Halvorsen, Vidar. 2002. *Ethics, Force and Violence in Policing*. Oslo: Unipub forlag

Lenman, James. 2006. "Compatibilism and Contractualism: The Possibility of Moral Responsibility." *Ethics* 117:7–31.

Levy, Neil. 2007. *Neuroethics. Challenges for the 21st Century*. Cambridge: Cambridge University Press.

Morse, Stephen. 2004. "New Neuroscience, Old Problems." In *Neuroscience and the Law. Brain, Mind and the Scales of Justice*, edited by Brent Garland, 157–98. Washington, DC: Dana Press.

Pereboom, Derk. 2007. "Hard Incompatibilism." In *Four Views on Free Will*, edited by John M. Fischer, Robert Kane, Derk Pereboom, and Manuel Vargas, 85–125. Singapore: Blackwell.

Rawls, John. 1999a. "Two Concepts of Rules." In *John Rawls. Collected Papers*, edited by Samuel Freeman, 20–26. Cambridge, MA: Harvard University Press.

Rawls, John. 1999b. "The Sense of Justice." In *John Rawls. Collected Papers*, edited by Samuel Freeman, 96–116. Cambridge, MA: Harvard University Press.

Rawls, John. 1999c. "The Justification of Civil Disobedience." In *John Rawls. Collected Papers*, edited by Samuel Freeman, 176–89. Cambridge, MA: Harvard University Press.

Scanlon, Thomas M. 2000. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Smart, John J.C. 2003. “Free Will, Praise and Blame.” In *Free Will* (2nd Edition), edited by Gary Watson, 58–71. New York, NY: Oxford University Press.
Smilansky, Saul. 2000. *Free Will and Illusion*. New York, NY: Oxford University Press.
Strawson, Galen. 1986. *Freedom and Belief*. Oxford: Clarendon Press.
Strawson, Galen. 2003. “The Impossibility of Moral Responsibility.” In *Free Will* (2nd Edition), edited by Gary Watson, 212–28. New York, NY: Oxford University Press.
Strawson, Peter. 2003. “Freedom and Resentment.” In *Free Will* (2nd Edition), edited by Gary Watson, 72–93. New York, NY: Oxford University Press.
Vargas, Manuel. 2005. “The Revisionist’s Guide to Responsibility.” *Philosophical Studies* 125:399–429.
von Hirsch, Andrew. 1994. “Censure and Proportionality.” In *A Reader on Punishment,* edited by R. Antony Duff and David Garland, 112–32. New York, NY: Oxford University Press.
Wegner, Daniel M. 2002. *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.
Williams, Bernard. 1981. *Moral Luck*. Cambridge: Cambridge University Press.

Chapter 10
Moral Responsibility and Jointly Determined Consequences

Alexander Brown

Abstract In *Responsibility and Control: A Theory of Moral Responsibility*, John Fischer and Mark Ravizza argue against incompatibilist principles of moral responsibility and offer a compatibilist account of moral responsibility. The book has sparked much discussion and criticism. In this article I point out a significant flaw in Fischer and Ravizza's negative arguments against the incompatibilist Principle of the Transfer of Non-Responsibility. I also criticise their positive argument that moral responsibility for consequences depends on action-responsiveness. In the former case I argue that their putative counterexamples against Transfer NR and Transfer NR* are underdescribed but once fully described depend upon consequence-particulars and not consequence-universals as they claim. In the latter case I argue that their account is unable to cope with quite ordinary cases of jointly determined consequences.

10.1 Introduction

In *Responsibility and Control: A Theory of Moral Responsibility*, John Fischer and Mark Ravizza offer a compatibilist account of moral responsibility which purports to explain both how it is that "inner mechanisms" can lead to acts or omissions for which agents are morally responsible *and* how it is that "outer processes" can lead to consequences for which agents are morally responsible (Fischer and Ravizza 1998).[1] The book has sparked much discussion and criticism, particularly with respect to Fischer and Ravizza's claim that inner mechanisms (that is, mental states leading to acts or omissions) attract moral responsibility just in case they exhibit "guidance control" (see Glannon 1997; Stump 2000; Levy 2002; Ginet 2006; Judisch 2007; McKenna 2008). This account (they claim) is superior to accounts which ground moral responsibility in "second-order volitions" (cf. Frankfurt 1971, 1987). In this article, however, I wish to focus on the outer processes between agents' acts or omissions and the consequences in question. I accept that moral responsibility for consequences must depend on both inner mechanisms and outer

[1] The book brings together and refines material from a series of earlier articles.

A. Brown (✉)
School of Political, Social and International Studies, University of East Anglia, Norwich, UK
e-mail: alexander.c.brown@uea.ac.uk

N.A Vincent et al. (eds.), *Moral Responsibility*, Library of Ethics and Applied Philosophy 27, DOI 10.1007/978-94-007-1878-4_10,

processes, but I wish to bracket off internal mechanisms to focus on external processes.

Having introduced Fischer and Ravizza's negative argument that the inevitability of certain events or consequences does not rule out the possibility of an agent's moral responsibility for those consequences, and their positive argument that moral responsibility for consequences depends on action-responsiveness, I set out to do two main things. First, I attempt to highlight a flaw in Fischer and Ravizza's account of simultaneous overdetermination of consequence-universals. They present some examples in which two actual paths lead to the occurrence of the same consequence, where each path would have been sufficient to make the consequence obtain had the other path not existed. According to Fischer and Ravizza, such "joint" cases undermine "the Principle of the Transfer of Non-Responsibility" (Transfer NR). I shall argue that this is not the case. Since the consequence-universals in the putative counter-examples are underdescribed, the examples in question do not qualify as counter-examples to Transfer NR, but once the relevant consequences are properly fleshed out, they look more like consequence-particulars than consequence-universals.

Second, I investigate cases of jointly determined consequences, where each path is *not* sufficient by itself to bring about the relevant consequence. While examples of simultaneous overdetermination appear in philosophy books, examples of jointly determined consequences are a more familiar part of ordinary moral experience. Take the following description of the execution of a man on death row in Texas:

> The warden told Willingham that it was time. Willingham, refusing to assist the process, lay down; he was carried into a chamber eight feet wide and ten feet long. The walls were painted green, and in the center of the room, where an electric chair used to be, was a sheeted gurney. Several guards strapped Willingham down with leather belts, snapping buckles across his arms and legs and chest. A medical team then inserted intravenous tubes into his arms. Each official had a separate role in the process, so that no one person felt responsible for taking a life. (Grann 2009)

Fischer and Ravizza declare that they intend to establish a wide reflect equilibrium within the domain of phenomena associated with moral responsibility (Fischer and Ravizza 1998:34). If so, then it is important that their account of moral responsibility for consequences can explain our intuitive judgements about examples of jointly determined consequences. I shall argue that as it stands their account is not competent in this regard. If (as they propose) action-responsiveness is a necessary condition for an agent being morally responsible for certain consequences, then an act or omission that only jointly determines a consequence cannot attract moral responsibility given Fischer and Ravizza's definition of action-responsiveness. Yet it seems intuitive to say that such acts or omission can attract moral responsibility for the agents concerned. Therefore, either their definition of action-responsiveness is defective or action-responsiveness is not a necessary condition after all. At the end of this article I put forward my own dedicated accounts of joint moral responsibility for consequences and joint action-responsiveness for consequences which can deal appropriately with these examples.

10.2 Preemptive Overdetermination of Consequences

Typically, a compatibilist approach to moral responsibility will contain both negative and positive arguments: first, it will make the negative argument that it is false to suppose that the thesis of causal determinism is incompatible with attributions of moral responsibility (most notably by identifying counter-examples to incompatibilist principles); second, it will present a positive account of what moral responsibility is grounded in (such as an account of the internal mechanisms in virtue of which agents can be deemed morally responsible for their acts or omissions and an account of the external processes in virtue of which agents can be deemed morally responsible for the consequences of their acts or omissions).[2] I begin with the negative argument before turning to the positive argument.

Fischer and Ravizza reject the incompatibilist principle that a person cannot be held morally responsible for bringing about an event if that event was inevitable, where "inevitable" means that the agent could not have prevented the event from obtaining even if he had acted differently than he did. They do so by evoking Frankfurt-type cases.[3] In their example "Avalanche" we are asked to imagine that Betty is a double agent who has been instructed to explode dynamite at the top of a mountain in order to trigger an avalanche that will destroy an enemy camp in the valley below. However, unbeknownst to Betty, Ralph has been asked by his superiors (who are also Betty's superiors) to place explosives a few feet lower down the mountain than Betty's explosives, and to trigger the avalanche if Betty fails to do so. The superiors are unsure of Betty's commitment to the cause. Fischer and Ravizza argue that even though Ralph's presence is enough to ensure that the avalanche and camp-destruction take place, this does not detract from Betty's moral responsibility. Provided she has guidance control over her own actions, she *is* morally responsible for the consequences (Fischer and Ravizza 1998:155–56). Indeed, they argue that the same could be said even if the story is modified so that if Betty had not placed the charge, then a natural event (erosion) would have triggered the avalanche (157).

The nub of the example is to show that mere inevitability of consequences does not rule out the possibility that an agent could be morally responsible for those consequences. Fischer and Ravizza believe that other Frankfurt-type cases illustrate

[2] I do not mean to say that a compatibilist must attempt to do both of these things, only that a complete compatibilist approach would. Nor do I mean to suggest that the business of developing a positive account of moral responsibility is idiosyncratic to compatibilists. For, it is available to an incompatibilist to combine such an account (even the positive account of Fischer and Ravizza) with the requirement that the agent must have genuinely open alternative possibilities (see Ginet 2006:241).

[3] The original Frankfurt case runs as follows. Suppose Black places Jones in a set of circumstances which leave Jones no alternative but to do the thing that Black wants him to do. If Jones begins to look as though he is not going to do as Black desires, then Black will take "effective steps" to ensure that he does, although Jones does not know that Black will intervene. In the end Jones decides to do the thing without Black actually having to step in to make him do it. Although Jones could not have acted otherwise in that sense, we nonetheless tend to think that he is morally responsible for his actions (Frankfurt 1969:835).

the same point (155). Take the example of "Assassin". Imagine that Sam tells his friend Jack of his plan to assassinate the mayor. Jack has his own reasons for wanting to assassinate the mayor and is nervous. He wants to make doubly sure that the mayor is shot and killing. So he hatches a plan to shoot and kill the mayor himself but only if Sam fails to do so for some reason. As it turns out, Sam does shoot and kill the mayor, and Jack (the assassin who ensures that that consequence will obtain no matter what) did not actually play a role in bringing about the shooting and death of the mayor. Since Sam *did* shoot and kill the mayor, it makes sense to say that Sam *is* morally responsible for this consequence, and this holds true in spite of the fact that the mayor being shot and killed was inevitable (given Jack's intentions and abilities).

Fischer and Ravizza refer to what happens in these sorts of cases as *preemptive overdetermination of consequences*, meaning that the ensuring conditions (as in, that Ralph or erosion would have triggered the avalanche and crushed the enemy camp if Betty had not done so, and that Jack would have shot and killed the mayor if Sam had not done so) are preempted by the actual conditions (that Betty did detonate the dynamite and crush the camp, and that Sam did shoot and kill the mayor). These examples demonstrate that, contrary to the incompatibilist principle that the inevitability of consequences rules out the possibility of moral responsibility, it *can be* appropriate to hold agents morally responsible for consequences even in scenarios where those consequences were somehow inevitable.[4]

There is, however, something about these cases that troubles Fischer and Ravizza; and rightly so I think. Let us say that an incompatibilist is someone who holds the following nuanced view about moral responsibility: if an agent is not morally responsible (or no agent is morally responsible) for certain events, and for the fact that those events causally determine other events or consequences, then that agent is not morally responsible (or no agent is morally responsible) for the other events or consequences. This is a view *only about* cases in which events for which an agent is not morally responsible (or no agent is morally responsible) causally determine other events or consequences about which the question of moral responsibility is asked. Yet the examples of Avalanche and Assassin are not cases of this sort. Here the relevent (counterfactual) events for which the agents in question (Betty and Sam) are not morally responsible (or no agent is morally responsible), namely, that Ralph (or erosion) would trigger an avalanche to destroy the camp if Betty does not and that Jack would shoot and Kill the mayor if Sam fails to do so, do *not* causally determine the other events or consequences about which the question of moral responsibility is being asked. The events or consequences in question are *preempted* by the actions of Betty and Sam. As such, the examples of Avalanche and Assassin are strictly speaking beside the point. They do not test the nuanced incompatibilist view. The right way to establish genuine counterexamples to the nuanced incompatibilist view is to present cases in which events for which the agents in question are not morally

[4] For alleged counterexamples in which the preempting agent is not morally responsible after all, see Levy (2002).

responsible (or no agent is morally responsible) *do* bring about consequences for which the agents in question are morally responsible (159–60).

10.3 Simultaneous Overdetermination of Consequences

Fischer and Ravizza take this objection seriously. Nevertheless, they believe that it is possible to construct other cases that are pertinent to the nuanced incompatibilist view. Specifically, they think that cases of *simultaneous overdetermination of consequences* are powerful counterexamples to the nuanced incompatibilist view. Such cases involve two-paths to an event or consequence – one path for which the agent is putatively morally responsible and another path for which the agent is not morally responsible (or no agent is morally responsible).[5] Before introducing these cases, however, I need first to say something about the intended target.

The intended target for cases of simultaneous overdetermination is the Principle of the Transfer of Non-Responsibility (Transfer NR), or two versions thereof:

> Transfer NR
>
> (1) If p obtains, and no one is even partly morally responsible for p; and
> (2) if p obtains, then q obtains, and no one is even partly morally responsible for the fact that if p obtains, then q obtains; then
> (3) q obtains, and no one is even partly morally responsible for q. (152)

and

> Transfer NR*
>
> (1) If S is not morally responsible for p; and
> (2) S is not morally responsible for the fact if p obtains, then q obtains; then
> (3) S is not morally responsible for q. (157)

Fischer and Ravizza claim that respectively Transfer NR and Transfer NR* are susceptible to the following counterexamples.

In "Joint Avalanche" we are asked to imagine that Betty sets off an explosive charge at T1 which triggers an avalanche that crushes an enemy camp at time T3. However, unbeknownst to Betty, whilst she is setting her explosives a nearby glacier is melting, shifting and eroding, and it releases another avalanche that crushes the enemy camp at the exact same time, T3. According to Fischer and Ravizza, Betty

[5] The path for which the agent is putatively responsible must, on Fischer and Ravizza's account, involve guidance control with respect to both inner mechanisms (from mind to act or omission) and outer processes (from act or omission to consequence). Since my interest in this article is with outer processes and not internal mechanisms, I shall not discuss Eleonore Stump's objection that if causal determinism is true and inner mechanisms are constituted by physical matter (i.e. neurons), then the path for which the agent is putatively responsible fails (Stump 2000). For a reply, see Fischer and Ravizza (2000).

is morally responsible for the consequence, *that the enemy camp is crushed at T3*, despite the fact that no one (including Betty) is even partly morally responsible for the fact that if erosion triggers an avalanche, then the enemy camp is crushed at T3. This is presented as a counterexample to Transfer NR, since here we have a case in which there is an event p for which no one is even partly morally responsible that causes another event or consequence q for which an agent *is* morally responsible (i.e. Betty) (162).

In "Joint Assassins" we are to suppose that Sam and Jack simultaneously shoot and kill the mayor. Unlike the case of Assassin, Jack does not wait to see what Sam will do. Rather, he decides to shoot and kill the mayor himself and does so at the exact same moment as Sam (time T2). Although the actions of Sam and Jack are independently sufficient to bring about the consequence, *that the mayor is shot and killed at T2*, in fact both Sam and Jack shoot and kill the mayor. Fischer and Ravizza present Joint Assassins as a counterexample to Transfer NR*. The suggestion is that although Sam is not even partly responsible for the event, *that Jack pulls his own trigger at T1*, nor for the state of affairs, *that if Jack pulls his own trigger at T1, then the mayor is shot and killed at T2*, Sam is morally responsible for the consequence, *that the mayor is shot and killed at T2* (161). (They also show sensitivity to the fact that a change to the timings of Joint Assassins will change the judgement. They imagine a case, "Joint Assassins 2", in which Jack actually fires his gun a few seconds after Sam, and Jack's bullet hits the mayor a few seconds after. In this case they judge that Sam is morally responsible for the consequence, *that the mayor is shot and killed at T2*, whereas Jack is only morally responsible for the consequence, *that the mayor is shot a second time* (118–19). They also imagine a case, "Joint Assassins 3" in which Sam fires first and then Jack second, but because Jack is much closer to the mayor, Jack's bullet hits the mayor first. Here they determine that Jack is morally responsible for the consequence, *that the mayor is shot and killed at T2*, whereas Sam is only morally responsible for the consequence, *that the mayor is shot a second time* (119).)

The alleged power of Joint Avalanche and Joint Assassins lies in the fact that they involve the simultaneous overdetermination of events or consequences. That is to say, they involve "the actual occurrence of two triggering events that both bring about the terminal events simultaneously" (118). This feature makes it possible for there to exist events for which some agent S is not morally responsible (or no agent is morally responsible) that causally determine a consequence for which S *is* morally responsible (or at least one agent *is* morally responsible).[6] This is not possible in cases of preemptive overdetermination because here it is only ever

[6] At first glance, it is not clear why it is important in the case of Assassins that the two agents are acting independently of each other. But perhaps the importance can be underlined as follows. Suppose a married couple are living beyond their means. They rack up large debts they cannot afford to pay off. They each know what the other is doing. Nevertheless, the actions of each agent would have been sufficient to produce debts which they cannot repay. This is a case of simultaneous overdetermination of consequence, but it is not a counterexample of the right sort. That is because it might be argued that the husband is partly responsible for the actions of the wife and vice versa.

the case that one event actually brings about the terminal event or consequence in question.

It is also important to recognise that Fischer and Ravizza's cases of simultaneous overdetermination (along with their cases of preemptive overdetermination) are presented as involving "consequence-universals". A consequence-universal "can be brought about via different causal antecedents" (96). In contrast to this, "a consequence-particular is individuated more finely" (ibid.). In other words, "the actual causal pathway to a consequence particular is an essential feature of it" (ibid.). The more basic distinction here is between what is repeatable or multi-instantiable (universals) and what is non-repeatable or mono-instantiable (particulars). The property black, for example, is a universal because it can be exemplified or instantiated in more than one place at the same time, as when I make the truthful statement that both my shoes and my trousers are black. Moreover, two different properties can be instantiated at the same place and time, as when I make the truthful statement that my shoes are black and shiny. Particulars, by contrast, are non-repeatable. Any particular pair of shoes cannot be in more than one place at the same time. And no two pairs of shoes can occupy exactly the same space at the same time. It is safe to say that if properties like *being black* are simple universals, then consequence-universals are complex universals in the sense that they combine particulars with universal properties. Hence we can say that the consequence-universal, *that the mayor is shot and killed at T2*, is complex because it attributes the universal property of being shot and killed to a particular person, the mayor. Consequence-particulars, by contrast, involve particular properties and lend greater precision to how consequences come about. Sticking with the same example, a consequence-particular might specify who shot and killed the mayor at T2.

Fischer and Ravizza stress that their counterexamples to Transfer NR and Transfer NR* "treat consequences as consequence-universals" (154 n. 5). In the case of Joint Assassins, they propose that Sam and Jack are each morally responsible for the consequence-universal, *that the mayor is shot and killed at T2* (161). The counterexample would not have worked if it had been couched in terms of consequence-particulars. Sam cannot be held morally responsible for the consequence-particular, *that Jack shot and killed the mayor at T2*, just as Jack cannot be held morally responsible for the consequence-particular, *that Sam shot and killed the mayor at T2*. Fischer and Ravizza also distinguish between "descriptive" and "modal" consequence-universals (102). The relevant descriptive consequence-universal in the case of Joint Assassins would be something like, *that the mayor is shot and killed by someone or other at T2*. The relevant modal consequence-universals would be, *if contrary to fact Sam had not shot and killed the mayor at T2, then the mayor would have been shot and killed by Jack alone*, and *if contrary to fact Jack had not shot and killed the mayor at T2, then the mayor would have been shot and killed by Sam alone*. Plainly the counterexample would not have worked if it

---

This looks likely if each is aware of what the other is doing, each does not attempt to stop the other and each encourages the spending of the other.

had been couched in terms of a modal consequence-universal. This is because Sam, say, cannot be morally responsible for the modal consequence-universal, *if contrary to fact Sam had not shot and killed the mayor at T2, then the mayor would have been shot and killed by Jack alone*.

The upshot is that if cases of simultaneous overdetermination of consequences are to attract moral responsibility in the right way, then it is only because one and the same consequence can be brought about in more than one way, and this is only possible if the relevant consequences are consequence-universals of the right sort. However, it is one thing for Fischer and Ravizza to acknowledge the fact that for their argument about moral responsibility to work they *must* appeal to consequence-universals rather than consequence-particulars; it is quite another to show that consequence-universals rather than consequence-particulars *do* figure in our ordinary attributions of moral responsibility; and yet another thing to demonstrate that the argument *can* work by appealing to consequence-universals rather than consequence-particulars. It is to these further necessary steps in the argument that I now turn.

10.4 Consequence-Universals and Consequence-Particulars

Do consequence-universals figure in ordinary attributions of moral responsibility? Some people might think that in ordinary situations we tend to hold agents morally responsible not for consequence-universals but for instantiations of consequence-universals (i.e. the actual obtaining of consequence-universals) *vis-à-vis* consequence-particulars. Of course, Fischer and Ravizza hold that agents can be morally responsible for both consequence-universals and consequence-particulars (121). In addition to their counterexamples against Transfer NR and Transfer NR* they have another set of arguments to show that agents can be morally responsible for consequence-particulars even in the absence of genuinely open alternative possibilities (98–101).[7] However, the present objection is that their argument that agents can be morally responsible for consequence-universals is misplaced since generally speaking we do not take agents to be morally responsible for consequence-universals just as we do not ordinarily take agents to be morally responsible for Platonic forms.

Even if some people might be tempted to think that consequence-particulars are the only things that matter as far as moral responsibility is concerned, I do not think that this is actually the case. On the contrary, consequence-universals are a feature of ordinary moral discourse and the attribution of moral responsibility. Suppose two brothers are playing football outside of their house. Just as the first brother is about to kick the ball, the second brother shouts out, "There is a spider crawling up your neck!" The first brother is distracted for a moment and accidentally kicks the ball toward the front window, smashing it. Although he kicked the ball into the window,

[7] For an interesting discussion of a possible tension between Fischer and Ravizza's account of guidance control over consequence-particulars and their account of taking responsibility, see Judisch (2007).

he says to the second brother, "It was your fault that the window got smashed because you distracted me." The other replies, "No, you were responsible for the fact that the window got smashed because you took your eye off the ball." Finally, the boys' mother appears on the scene and offers the final word, "Never mind who actually kicked the ball, you were both playing football outside of the house even though I have told you many times not to do so, therefore you are both jointly responsible for smashing the window." Presumably all of these utterances make sense, and they do so because the interlocutors are referring to the consequence-universal, *that the window is smashed by the ball*. (It also seems to me that the attribution of joint responsibility in this case is made plausible by the fact that the boys were engaged in a joint enterprise. I shall have more to say about joint enterprises below.)

Quite apart from the question of whether or not it is appropriate to attribute moral responsibility to agents for consequence-universals, I believe that a more pressing issue is whether or not consequence-universals can work in the way that Fischer and Ravizza intend them to work; whether or not Fischer and Ravizza are able to rely on consequence-universals in the cases of Joint Avalanche and Joint Assassins. I have serious doubts on this score.

Given their description of Joint Avalanche, it is clear that Fischer and Ravizza have in mind two simultaneously occurring events which are both counterfactually sufficient to crush the camp, meaning that if contrary to fact only one of the two avalanches had hit the camp, then it would have been sufficient to crush the camp by itself. Since the two avalanches did hit the camp at the same time and each avalanche would have been sufficient to crush the camp counterfactually, at first glance it seems plausible to say that both avalanches crushed the camp. But this idea is more difficult than it first appears. We are told that two events simultaneously overdetermined the same consequence, but this is metaphorical. It does not tell us *how it is that* the two events actually caused the same consequence at the same moment given the laws of nature in our world – the world for which the theory is intended. Without a proper physical description of the events and consequences in question it is mysterious how two events can simultaneously overdetermine the same consequence.

It deserves mention that avalanches are constituted out of physical things, including volumes of snow, lumps of rock, chunks of ice, tree roots and branches, man-made objects and dust particles. Hence Betty's avalanche B and the natural avalanche A are made up of constituents B/C1, B/C2 to B/Cn and A/C1, A/C2 to A/Cn respectively. The camp is also made up of a number of component parts P1, P2 to Pn and locations L1, L2 to Ln. With this in mind, let us imagine that the camp was being monitored by a network of slow motion cameras. Looking at the camera footage avalanche scientists discover that when the two avalanches hit the camp at time T3, it just so happened that all of the constituents from Betty's avalanche B, as in, B/C1, B/C2 to B/Cn, and only constituents from Betty's avalanche B, crushed all parts of the camp P1, P2 to Pn and occupied all locations L1, L2 to Ln. All of the constituents of the natural avalanche, A/C1, A/C2 to A/Cn, were pushed to the side at the point of impact. The avalanche scientists might account for this surprising set of facts with the hypothesis that Betty's avalanche B had greater velocity than

the natural avalanche A. Of course, had the constituents of Betty's avalanche B not crushed the camp and occupied its locations at T3 the constituents of avalanche A would have done so. But this is not what *actually* happened. Given this scientific discovery, the question is whether or not it still makes sense to say that both avalanches crushed the camp. The answer is surely that Betty's avalanche B and not the natural avalanche A crushed the camp. So here we have a case in which at least one person (Betty) *is* morally responsible for an event that causally determined (in part) another event or consequence for which someone (Betty) is morally responsible. So this is not a counterexample to Transfer NR.[8]

At this stage Fischer and Ravizza might insist that I have described a highly unusual turn of events. In most instances of simultaneous crushing of camps (so the counter runs) avalanche scientists expect that the constituents of both avalanches, B/C1, B/C2 to B/Cn and A/C1, A/C2 to A/Cn, will crush different parts of the camp and occupy its different locations. Not only that but snow, ice, rock, trees, objects and dust particles from the two avalanches will merge, meld and combine to form new, mixed constituents, M/C1, M/C2 to M/C3. If this is the case, then we can say that the two avalanches were jointly responsible for crushing the camp. What is more, we can say this even if each avalanche would have been sufficient to crush the camp without the presence of the other avalanche. So once again we do have a counter-example to Transfer NR, namely, a case in which there is an event for which no one is even partly morally responsible (erosion) that causally determined another event or consequence for which at least one person (Betty) *is* morally responsible. However, in order to make this claim Fischer and Ravizza have no choice but to rely on a consequence-particular rather than a consequence-universal. It would not be enough to say that there obtains a consequence-universal, *that the camp is crushed by an avalanche at T3*, which is caused (in part) by an event for which no one is even partly morally responsible. Why? Because that description is quite consistent with the story I told in which constituents from Betty's avalanche B crush all parts of the camp and occupy all locations. In order to rule out this possibility, Fischer and Ravizza must appeal to a consequence-particular, such as, *that a combination of avalanche constituents B/C1, B/C2 to B/C3, A/C1, A/C2 to A/Cn and M/C1, M/C2 to M/C3 crush parts of the camp P1, P2 to Pn at locations L1, L2 to Ln at T3*.

Much the same point applies to Joint Assassins. It is not clear that appealing to the consequence-universal, *that the mayor has been shot and killed at T2*, suffices to show that Sam and Jack are morally responsible. To see why, consider two additional

[8] I do not consider here the fact that there is an ambiguity in the concept of crushing a camp. On one reading, the camp is not crushed until every last part of the camp, P1, P2 to Pn, is crushed. This is "crushed" in the sense of complete obliteration. According to another reading, the camp is crushed only provided that all of the core parts, P*1, P*2 to P*n, are crushed, where the core parts might be the foundation blocks of the main living quarters, the supporting legs of the observation tower, the camp flag, the fence posts, the walls of the ammunition store and the communications dish. This is "crushed" in the sense of rendered useless as a camp but not completely obliterated. I believe that the example could be made to work whichever of the two readings of the concept of crushing a camp is given.

versions of Joint Assassins. In Joint Assassins 4, a forensic pathologist carries out a painstaking autopsy of the mayor's body. The autopsy along with CCTV footage, reveals that both Sam's bullet and Jack's bullet hit the mayor at the exact same moment, but Sam's bullet entered the mayor's heart, while Jack's bullet entered the mayor's stomach. Although either bullet would have been sufficient by itself to cause death, the forensic pathologist determines that it was the bullet entering the heart that caused death. Given this fact, we are inclined to say that Sam killed the mayor, not Jack. This would not count as a counterexample to Transfer NR*, since in this case there is no path for which an agent S is not responsible and that leads to a consequence for which S is responsible. Now consider Joint Assassins 5. Suppose the autopsy reveals that the two bullets entered and exited the mayor's heart at locations L1/L2 and L3/L4 respectively. The forensic pathologist determines that the mayor was killed by virtue of a massive trauma to the heart caused by two bullets entering and exiting the heart at locations L1/L2 and L3/L4 at T2. This implies the actual occurrence of two triggering events which brought about the terminal event simultaneously. And so the counter-example is back in play. However, we can *only* say this because we rely on a consequence-particular, namely, *that the mayor is shot and killed by virtue of a massive trauma to the heart caused by one bullet entering and exiting the heart at locations L1/L2 and a second bullet entering and exiting the heart at locations L3/L4 both at T2*.

The moral of the story is that Fischer and Ravizza's argument against Transfer NR and Transfer NR* does not go through unless the consequences of the putative counterexamples are properly fleshed out, but once the consequences are properly fleshed out we find that they are actually consequence-particulars rather than consequence-universals. Fischer and Ravizza are not unaware of this sort of counter-argument. They refer to it as "the Divide and Conquer strategy" (98). However, they seem to think that the strategy fails in respect of consequence-universals merely because it is possible for agents to exercise guidance control over consequence-universals (101, 106). Yet my objection is not that it is somehow impossible for an agent to exercise guidance control over a consequence-universal. Rather, my objection is that it is difficult to make sense of the claim that two acts can jointly overdetermine the same consequence given our laws of nature without knowing more about the consequences in question. This claim remains mysterious until the contents of the consequences have been fleshed out. But once the contents of the consequences have been fleshed out, they turn out to be consequence-particulars rather than consequence-universals.

Fischer and Ravizza affirm that their counterexamples to Transfer NR and Transfer NR* treat consequences as consequence-universals. If I am correct, then they are mistaken. But so what? Why does it matter if the consequences in question are consequence-particulars as opposed to consequence-universals? Apart from showing that Fischer and Ravizza are mistaken about the nature of their own counterexamples, it matters because telling the stories in terms of consequence-particulars changes our intuitive judgements. If the relevant consequence in the case of Joint Avalanche is the consequence-universal, *that the camp is crushed by an avalanche at T3*, then it opens up the possibility of saying that Betty is fully morally

responsible for what happened. But if the relevant consequence turns out to be the consequence-particular, *that the camp is crushed by a combination of avalanche constituents B/C1, B/C2 to B/C3, A/C1, A/C2 to A/Cn and M/C1, M/C2 to M/C3 at T3*, then Betty cannot be fully morally responsible for that consequence, since she is responsible for only some and not all of these avalanche constituents. Likewise, if the relevant consequence in the case of Joint Assassins is the consequence-universal, *that the mayor is shot and killed at T2*, then it seems that both Sam and Jack can be fully morally responsible. However, if the relevant consequence is the consequence-particular, *that the mayor is shot and killed by virtue of a massive trauma to the heart caused by one bullet entering and exiting the heart at locations L1/L2 and a second bullet entering and exiting the heart at locations L3/L4 both at T2*, then neither Sam nor Jack can be fully morally responsible for that consequence. As the case is described above, Sam is not responsible for both bullets, just as Jack is not responsible for both bullets.

Let us now turn from Fischer and Ravizza's negative argument against Transfer NR and Transfer NR* to consider their positive account of what it means for an agent to exercise guidance control over a consequence-universal.

10.5 Jointly Determined Consequences

On Fischer and Ravizza's positive account of moral responsibility, for an agent to be morally responsible for the consequences of his or her acts or omissions it is necessary that the inner mechanism leading to the relevant act or omission is "moderately reasons-responsive" (chap. 3), where the inner mechanism is the "agent's own" (chap. 7) and the agent "takes responsibility" (chap. 8) for it. Furthermore, it is necessary that the act or omission is part of a process of type-P which leads from the act or omission to the consequence in question and is "sensitive to action" or is action-responsive (107). In what follows I wish to concentrate on this notion of action-responsiveness.

According to Fischer and Ravizza, to say that an act or omission is part of a process of type-P that is action-responsive with respect to a consequence C is to say that the process in question involves a person S who moves his body in way B at time T and this is actually sufficient to cause C to obtain at $T+i$ given ordinary background conditions, *and* "If S were to move his body in way B* at *T*, all triggering events (apart from B*) which do *not actually* occur between *T* and $T+i$ or which *actually* occur and bring about C *simultaneously or subsequently* to $T+i$ were *not* to occur, and P-type process were to occur, then C would not occur" (120). So, for example, in the case of Joint Assassins we can say that Sam's act of pulling the trigger is action-responsive with respect to the consequence, *that the mayor is shot and killed at T2*, if and only if Sam's act is part of a process leading from the act to the consequence which is such that Sam's pulling the trigger is actually sufficient for the shooting and killing of the mayor at T2 given ordinary background conditions, and in the absence of other triggering events (i.e. another assassin) Sam's not pulling the trigger does not lead to the shooting and killing of the mayor at T2. Fischer and

Ravizza assume that the bullets are fired under ordinary atmospheric conditions and that such ordinary background conditions are part of the process that actually leads to the consequence (113).

Fischer and Ravizza believe that action-responsiveness is a necessary condition for attributing moral responsibility for consequences. To support this belief, they offer the example of "Missile 3". Imagine that an evil woman, Elizabeth, has launched a missile toward Washington, D.C. However, a second woman, Joan, has another weapon which she can fire at Elizabeth's missile so as to deflect it to a less populous area of the city. Due to her position, the missile's trajectory and the nature of her own weapon, Joan cannot actually prevent the incoming missile from hitting the city. But she can deflect Elizabeth's missile between neighbourhoods, say, from Columbia Heights to Glenmont. Fischer and Ravizza contend that because Joan is involved in a process which is *not* action-responsive with regard to the consequence-universal, *that Washing, D.C. is bombed*, she is not morally responsible for that consequence. This holds even if it *is* the case that Joan is responsible for the consequence-particular, *that Washing, D.C. is bombed at Columbia Heights rather than Glenmont* (95).

However, I shall now try to show that Fischer and Ravizza's account of action-responsiveness cannot capture our intuitions in an important range of cases. To be more specific, I will offer three examples that although do *not* contain a person S who moves his body in a way that is sufficient to cause C to obtain, and as such do not qualify as examples of action-responsiveness according to the proposed definition, do nevertheless seem to involve moral responsibility at an intuitive level.[9] All three examples involve persons whose acts *jointly determine* certain consequences.[10]

The first example, which I call Revolutionaries, involves *cumulative* shooting and killing. Imagine that two revolutionaries, working independently and without knowledge of each other's plans, both set out to kill a government official, each with a single bullet. However, because they are both farmers and not trained soldiers they are not very accurate. Consequently, they each shoot the mayor in parts of his body that other things remaining equal would not cause death. Yet the cumulative effect of both bullets is enough to cause death. In this example I think it makes sense to say that each revolutionary is *jointly* responsible for shooting and killing the government official even if the actions of each assassin, taken in isolation, did not cause death.

[9] I shall not discuss here an example put forward by Glannon (1997) in which an agent is in a situation such that even though her act was not, and could not have been, sufficient to bring about or prevent the occurrence of an undesirable consequence directly, she is nevertheless morally responsible for the consequence through her own negligent interference in the actions of another person who is in a position to determine what happens directly.

[10] I do not intend to consider cases of analytic joint responsibility; that is to say, cases in which it is true by virtue of the meanings of the terms that two agents are responsible for the consequence in question since the consequence must involve joint action. For example, we might say that two people must be jointly responsible for the consequence, *that the boss had an affair with his secretary*, since given the ordinary meaning of "an affair" the consequence must be the result of the joint actions of two agents. This is captured by the popular dictum, "it takes two to tango".

The second example, which I shall call Fundamentalists, involves shooting and killing as a *joint enterprise*. Suppose two religious fundamentalists set out to kill the leader of their country on the grounds that they believe he is standing in the way of the creation of a religious state. Due to the heavy security protecting the leader the fundamentalists have to work together in order to smuggle a rifle into the building opposite the leader's residence. The first fundamentalist smuggles in the frame, barrel and magazine of the rifle, while the second smuggles in the sights and the ammunition. The collaboration does not end there. While the first fundamentalist loads the rifle and holds it steady, the second lines up the crosshairs and pulls the trigger. Here we have a case in which two agents act in conjunction in pursuit of a common goal that they simply could not achieve working alone. Although only one of the fundamentalists pulls the trigger, it seems that they are both *jointly* morally responsible for shooting and killing the leader in the sense that they were both part of a joint enterprise. That they are both jointly morally responsible holds true even though there is *not* one person S whose acts caused the consequence C to obtain.

The third example, call it Firing Squad, involves *probabilistic* shooting and killing. Imagine that an army training facility commandant has set up a firing squad whose task it is to execute a deserter. The firing squad is composed of six marksmen all of whom are good shots. The commandant instructs each of the marksmen to aim at the deserter's head from a close range and to shoot at the same moment. However, unbeknownst to the marksmen, the commandant only loads one of the six rifles with live ammunition; the other five rifles are loaded with blanks. The marksmen select the rifles at random and out of sight of the commandant. They are all wearing hearing protectors. Neither the commandant nor the marksmen know which particular marksman had been given the rifle loaded with live ammunition such that in the end nobody knows who actually shot and killed the deserter. Needless to say, all we know is that for each marksman there is a 16.67 percent probability that he shot and killed the deserter. But this is not the same as saying for each marksman that his actions caused the death of the deserter. This is unlike the case of Assassins in which, according to Fischer and Ravizza's description, there is a 100 percent probability that each assassin shot and killed the mayor. As a way of reflecting what is unknown about the shooting and killing of the deserter in Firing Squad, we might say that the marksmen *jointly* shot and killed the deserter.

On Fischer and Ravizza's positive account of moral responsibility and action-responsiveness, in order to attribute moral responsibility to a person it is necessary that he moves his body in way B at time T and this is actually sufficient to cause C to obtain. In the foregoing examples, each particular person did not move his body in a way that was actually sufficient to cause the relevant consequences to obtain. Not only that, but there are (by stipulation) no other acts or omissions which are sufficient to cause the consequence. That is to say, there are no other genuine trigger events to be screened out assuming that a trigger event is "an event which is such that, if it were to occur, it would *initiate* a causal sequence leading to C" (110–11). Nevertheless, these examples intuitively involve persons who *are* morally responsible for the consequences in question or at least are *jointly* morally responsible for the consequences in question. If I am right in this intuitive judgment,

then these examples directly challenge Fischer and Ravizza's account of action-responsiveness or their assertion that action-responsiveness is a necessary condition of moral responsibility.

What is interesting about these examples is that they exist on the margins of what Fischer and Ravizza have in mind when they say that moral responsibility for consequences may only be attributed to persons when their acts or omissions are involved in consequence-obtaining processes that are themselves "sensitive to action". This is another way of saying that the relevant acts or omissions must make a difference to what actually happens. All of my examples involve particular acts that can make a difference to what actually happens but only in the sense that they make a difference to what actually happens (as in, they cause the terminal event to obtain) provided that they are combined with other acts or omissions. To make the same point slightly differently, the acts do not qualify as triggers in the strict sense, since if the act occurs by itself, then it is not enough to initiate the consequence. Nevertheless, each act is crucial to a joint act or a set of acts that is a trigger in that sense.

10.6 Possible Replies

How might Fischer and Ravizza respond? I do not think that they have the option of accepting that action-responsiveness is not a necessary condition for attributing moral responsibility. This would be to significantly change the character of their view of moral responsibility. It would also render their account of inner mechanisms discontinuous with their account of outer processes. But neither, I think, do they have the option of simply ignoring the above examples. Examples of jointly determined consequences do not fall on the penumbra of our ordinary moral experience. On the contrary, such examples are fairly typical of the sorts of questions of moral responsibility that we face every day. Consider the joint responsibility of prison guards for death row executions, the joint responsibility of parents for the bad behaviour of their children, the joint responsibility of a group of thugs for beating up and killing a neighbour who dares to challenge them, the joint responsibility of members of terrorist organisations for planning and carrying out atrocities, the joint responsibility of multiple generations for global warming.

Alternatively, Fischer and Ravizza might try to insist that the foregoing examples are not counterexamples to their account on the grounds that their definition of action-responsiveness *does* apply to these examples after careful reflection. Recall that according to Fischer and Ravizza's definition, an act or omission is part of a process of type-P that is action-responsive with respect to a consequence-universal C if and only if the act or omission is actually sufficient for C to obtain and a different consequence would have obtained if the act or omission had not occurred and all other triggering events that would have been sufficient to cause C are not in play. Much depends on what the relevant process is. If the relevant process for each act has narrow scope, then it involves that act and only that act along with its ordinary background conditions. In which case, when it comes to the foregoing examples the

processes in question are not action-responsive with respect to their consequences. Yet if the relevant process for each act has wide scope, then it involves that act along with its ordinary background conditions and possibly other people's acts as well. In which case, the processes in question can be action-responsive in these examples.

However, this raises the difficulty of identifying in advance which is the operative process in any given example: is it the narrow scope process or the wide scope process? After all, someone might attempt to abdicate moral responsibility for the consequences of an act on the grounds that his or her act was part of a process that had narrow scope and was thereby not action-responsive in relation to the consequence, while others will argue (perhaps the victims of the act) that the person should be held morally responsible for the consequences on the grounds that the act was part of a process that had wide scope and was action-responsive.

In fact, Fischer and Ravizza "concede both that process individuation is problematic and that [they] do not have an explicit theory of process individuation" (113). Nevertheless, they insist that "all that is required for our purposes is that there be agreement about some fairly clear cases" (ibid.). But what would that agreement look like? Perhaps the idea is that the operative process will be defined as being whichever process produces an attribution of moral responsibility that most people find intuitive. But then the task of process identification becomes ad hoc. Furthermore, where it is vague whether or not someone is morally responsible, it remains vague what the relevant process is. Fischer and Ravizza seem willing to accept this outcome. "If we are unsure about an agent's moral responsibility for a consequence in precisely those cases in which we are unsure about process individuation, then at least the vagueness of our theory will match the vagueness of the phenomena it purports to analyze" (ibid.). Nevertheless, this outcome may seem disappointing to some people. Ideally (they might say) we want a theory of moral responsibility that has the power to make us sure about examples concerning which we had been previously unsure.

In light of all this, it seems to me that in order to capture our ordinary intuitions about jointly determined consequences we need new accounts of moral responsibility and of action-responsiveness which are especially designed for examples of that sort. As well as accounts of individual moral responsibility for consequences and individual action-responsiveness for consequences we need dedicated accounts of joint moral responsibility for consequences and joint action-responsiveness for consequences. With this in mind, let us say that an individual is jointly morally responsible for a consequence C if and only if he or she is morally responsible for the internal mechanism leading to his or her own act or omission *and* his or her act or omission is jointly action-responsive with respect to C along with the act or omission of at least one other individual who is also morally responsible for the internal mechanism leading to his or her own acts or omissions. An individual's act or omission is jointly action-responsive with respect to a consequence C if and only if it along with the acts or omissions of at least one other individual is part of a process of type P which is sufficient for C to obtain, and a different consequence would have obtained if it along with the other acts or omissions had not occurred and all other triggering events that would have been sufficient to cause C are not in

play. Taking into consideration my earlier discussion of consequence-universals and consequence-particulars, I leave it open as to whether C is a universal or particular.

The reader might wonder why I have not mentioned the possibility of an individual being jointly morally responsible for a consequence that has been jointly causally determined by his or her acts or omissions along with a *natural event* as opposed to the acts or omissions of at least one other person or moral agent. My hunch is that it would be linguistically odd to speak of *joint* moral responsibility in such cases, since here moral responsibility would be shared between a person and a natural event. This hunch is supported by the view that reason-responsiveness is a necessary condition for moral responsibility and by the fact that this condition cannot be satisfied by a natural event. So, I propose that in such instances we speak instead of a person being *partly* morally responsible for the given consequence, where this is by virtue of her being morally responsible for the internal mechanism involved *and* her acts or omissions being jointly action-responsive with respect to that consequence along with the natural event. The definition of joint action-responsiveness remains the same in this proposed account of *part* moral responsibility for consequences except that natural events are added to the right hand side of the biconditional.

With these accounts of joint moral responsibility for consequences and joint action-responsiveness for consequences in place, I believe that it is now possible to make intuitive judgements about my three examples. In Revolutionaries, we can say that each of the revolutionaries is jointly morally responsible for the consequence, *that the government official was shot and killed*, because (by stipulation) each revolutionary was morally responsible for his own internal mechanisms *and* both of their actions were jointly action-responsive with respect to that consequence. One might say that their actions each contributed – in a cumulative fashion – to a process that was action-responsive with respect to the consequence. I propose similar judgements for Fundamentalists and Firing Squad. In the example of Fundamentalists, the two sets of actions were jointly action-responsive – by dint of the fundamentalists' joint enterprise – with respect to the consequence, *that the leader was shot and killed*. And in Firing Squad, the acts of each of the marksmen were jointly action-responsive in respect of the consequence, *that the deserter was shot and killed* – by virtue of the fact that together their acts had a 100 percent probability of bringing about that consequence.

Before concluding, I offer the following comment on how Fischer and Ravizza might respond to my accounts. They might argue that there is no need to augment their theory with my new accounts, since they can deal with my three examples of jointly determined consequences in the following way. When an individual agent is part of a joint activity, the question of moral responsibility pertains to the group or corporate entity. We can then apply the existing account of moral responsibility to that corporate entity as the relevant *agent*. Hence, Fischer and Ravizza might say that a corporate entity is morally responsible for a certain consequence just in case its behaviour issues from its own, suitably reasons-responsive internal mechanism *and* its behaviour is part of a process which is action-responsive with respect to that consequence. This account would not be applicable to examples where an

individual's acts or omissions jointly determine a consequence along with a natural event. That is because in such situations there is no corporate entity with a suitably reasons-responsive internal mechanism. Yet it could potentially work for my three examples.

Now there certainly are occasions when it is fitting to inquire into the moral responsibility of a corporate entity or group. We often say that nations are morally responsible in this way. But what happens to the moral responsibility of individuals in all of this? Surely we do not want to lose that dimension altogether. At this point Fischer and Ravizza could try to argue that the moral responsibility of the individual flows from the moral responsibility of the group. Provided they can explain how it is that groups have suitably reasons-responsive internal mechanisms, they might argue that an individual is morally responsible for the consequences of a group's behaviour just in case he or she is a voluntary member of the group, he or she played some part in the group's reasons-responsive internal mechanism and the group's behaviour is part of a process that is action-responsive in relation to the consequences. However, none of this means that we are somehow uninterested in the particular part played by individual members in bringing about the relevant consequences. For, we cannot take it as read that every member's behaviour is action-responsive with respect to the consequences. And arguably we might want to draw a moral distinction between someone whose acts or omissions do actually figure in the relevant action-responsive process and someone whose acts or omissions do not. In addition to this, one cannot rely on the existence of group moral responsibility to explain individual moral responsibility in the case of Revolutionaries (and other similar cases) because here the individuals under discussion have no knowledge of each other and are not associates.

Hence the purpose of my schema is to account for joint moral responsibility of individuals by making it explicit that an individual's acts or omissions must be jointly action-responsive in relation to the relevant consequences along with the acts or omissions of other individuals with whom that joint moral responsibility is to be shared. What is more, I have provided a story or set of stories that hopefully explain *how it is possible* for the acts or omissions of individuals to become parts of jointly action-responsive processes along with the acts or omissions of other individuals. These stories involved associates and members of groups as well as non-associates and isolated agents. In my three examples I explained the occurrence of joint action-responsiveness in terms of *accumulation*, *joint enterprise* and *probability*. This means that I can use my schema to account for the possible existence of joint moral responsibility without necessarily having to rely on the existence of corporate entities or groups with their own suitably reasons-responsive internal mechanisms.

In this article I have put to one side the question of internal mechanisms and focused instead on external processes. (There might be a need for an account of joint moral responsibility in respect of joint internal mechanisms in the event that examples come to light which call for such an account.) I have pointed out a flaw in Fischer and Ravizza's negative argument that the inevitability of certain events or consequences does not rule out the possibility of an agent's moral

responsibility for those consequences. I have also been critical of their positive argument that moral responsibility for consequences depends on action-responsiveness. In the former case I argued that their putative counterexamples against Transfer NR and Transfer NR* are underdescribed but once fully described depend upon consequence-particulars and not consequence-universals as they claim. In the latter case I argued that their account is unable to cope with quite ordinary cases of jointly determined consequences.

References

Fischer, John, and Ravizza, Mark. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Fischer, John, and Ravizza, Mark. 2000. "Replies." *Philosophy and Phenomenological Research* 61:467–80.

Frankfurt, Harry. 1969. "Alternative Possibilities and Moral Responsibility." *Journal of Philosophy* 66:829–39.

Frankfurt, Harry. 1971. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68:5–22.

Frankfurt, Harry. 1987. "Identification and Wholeheartedness." In *Responsibility, Character, and the Emotions: New Essay in Moral Psychology*, edited by Ferdinand Schoeman. Cambridge: Cambridge University Press. pp. 27–45.

Ginet, Carl. 2006. "Working with Fischer and Ravizza's Account of Moral Responsibility." *Journal of Ethics* 10:229–53.

Glannon, Walter. 1997. "Sensitivity and Responsibility for Consequences." *Philosophical Studies* 87:223–33.

Grann, David. 2009. "Trial by Fire: Did Texas Execute an Innocent Man?" In *The New Yorker*, September 7. Available at: www.newyorker.com/reporting/2009/09/07/090907fa_fact_grann, last accessed July 1, 2011.

Judisch, Neal. 2007. "Reasons-Responsive Compatibilism and the Consequences of Belief." *The Journal of Ethics* 11:357–75.

Levy, Neil. 2002. "Excusing Responsibility for the Inevitable." *Philosophical Studies* 111:43–52.

McKenna, Michael. 2008. "Saying Good-Bye to the Direct Argument the Right Way." *Philosophical Review* 117:349–83.

Stump, Eleonore. 2000. "Review: The Direct Argument for Incompatibilism." *Philosophy and Phenomenological Research* 61:459–66.

Chapter 11
Joint Responsibility Without Individual Control: Applying the Explanation Hypothesis

Gunnar Björnsson

Abstract This paper introduces a new family of cases where agents are jointly morally responsible for outcomes over which they have no individual control, a family that resists standard ways of understanding outcome responsibility. First, the agents in these cases do not individually facilitate the outcomes and would not seem individually responsible for them if the other agents were replaced by non-agential causes. This undermines attempts to understand joint responsibility as overlapping individual responsibility; the responsibility in question is *essentially* joint. Second, the agents involved in these cases are not aware of each other's existence and do not form a social group. This undermines attempts to understand joint responsibility in terms of actual or possible joint action or joint intentions, or in terms of other social ties. Instead, it is argued that intuitions about joint responsibility are best understood given the *Explanation Hypothesis*, according to which a group of agents are seen as jointly responsible for outcomes that are suitably explained by their motivational structures, invoked collectively: something bad happened because they didn't care enough; something good happened because their dedication was extraordinary. One important consequence of the proposed account is that responsibility for outcomes of collective action is a deeply normative matter.

11.1 Joint Moral Responsibility Without Individual Control

Sometimes a number of individuals seem *jointly* morally responsible for events over which they, as individuals, had no control. Consider a simplified case:

> *The Lake:* Alice, Bill and Cecil each have a small boat in East Lake outside their town. One day last spring, each painted the boat and, unknown to the others, poured excess solvent into the lake. In the back of their heads, they all knew that this could affect the wildlife, but each of them decided that it would be a hassle to dispose of the solvent in a safe way and hoped that nothing bad would happen. However, as the solvent from all three diffused throughout the lake over the next few days, its concentration became high enough everywhere to prevent micro-organisms in the lake from reproducing during the next few weeks, thus leaving higher organisms without food and effectively wiping out all fish in the lake.

G. Björnsson (✉)
Linköping University, Linköping, Sweden

University of Gothenburg, Gothenburg, Sweden
e-mail: gunnar.bjornsson@liu.se

N.A Vincent et al. (eds.), *Moral Responsibility*, Library of Ethics and Applied Philosophy 27, DOI 10.1007/978-94-007-1878-4_11,

> The concentration of solvent exceeded the threshold for the microorganisms by quite some margin: although the solvent from only one of the three would not have been enough to kill off the fish, the solvent from any two would.

Let us assume that all three agents satisfied conditions of moral accountability. They were not being forced or manipulated to do what they did and they had both the capacity to reason and reflect on the values involved and the relevant sort of control over their own decisions and actions. Then it seems that we can rightly hold them responsible for recklessly pouring solvent into the lake. But to just about everyone that I have confronted with the case, it also seems clear that they are morally responsible *for the death of the fish*, that is, for an outcome of their actions over which they had no control as individuals. Similarly, it seems that voters can be morally responsible for the outcome of a referendum, citizens for toppling a dictatorial regime, consumers for good or bad practices of companies they patronize, and frequent flyers and drivers of SUVs for climate effects, even though, as individuals, they could not have significantly affected those outcomes, practices or effects.

The question of this paper concerns the conditions for such joint responsibility for outcomes of collective actions. In the next section, I explain why a case like *The Lake* provides difficulties for standard ways of understanding collective responsibility. In Section 11.3, I propose a preliminary analysis of joint responsibility based on variations on *The Lake*. To support this analysis, Section 11.4 introduces the *Explanation Hypothesis*, a model of our concept of moral responsibility that was developed to account for various aspects of individual moral responsibility for decisions, actions and outcomes. In Section 11.5, I show how the Explanation Hypothesis subsumes and deepens the analysis of Section 11.3. In Section 11.6, finally, I suggest a way of turning the Explanation Hypothesis' characterization of our *concept* of moral responsibility into an account of moral responsibility as such. One of the important consequences of the proposed account is that responsibility for outcomes of collective action is a deeply normative matter.

Some caveats are in order. First, the concern of this paper is *moral, retrospective responsibility for events*. Space prevents me from saying anything about the tight and interesting connections between this topic and other questions discussed under the heading of "responsibility" – questions concerning legal liability, moral or legal obligations to *ensure* outcomes or to *take* responsibility for outcomes by compensating those harmed, and questions about what characterizes responsible persons, or responsible decision procedures. Second, since the concern is with joint responsibility of *individual* agents, I will not say anything about the claim that collectives can be responsible for an outcome when *none* of their members are. (For recent defences of "autonomous" corporate responsibility, see Arnold 2006; Pettit 2007; Tännsjö 2007; Copp 2007; for criticism see Corlett 2001; Haji 2006; McKenna 2006; Miller 2007.) Third, the primary concern here is with *outcome* responsibility rather than responsibility for decisions. The conditions under which individuals are responsible for their decisions are themselves highly contestable, but I will assume that all individuals in the cases discussed are autonomous, in control of their own decisions and actions, capable of rational deliberation, suffering from no motivational maladies, and, as a result, responsible for their own acts or failures to act. Fourth, since

our concern is with difficulties pertaining specifically to the understanding of how individuals are *jointly* responsible for outcomes, I will assume that other difficulties pertaining to outcome responsibility can be overcome, in particular the fact that outcomes often depend on factors outside the agent's control. (For discussion, see Feinberg 1968:681–82; Nagel 1976; Sverdlik 1987:74; May 1992:42–45; Enoch and Marmor 2007). Finally, although it is clear that individuals can be jointly responsible for good outcomes, I will follow most of the literature and focus on responsibility for *bad* outcomes. It should be clear, however, that the discussion generalizes.

11.2 Difficulties

As we shall see, neither the standard notion of individual responsibility for outcomes, nor typical strategies for making sense of collective moral responsibility explain the intuition that the agents in cases like *The Lake* are responsible for the outcomes in question.

On a standard conception, an individual agent is morally responsible for a harm to the extent that some morally faulty aspect of her behaviour played a significant causal role in producing that harm (Feinberg 1968:674; May 1992:15). The difficulty is to see how the reckless acts of the agents in *The Lake* play a significant causal role.

We have already noted that no one agent made any difference to the survival of the fish given the other acts, so significance cannot require such *difference making*. On the other hand, the solvent contributed by each agent was causally involved in bringing about the outcome. But causal *involvement* cannot in itself be what accounts for individual responsibility for the collective outcome. Suppose that there are two solvents. Solvent X works as before, preventing microorganisms from reproducing, but it can do so by means of either of two distinct but equally powerful chemical processes, X1 and X2, depending on whether solvent Y is present. Solvent Y is itself incapable of doing any damage except in extreme concentrations, but will favour process X2 in the presence of solvent X. Suppose further that whereas Bill and Cecil poured solvent X into the lake, Alice contributed solvent Y, thus slightly changing the way the solvents from Bill and Cecil prevented micro-organisms from reproducing. Then it is not clear that she would be morally responsible for the outcome.

Intuitively, it might seem that the relevant causal involvement would have to be one of at least *facilitating* the causal process, or make it more likely to produce the outcome (cf. Petersson 2004). But while that might be true for responsibility for outcomes of individual actions, it is not required in *The Lake*. Suppose that when the concentration of solvent reaches above what would be provided by two polluters, the process by which the microorganisms are prevented from reproducing is both slowed down and made more open to possible disturbances, thus slightly decreasing the objective probability of the outcome. Then it is true of each of the polluters that he or she actually (but unwittingly) lowered the probability that the fish would die and obstructed that process to some degree, given the actual contribution from the

other two. Nevertheless, the three polluters would still seem to be jointly responsible for the death of the fish; it still died because of their actions.

Now consider the corresponding case with only one agent involved:

> *Adam's Lake*: Because of rare but naturally occurring processes, a poisonous substance is produced in the mud at the bottom of the lake. The amount would be just enough, by itself, to prevent the microorganisms from reproducing. Over the same period, Adam is painting his boat, recklessly pouring excess solvent into the lake, solvent containing the very same poisonous substance. The overall result is that the lake contains more than enough to kill off the microorganisms. In fact, at this concentration, the processes preventing the reproduction are a little slower than they would have been if Adam had not disposed of his solvent this way. In the end, though, the microorganisms are wiped out.

Though it is clear that Adam is morally responsible for recklessly pouring solvent into the lake, most people seem reluctant to say that he is responsible for the death of the fish. At the very least, it was much clearer that Alice, Bill and Cecil were so responsible in *The Lake*. This strongly suggests that the responsibility attributed to the three is fundamentally collective. *Taken together*, the faulty behaviours of Alice, Bill and Cecil clearly played a significant causal role in wiping out the fish; *individually*, they did not.

The problem posed by *The Lake* for standard accounts of responsibility for outcomes of individual action is equally a problem for attempts, like that of Stephen Sverdlik (1987), to reduce collective or shared outcome responsibility to individual outcome responsibility. But it also poses a problem for other standard attempts to understand forms of collective or shared responsibility, whether reductive or not. Since the most obvious cases in which we hold agents responsible for an outcome *as a group* are cases where they have either worked together towards some goal or failed to do so, such attempts are often cast in terms of actual or possible joint agency or joint intentions (e.g., Held 1970; Rescher 1998; Kutz 2000; Miller 2006; Sadler 2006; Shockley 2007). Less obvious and more controversial are cases where members of a community are responsible for outcomes of acts by other members because members empower and are empowered by each other, and thus "shares in what each member does, and . . . should feel responsible for what the other members do" (May 1992:11).

The Lake fits neither of these patterns. Since Alice, Bill and Cecil performed their acts independently and without knowledge of the others, they had no intentions to act together with the others. Nor is it likely that our ascription of joint responsibility relies on the assumption that they could reasonably have formed such intentions. Moreover, we have no reason to think that they form a group the members of which empower each other. For all we know, they might see each other as enemies. Still, they seem jointly morally responsible for the death of the fish.

What is clear from *The* Lake and similar examples is that a number of individuals can be jointly responsible for an outcome if, *together*, they play a significant causal role for that outcome. Structurally, this relation between the actions of the individuals and the outcome is similar to well-known attempts to analyse causes, not as necessary conditions or difference makers, but as non-redundant parts of nomically sufficient conditions for effects (Mackie 1974; cf. Wright 1988). In *The Lake*,

the actions of the three agents are pair-wise sufficient for the outcome, each action being a non-redundant part of such a pair. It might thus be tempting to explain the joint responsibility of the three agents in such terms (Braham and van Hees 2010). Unfortunately, any such attempt will run into deep problems with cases of what David Lewis (1986b) calls "causal preemption". Suppose that instead of pouring solvent into *East Lake*, Alice built a contraption that monitored the concentration of solvent in the lake and set it to empty her bucket of solvent into the lake should the level not rise high enough to kill the fish. Since Bill and Cecil contributed enough solvent, Alice's contraption was never triggered. In this case, she clearly would not be responsible for the outcome, even though her action would be a non-redundant part of sufficient conditions for the death of the fish (conditions including her action and the contribution of either Bill or Cecil).[1]

Elsewhere I have defended a way for theories of causation dealing in sufficient conditions to adequately account for cases of causal preemption (Björnsson 2007). But something more would need to be said even with such an account at hand. The fact that Adam poured solvent into the lake was a non-redundant part of a sufficient condition for the death of the fish, together with the fact that some volume of mud at the bottom of the lake emitted the same amount of poisonous substance; yet Adam's responsibility for that outcome is much less obvious than that of Alice, Bill and Cecill in *The Lake*. Apparently it matters whether the actions of other agents are involved; the fundamental problem of joint responsibility is *why*. This is where I hope to make progress.

11.3 A Preliminary Analysis of Responsibility for Outcomes of Collective Action

To understand joint responsibility, the first thing to be clear about is the required relation between the collective and the outcome for which they are responsible. As a first approximation, what is required seems to be that, together, the responsible agents play a significant role in the *explanation* of the outcome: the fish died *because of* Alice, Bill and Cecil. With some qualifications, this is very much in line with the idea that individual outcome responsibility requires that the individual's behaviour played a significant causal role in the outcome. However, talk about *causal* (as opposed to *explanatory*) role suggests that the responsible parties *brought about* or *produced* the outcome rather than merely *let it happen*, and we know that production is not required for outcome responsibility:

> *The Well:* Eric, Fiona and George are spending a Sunday afternoon in the woods, each thinking that he or she is the only person within miles. Suddenly they hear cries for help

[1] Other problems are provided by probabilistic case where there are no causally sufficient conditions for outcomes, and so-called "switching" cases, where necessary parts of sufficient conditions seem to change the way an outcome happens without being causally responsible for it (cf. the case where Alice contributes solvent Y). These are also problems for counterfactual analyses in the tradition of David Lewis (1973); for discussion, see e.g. (Collins et al. 2004; Björnsson 2007).

> coming from an area with especially dense vegetation. Although the cries are disturbing and continues for a long while, each ignores them while thinking that they could be part of a prank, or that whatever might be going on is none of their business. Had they walked in the direction of the cries, however, they would have found a woman, Hannah, who had accidentally fallen into a partially overgrown old well but was hanging onto a ledge a meter or so down, screaming for help and slowly losing her grip. Since no one came to her help, Hannah eventually fell down into the dried up well and died as she hit the rocks at the bottom. The story could have ended differently, however. One person would not have been able to pull her up without help, but had any two of those who heard her cries come to her rescue, they would have been able to save her.

It seems that if they learned the truth of what happened, Eric, Fiona and George could rightly blame themselves for not having investigated the call closer. But it also seems that they are to some extent morally responsible for the fatal *outcome* of the accident (though not, of course, for the accident itself), and they certainly seem responsible for the fact that Hannah wasn't saved. They could have saved her, but they did not. As in *The Lake*, the responsibility involved seems to be essentially collective. In a version of *The Well – Esther's Well* – Esther is the only person in place to hear Hannah's cries. Like Eric, she ignores the cries for dubious reasons; like Eric she would have been unable to save Hannah even if she had responded. But whereas Eric, Fiona and George seemed clearly responsible for the fact that Hannah wasn't saved, Esther clearly is not. *Esther's Well* highlights the essentially joint nature of Eric's, Fiona's and George's responsibility in *The Well*, just as *Adam's Lake* did in relation to *The Lake*.

In *The Well*, unlike in *The Lake*, there is a sense in which none of the three were *involved* in the process leading to the final outcome: indeed, it seems that they could all have been absent and nothing in that process would have been different (ignoring minute differences in the gravitational field and the like). Nevertheless, it seems that their inaction *explains why* Hannah wasn't saved. This is the notion of "explaining why" that seems relevant for our ordinary attribution of moral responsibility in these cases.

Thus far I have suggested that *the agents* should play a significant role in the explanation of the outcome. But more needs to be said about the required sort of involvement. As we have already seen from *The Well*, the relevant involvement need not consist of any particular sort of positive *intentional action*: perhaps Eric was sitting on a rock, Fiona climbing a tree, and George running across a meadow instead of helping Hannah. Similarly, no *decisions* on part of members of the group need to be involved in the explanation. Perhaps none of the three even considered the possibility of finding out whether they could help; perhaps they just noted, absent-mindedly, that someone seemed to be in need of help but failed to see any reason to take action. That would not seem to remove their responsibility as long as they could have considered the possibility to help, and would have done so if they had cared more about the needs of others. That no decision is needed can be made even clearer with a case involving negligent ignorance where there is no awareness of risk involved. Suppose that Alice, Bill and Cecil poured the solvent into the lake while being unaware of its lethal potential. They could still be responsible for the outcome if the reason they were unaware was that they lacked concern for the environment

or for taking in relevant information, and if that explained why they failed to react to the warning signs on the cans of solvent.

In all these variations, we might say that some morally "faulty" aspect of behaviour explains the outcome, but the behaviour seems faulty only because it is explained by the wrong sensitivity to values, or the wrong motivational structure. If Alice, Bill and Cecil were ignorant of the solvent's lethal potential due to other factors than a lack of appropriate concern, their responsibility for the death of the fish is undermined. Similarly, suppose that George was wearing headphones and did not hear Hannah's cries for help. Or suppose that he heard the cries and started walking towards the well but was trapped by impenetrable vegetation blocking his way and delaying him until it was too late. In neither case would he seem to be responsible for the outcome. The best explanation for that, it seems, is that in these cases, unlike in the original scenario, George's concern or lack of concern fails to explain why he didn't reach the well in time.

Another thing to notice is that the outcome needs to be explained by the motivational structure in a "normal" way. If Dave finds out that Alice, Bill and Cecil lack appropriate concern for the environment and draconically proceeds to poison their lake to teach them a lesson, their lack of concern might be part of the explanation of the death of the fish in the lake, but they are not thereby morally responsible for it. Similarly, if George's lack of concern for others had made him ignore a discussion of feasible paths through the forest, and if as a result he was stuck in the mud and unable to heed Hannah's call, it is not clear that he is thereby morally responsible for not having come to her rescue.

Judging from the variations of *The Lake* and *The Well*, it seems that the two groups of people are responsible for the outcomes because the outcomes are explained (in a "normal" way) by the agents' motivational structures. The fish died because Alice, Bill and Cecil lacked appropriate concern for the environment; Hannah's accident had a fatal outcome because Eric, Fiona and George lacked appropriate concern for their fellow human beings. The same seems to hold for cases of moral responsibility for *good* outcomes. Suppose that each member of a trio discovers and mends a leaking sewer out of concern for the environment and that the reduction of pollution secured by any two of them would have been enough to save the fish in the nearby lake, but not the reduction secured by only one agent. Then it would seem reasonable to say that the fish survived because these three individuals cared about the environment, and they would seem to be correspondingly (jointly) responsible for that outcome.

The question remains, however, whether we can expect this analysis to survive still further variations, and whether it generalizes to other cases of collective responsibility. Moreover, we have yet to explain why the *individuals* are jointly responsible for the outcomes, given this diagnosis. It is one thing to say that the group is responsible, another to say that the members of the group are, and it might be thought that attributions of moral responsibility in cases like these involve some kind of mistake. Perhaps our desire to hold someone responsible prompts us to *confusedly* assign joint responsibility for outcomes on the ground that (a) each individual is responsible for wrongfully risking some bad effect – an adverse environmental effects,

say – and (b) what they risked actually took place because of these wrongdoings, taken collectively. The suspicion that there is something amiss with our judgments gains force from a comparison of Alice's responsibility in *The Lake* and Adam's in *Adam's Lake*. In spite of performing identical actions the upshots of which are causally involved in bringing about the death of the fish in the same way, and in spite of the fact that their actions resulted from identical motivational structures, Alice's responsibility was *much* clearer than Adam's. And in spite of acting in the very same way as Esther for the very same reasons and having exactly the same possibility to save Hannah – i.e. none – only Eric seemed responsible for the fact that Hannah wasn't saved. This is bound to strike some readers as arbitrary.[2]

I address these issues in the next three sections. Section 11.4 introduces an independently motivated hypothesis about our concept of individual retrospective moral responsibility, the *Explanation Hypothesis*. In Section 11.5, I explain how it subsumes the analysis of joint responsibility developed in this section. This gives us reason to think that our present analysis will generalize to further cases. Moreover, it suggests that the different attributions of responsibility to Alice and Adam are no more arbitrary than attributions of outcome responsibility in general. Although the Explanation Hypothesis is primarily an empirical hypothesis about our concept of responsibility, supported by its predictive power, it strongly suggests an account of moral responsibility. In Section 11.6, finally, I introduce that account – *Explanatory Responsibility* – and discuss how it makes issues of outcome responsibility deeply normative.

11.4 The Explanation Hypothesis

In two recent papers (Björnsson and Persson 2009, 2011), Karl Persson and I have argued that a wide variety of intuitions about individual responsibility for decisions, actions and outcomes can be explained if we understand our concept of moral responsibility as shaped by our interest in holding people responsible. What follows is a brief and simplified version of that story.

People hold each other responsible for a variety of events in a variety of ways. We blame or express indignation towards people who have brought about or failed to prevent something bad for lack of proper concern, and praise or express moral admiration towards those who have brought about or let happen something good at remarkable costs to themselves. Sometimes our expressions of so-called "reactive" attitudes are as simple as a frown or a smile. At other times we are more elaborate, punishing or demanding explanation or compensation, or distributing rewards and honours. And we direct analogues of all these reactions towards ourselves.

Our interest in holding people responsible is largely an interest in shaping motivational structures – values, preferences, behavioural and emotional habits, etc – in order to promote or prevent certain kinds of actions or events that we like or dislike.

[2] See (Zimmerman 1985:116–17) for an argument that seems to assume that differences of this sort cannot make for different degrees of responsibility.

Consciously or unconsciously, we often hold ourselves and each other responsible for various outcomes so that we will behave responsibly and take into account possible outcomes of the sort that we have been held responsible for. This is not to deny that we often hold people responsible for reasons of desert, without an eye to deterring or encouraging agents or third parties. The claim is merely that general reformative interests very much drive and shape our practices of holding people responsible. (For instance, consider the way expressions of indignation are placated when agents express regret and real motivation to avoid repeats, and consider plausible evolutionary rationales for our reactive attitudes.)

In order for our practices of holding people responsible to reliably affect outcomes in this way, they need to be targeted at motivational structures of types that are (a) systematically tied to those outcomes and (b) tend to be amenable to modification when targeted by these practices, and need to be so when (c) instances of the motivational structure type explain the outcome in a salient straightforward way that supports learning.

Undoubtedly, our concept of moral responsibility plays a central role in determining whom to hold responsible for what. In particular, expressions of indignation and requests for explanation are withheld when we conclude that the putative target of these practices was not responsible for the objectionable decision, action or outcome. Since our concept of moral responsibility plays this role, it would not be surprising if it has been shaped by the need to identify proper targets for our practices of holding people responsible, identified by conditions (a) through (c) above.[3]

This provides motivation for what we call the "Explanation Hypothesis", an empirical hypothesis about the conditions under which we take people to be retrospectively morally responsible for some event:

> *The Explanation Hypothesis*: People take P to be morally responsible for E to the extent that they take[4] E to be an outcome of a type O and take P to have a motivational structure S of type M such that GET, RR and ER hold:
>
> > *General Explanatory Tendency (GET)*: Type M motivational structures are part of a reasonably common sort of significant explanation of type O outcomes.
> > *Reactive Response-ability (RR)*: Type M motivational structures tend to respond in the right way to agents being held responsible for realizing or not preventing type O outcomes.

[3] In connecting moral responsibility to reactive attitudes and practices of holding responsible, this hypothesis is closely related to a category of accounts starting with Peter Strawson's (1962) paper "Freedom and Resentment". In (Björnsson and Persson 2011) we indicate how our particular way of spelling out this connection avoids some of the standard objections raised against such accounts.

[4] In saying that people "take" GET, RR and ER to hold, I do not mean that they are consciously aware of the considerations defined by these conditions in making their judgments of responsibility under these descriptions, only that judgments are in fact determined by such considerations.

Explanatory Responsibility (ER): S is part of a significant explanation of E of the sort mentioned in GET.

My focus here will be on the two explanatory requirements, GET and, in particular, ER, but a few words are needed to avoid misunderstanding of RR. It is meant to capture the idea that certain types of motivational structures are impervious to blame, praise or other practices of holding people responsible, and that this undermines moral responsibility. RR thus explains why we typically take moral responsibility to be diminished when behaviour is driven by compulsion, phobias, severe personality disorders and extreme stress.

Since RR concerns how motivational structures respond to blame, praise, etc., it is easy to think that the Explanation Hypothesis understands judgments of moral responsibility as forward-looking, concerned with whether holding someone responsible would reform her behaviour. That would be a misunderstanding, however. The fact that someone's motivational structure is of a *type* that tends to respond in the right way does not mean that it is likely to do so in this case. A particular instance of a type that tends to respond appropriately might resist reform: disdain might satisfy RR, but disdain for morality might be self-protecting. Moreover, various extraneous factors might mask the motivational structure's disposition to react in the right way: perhaps the agent is disposed to react adversely to criticism, say, or perhaps she suffered from a stroke immediately after her action and no longer has the cognitive capacity to understand what she is held responsible for. To be directly forward-looking, judgments of moral responsibility would have to be sensitive to such masks, but they clearly are not; they are essentially backward-looking, concerned with what *explained* the outcome in question.

Among motivational states and outcomes that satisfy RR, there are basically two kinds of explanation that also satisfy GET: First, events are often explained by the fact that we want them sufficiently, as our desires guide our goal-directed cognitive mechanisms ("The trial was all due to Dr. Ortega's relentless passion for justice"; "Her tragic death was due to Mr. Inza's obsession with revenge"). Second, the fact that we do not sufficiently want something not to happen often explains why we let it happen ("The new factory was allowed to pollute the river because the CEO didn't care about the environment"; "He missed his daughter's game because he cared more about his work than about her").[5] Consequently, we take people to be responsible for a bad outcome when we think that it happened because they wanted them ("Mr. Inza is to blame for her death") or because they didn't care enough to prevent them ("The pollution is the CEO's fault"), and take people to be responsible

[5] It is an interesting question whether GET-satisfying explanations require awareness on part of the agent that the sort of outcome in question might take place or whether it can be enough that the person would have been aware and acted on the information if the person had possessed a different motivational structure. We are currently investigating this, and preparatory studies suggest that most people come down on the latter side. For some of the philosophical controversy, see (Zimmerman 2008:chap. 4; Sher 2009).

for a good outcome when it happened because they wanted it ("Dr. Ortega deserves all credit for the trial").[6]

According to the Explanation Hypothesis, our everyday concept of an *explanation why something happened* is at the core of our thinking about moral responsibility. One key feature of that concept is that it is highly *selective*. Suppose that a house has just burned down and that we are asked why. In answering, we could list a number of conditions, each of which might be a necessary part of complex sufficient condition for the outcome: there was a thunderstorm, the house was hit by lightning an hour earlier, the house consisted largely of combustible matter, there was oxygen in the air, etc.[7] All of these conditions, and countless more, might be part of a *full* causal story leading up to the fact that the house burned down, but only a small subset will stand out when we want to give a condensed explanation of that fact. When we do, the fact that the house was hit by lightning will likely grab our attention, whereas the fact that the house consisted of combustible matter or that there was oxygen in the air would be taken for granted as part of what we might call the explanatory "background". Typically, the explanatory background consists of conditions that are generally to be expected whereas attention grabbers are conditions that violate such expectations. Generally speaking, we expect houses to be built from some amount of combustible material, and we certainly expect there to be oxygen in the air, but we do not in the same way expect houses to be hit by lightning at some given time.

Our everyday notion of explanation is selective in another way too. The bolt of lightning that hit the house itself had a causal genesis, and there were numerous causal intermediaries between the fact that the house was hit by lightning and the fact that it burned to the ground. These conditions are not likely to be seen as part of the explanans, however. When we provide explanations of an event, we cite a condition that we take to provide a particularly *telling* explanation among those leading up to that event, a condition that satisfies our explanatory interests without immediately raising new and urgent why-questions. If we wonder why the house burned down and are told that the attic insulation caught fire, we will probably wonder *why* the insulation caught fire, and if we are told that there was a separation of positive and negative charges in the neighbouring atmosphere, we are likely to ask how *that* explained that the house burned down. By contrast, if we are told that the house was hit by lightning, we will probably be satisfied: we take a house's being hit by lightning to be both the sort of thing that just happens and the sort of thing that causes houses to burn down.

---

[6] It is possible that GET should be restricted to these two broad kinds of explanation.

[7] In (Björnsson 2007) I argue that our causal reasoning is *primarily* directed towards sufficient rather than necessary conditions and that this is explained by the connection between causal thinking and instrumental reasoning: instrumental reasoning is primarily directed at ensuring certain states of affairs rather than making them possible. The priority of sufficiency over necessity explains why causation is compatible with many varieties of overdetermination and ultimately explains why responsibility is not a matter of difference making. (All this simplifies matters by ignoring probabilistic causation and explanation.)

When condition ER in the Explanation Hypothesis refers to a *significant* explanation, that means an explanation that satisfies our explanatory interests and background assumptions or, differently put, fits our *explanatory frame*. The selective nature of significant explanations makes the Explanation Hypothesis a surprisingly powerful account of judgments of moral responsibility. Obviously, the hypothesis can account for the fact that we take people to be responsible for most intended outcomes of their actions: because of our powerful goal-directed mechanisms, such outcomes are straightforwardly explained with reference to what we want to achieve, and most of our everyday preferences satisfy RR. But relying on the selective nature of significant explanations also provides a unifying account of how a wide variety of otherwise disparate phenomena affect judgments of responsibility. As I have argued elsewhere (Björnsson and Persson 2009, 2011), it explains why we take it that (a) external force, (b) threats and (c) ignorance mitigate moral responsibility to various degrees, as well as why we take it that (d) those who actively participate in the production of an outcome have a higher degree of responsibility for it than those who merely allow others do it, that (e) someone who takes initiative is more responsible than someone who tags along, and that (f) agents are more responsible for known negative than for known positive side-effects that the agent does not care about. It also explains why judgments of responsibility tend to be undermined by considerations suggesting that (g) our decisions are a matter of luck, (h) our actions are, ultimately, the upshots of events over which we have no control, (i) our behaviour can be given reductionistic, mechanistic explanations, or that (j) the felt conflict between determinism and moral responsibility is lessened when people consider concrete cases, and especially cases involving grave moral transgressions.

11.5 The Explanation Hypothesis and Collective Responsibility

The explanatory power of the Explanation Hypothesis, along with its etiological motivation, gives us reason to think that the everyday concept of retrospective moral responsibility has a structure that straightforwardly incorporates our preliminary analysis of joint responsibility in Section 11.3: in cases of joint responsibility, the motivational structures of all participants are seen as parts of a significant explanation of the outcome. This gives us independent reason to expect further cases of joint responsibility to conform to the same analysis, thus providing a first answer to the generalization worry.

More specifically, the Explanation Hypothesis explains both why we take the agents of *The Lake* to be responsible for the death of the fish and why we take them to be *jointly* responsible. We see them as *responsible* for the outcome because the three conditions GET, RR and ER are satisfied for each of them, and we see them as *jointly* responsible because their motivational structures are part of a significant explanans only taken together with the motivational structures of the other two.

Start with the last claim. Compare the following two answers to the question: why did the fish in the lake die?

(1) Alice, Bill and Cecil didn't care about the environmental effects of their actions.
(2) Alice didn't care about the environmental effects of her actions.

Whereas (1) sounds like a perfectly good explanation, (2) is clearly problematic, for two reasons. First it brings attention to the fact that Alice's carelessness made no difference to the outcome because there would have been enough solvent in the lake without it, and although difference making doesn't always undermine explanatory claims it might do so in this case.[8] But (2) is also problematic because it focuses on Alice at the exclusion of Bill and Cecil who played exactly the same role in killing off the fish. Both these defects are absent in (1). That the trio didn't care about the environmental effects of their actions straightforwardly explained why they poured solvent into the lake, and the resulting concentration of solvent explained why the fish died. Of course, not all their actions or all the solvent was needed for that outcome, but there is no privileged subset of these actions that would provide a better explanans. For example, if we explained the death of the fish by mentioning the carelessness of Alice and Bill, we would misleadingly suggest that Cecil had less to do with the outcome than the other two. For that reason, such a restricted explanans would not provide us with an acceptable straightforward explanation.

Now consider the claim that the motivational structure of *each* agent satisfies GET, RR and ER for the outcome in question. First, it satisfies GET because the outcome is explained by a lack of concern to avoid that sort of outcome in the normal way. The most common explanation of this type will be one in which an *individual's* lack of concern explains the outcome, but we frequently explain outcomes in terms of attitudes of members of a group: "The kids next door play loud music because they don't care about the neighbours"; "Sweden rejected the Euro because many Swedes were afraid of losing political independence"; etc. Second, the motivational structures also satisfy RR: we have assumed that the individuals involved satisfy conditions needed for individual responsibility for decisions and action. Finally, we have just seen that the individual agent's motivational structure satisfies ER, as it is alluded to in the joint explanation given by (1).

Contrast this case with *Adam's Lake*. Just like Alice's lack of environmental concern, taken on its own, Adam's lack of concern does not itself strike us as straightforwardly explaining the death of the fish. But whereas Alice's is *part* of a significant explanation that satisfies ER, expressed in (1), it is not clear that Adam's is. For example, the following answer to the question of why the fish died in *Adam's Lake* seems strained:

(3) Adam didn't care about the environment and a poisonous substance was produced at the bottom of the lake.

[8] The model of causal judgment developed in (Björnsson 2007) explains the restricted role of difference making or counterfactual dependence in causal judgments and shows why the lack of counterfactual dependence might undermine the claim that Alice's carelessness caused or explained the death of the fish in the lake. This effect would be even stronger in the version of *The Lake* where her contribution actually lowered the probability of the outcome.

Although both conjuncts mention conditions that are part of a complete causal explanation of the death of the fish, their conjunction does not form the most *salient* explanation of the outcome. It would be considerably more natural to appeal to the fact that the lake was poisoned, as the causes of the poisoning are diverse. Moreover, among those causes, the fact that a poisonous substance was produced at the bottom of the lake would likely be seen as more significant than Adam's contribution, since it actually made a difference to the outcome.

Intuitions about *The Well* are explained almost exactly as intuitions about *The Lake*. Eric, Fiona and George are seen as jointly responsible for the fact that Hannah wasn't saved because that fact is naturally explained with reference to *their* lack of concern, but not with reference to, say, Eric's lack of concern in particular. The defect of an explanation singling out one individual is more strongly marked than in *The Lake*. "Why wasn't Hannah saved?" "Because Eric didn't care to see whether he could help!" The answer invites the reply that Eric couldn't have saved Hannah on his own, and does so even more strongly than (1) invited the reply that the fish would have died without Alice's action: at least Alice's action was directly causally involved in blocking the reproduction of the microorganisms whereas Eric's inaction made no definite difference at all.[9] (This explanatory inadequacy is of course even more accentuated in *Esther's Well*, where Esther's lack of care clearly does not explain why Hannah wasn't saved.)

What we have seen, then, is how the Explanation Hypothesis supports the diagnosis of joint responsibility provided in Section 11.3. Given that so many other aspects of our thinking about moral responsibility is well understood given this account, we should expect further variations on the cases discussed here to conform to the same pattern.

For similar reasons, we should hesitate before saying that typical intuitions about cases like *The Lake* result from confusedly attributing joint responsibility based on (i) *individual responsibility for decisions and actions* and (ii) *non-distributive collective responsibility for an outcome*, that is, collective responsibility that does not imply corresponding responsibility for members of the collective. The argument given here suggests that intuitions of joint responsibility rely on the same sort of considerations as do intuitions about individual responsibility. From the point of view of our concept of retrospective moral outcome responsibility, then, the attribution of joint responsibility is in no way arbitrary. Nor is it arbitrary, from an etiological point of view, that we should have a concept that yields this pattern of judgments; a focus on cases with a straightforward explanatory connection between

[9] The Explanation Hypothesis also implies that subtle differences in characterizations of outcomes might yield different verdicts about moral responsibility. It is intuitively clear that Eric, Fiona and George are responsible for the fact that Hannah wasn't saved, but it is less clear that they are responsible for her death. If we ask why she wasn't saved, it is natural to cite, say, the trio's lack of concern, but if we ask why she died, it is considerably more natural to cite the fact that she fell into an old well or didn't watch where she was going than to cite the non-intervention. Different explananda yield different explanatory frames: unlike the fact that she died, the fact that she wasn't saved implies that she was in danger, thus relegating her initial fall into the well to the explanatory background.

suitable motivational structures and outcomes is crucial for the sort of moral reform that much of our everyday practice of holding people responsible is aimed at. One might worry, though, that it is *unfair* that Alice should be held responsible (together with Bill and Cecil) for the death of the fish whereas Adam is not, given that both were equally reckless and contributed solvents that were similarly causally involved in processes leading to the death of the fish. But this is a familiar problem for outcome responsibility in general, not specifically for joint responsibility or for the analysis proposed here. Factors outside the control of an agent are part of what determines the outcome of her behaviour: only one of two equally reckless drivers is responsible for the death of a child, because only one had a child run out into the street in front of him; only one of two equally courageous and skilled lifeguards is responsible for having saved a life, because only one had the opportunity.

Thus far, we have seen how the Explanation Hypothesis handles cases of joint responsibility without individual control. But it also predicts that people might be seen as jointly rather than individually responsible for an outcome even in cases where each individual could have prevented the outcome. Think of a version of *The Well* where any one of Eric, Fiona and George could have saved Hannah using a winch next to the well. We might still be reluctant to say that *Eric* is responsible for the fact that Hannah wasn't saved because it arbitrarily picks out Eric at the exclusion of the other two. The significant explanans is still that *none of the three* cared enough to go see whether help was needed; that corresponds to the most natural assignment of responsibility, namely jointly, to all of them.

Another prediction, borne out by almost every discussion of distributive collective responsibility, is that we will ascribe joint responsibility in many cases where agents act together, with joint intentions, since these tend to be cases where agents' motivational structures are involved in straightforwardly explaining the intended outcome. Similarly, intuitions about corporate responsibility bear out the prediction that we will ascribe moral responsibility for outcomes to corporations (organizations, nations, clubs) insofar as we take them to have structures that both straightforwardly explain their actions or omissions and corresponding outcomes and are open to modification by practices of holding these corporations responsible (see e.g. French 1984; May and Hoffman 1991).

For both cases of joint action and cases of corporate moral responsibility, the Explanation Hypothesis predicts attributions of quite different degrees of responsibility to different members of a collective that are causally involved in producing or failing to prevent some outcome. For example, we might think that a stream has been polluted because a certain company doesn't care about the environment, but we do not thereby think that the janitor at the company headquarters is responsible for the pollution. He might have somehow facilitated the process leading to the pollution, but his motivation is not thereby part of a *significant* explanation in the way that the motivational structures of the CEO or members of the board are likely to be. And the same might be true about a member of the board who voted against the polluting activity, or even about someone who voted for it because she thought that that was the way to minimize the harm by allowing her to minimize the resulting pollution.

11.6 Explanatory Responsibility and the Normativity of Retrospective Outcome Responsibility

As we have seen, the Explanation Hypothesis promises a unified account of our judgments of individual and collective responsibility, an account that sees our ascription of essentially *joint* responsibility in cases like *The Lake* or *The Well* as integral to our thinking about moral responsibility in general. Moreover, although it does not say what the relation of moral responsibility *is*, it strongly suggests an account of that relation. Given the Explanation Hypothesis' account of our concept of moral responsibility, it might seem reasonable to assume that the relation of moral responsibility corresponds to what is identified when the concept is applied without any mistakes, that is, when GET, RR and ER hold.

Things are not quite that simple, however, because the selective nature of our explanatory judgments makes them sensitive to differences in explanatory frames. For example, it seems that when people are encouraged to abstract away from the level of detail that we employ in everyday explanations of actions and to focus on causal factors outside agents' control, they are less inclined to find motivational structures explanatorily significant, and less inclined to ascribe responsibility (Björnsson and Persson 2009, 2011). In the same way, explanatory judgments often depend on *normative* expectations or ideals. Suppose that that a child falls and breaks an arm during some rough and tumble play. A person who thinks that mothers ought to be strongly protective of their children is more likely to explain this fact with reference to the mother's lack of protective concern, and thus more likely to take the mother to be responsible for the accident.[10]

This frame-dependence of our concept of moral responsibility means that if there is a determinate, objective, truth of the matter as to whether people are responsible for certain outcomes, the "significant explanations" referred to in GET and ER needs to be restricted. The most obvious way to do so is to require that they are significant relative to a *correct explanatory frame*: relative to *correct* normative ideals, *correct* background assumptions, and *relevant* explanatory interests and explanatory perspectives. "Objectifying" the Explanation Hypothesis, we would thus get the following characterization of moral responsibility:

> *Explanatory Responsibility:* P is morally responsible for E to the extent that E is an outcome of a type O and P has a motivational structure S of type M such that GET, RR and ER hold relative to a correct explanatory frame.

Obviously, Explanatory Responsibility only implies determinate judgments of responsibility given substantial assumptions about what the correct explanatory frames are. This is not the place to defend some such assumptions,[11] but the fact

[10] For empirical data illustrating some effects of normative expectations on explanatory judgments, see e.g. (Alicke 1992; Knobe and Fraser 2008; Hitchcock and Knobe 2009; Sytsma et al. 2010).

[11] In (Björnsson and Persson 2011) we argue that explanatory frames of the sort that motivate most of our everyday judgments of moral responsibility should be preferred to the frames that are induced by sceptical arguments against moral responsibility.

that moral responsibility would depend on the correctness of normative expectations is itself a highly significant consequence.[12] Because of it, fundamental issues in normative ethics are directly relevant to questions of moral responsibility.

As an example, consider how issues of joint responsibility are affected by the disagreement about the existence of reasons to do one's own part in a cooperative scheme even when others are known not to, or to "keep one's own hands clean". Thus far, I have discussed cases where, for all the agents knew, their acts could have made a difference individually to the outcome for which they are responsible. Moreover, this feature might seem essential to the cases. For example, if Alice had poured solvent into the lake knowing for sure that it would make no significant environmental difference or even slowed down ongoing damage, that could clearly undermine her responsibility for the death of the fish as her contribution would no longer be explained with reference to a lack of care. But suppose that there are moral reasons for people to do their part in appropriate cooperative schemes that do not depend on the possibility of actually significantly furthering the ultimate point of these schemes. Then people might be jointly responsible for bad outcomes that they, as individuals, *knew* they could not prevent: if they had *all* been more concerned to do their part, the outcome would have been different.

If there are non-consequentialist reasons of this sort, their strength will also have major impact on what we are responsible for. Given high enough normative expectations that people should avoid working for or purchase the goods of organizations that are responsible for certain bad outcomes, it will seem that great many people without direct causal influence on these outcomes are nevertheless responsible for them, i.e. for such things as the effects of a company's environmental policy, the persecution of members of organized labour in undemocratic countries, or the enactment of severe oppression of civilians on occupied territories. After all, if people had cared more and been more "principled", many such things could have been very different. This in turn raises difficult questions about the relation between normative expectation and psychological realism: since it seems unlikely that people will live up to these expectations under present circumstances, are they really reasonable? If correct, Explanatory Responsibility makes clear just how such questions are central to issues of collective responsibility, by being directly relevant for the identification of significant explanations.

Acknowledgments Earlier versions of this text have been presented and received valuable input at the *International Conference on Moral Responsibility* in Delft, August 2009 at the Centre for Applied Ethics at Linköping University, at the Department of Political Science and the Department of Philosophy, Linguistic and Theory of Science at University of Gothenburg, and at the Department of Philosophy, Lund University. I am also grateful to participants at the CEU 2009 summer school on moral responsibility, and for comments from Ibo van de Poel and an anonymous reviewer for this volume.

[12] For related discussions of how normative aspects affect judgments of responsibility, see (Smiley 1992).

References

Alicke, Mark D. 1992. "Culpable Causation." *Journal of Personality and Social Psychology* 63:368–78.

Arnold, Denis G. 2006. "Corporate Moral Agency." *Midwest Studies in Philosophy* 30: 279–91.

Björnsson, Gunnar. 2007. "How Effects Depend on Their Causes, Why Causal Transitivity Fails, and Why We Care About Causation." *Philosophical Studies* 133:349–90.

Björnsson, Gunnar, and Karl Persson. 2009. "Judgments of Moral Responsibility: A Unified Account." *Society for Philosophy and Psychology*, 35th Annual Meeting 2009 PhilSci archive. http://philsci-archive.pitt.edu/archive/00004633/.

Björnsson, Gunnar, and Karl Persson. 2011. "The Explanatory Component of Moral Responsibility." *Noûs*. 45. Doi: 10.1111/j. 1468-0068.2010.00813.x

Braham, Matthew, and Martin van Hees. 2010. "An Anatomy of Moral Responsibility." http://www.rug.nl/staff/martin.van.hees/Anatomy.pdf

Collins, John, Ned Hall, and L.A. Paul. eds. 2004. *Causation and Counterfactuals*. Cambridge, MA: The MIT Press.

Copp, David. 2007. "The Collective Moral Autonomy Thesis." *Journal of Social Philosophy* 38:369–88.

Corlett, J. Angelo. 2001. "Collective Moral Responsibility." *Journal of Social Philosophy* 32: 573–84.

Enoch, David, and Andrei Marmor. 2007. "The Case Against Moral Luck." *Law and Philosophy* 26:405–36.

Feinberg, Joel. 1968. "Collective Responsibility." *The Journal of Philosophy* 65:674–88.

French, Peter A. 1984. *Collective and Corporate Responsibility*. New York, NY: Columbia University Press.

Haji, Ish. 2006. "On the Ultimate Responsibility of Collectives." *Midwest Studies in Philosophy* 30:292–308.

Held, Virginia. 1970. "Can a Random Collection of Individuals Be Morally Responsible?" *The Journal of Philosophy* 67:471–81.

Hitchcock, Chrisopher, and Joshua Knobe. 2009. "Cause and Norm." *Journal of Philosophy* 106:587–612.

Knobe, Joshua, and Ben Fraser. 2008. "Causal Judgment and Moral Judgment: Two Experiments." In *Moral Psychology*, edited by Walter Sinnott-Armstrong, Vol. 2, 441–47. Cambridge, MA: MIT Press.

Kutz, Christopher. 2000. *Complicity*. New York, NY: Columbia University Press.

Lewis, David. 1973. "Causation." *Journal of Philosophy* 70:556–67. Reprinted in Lewis 1986a, 159–72.

Lewis, David. 1986a. *Philosophical Papers*, Vol. II. New York, NY: Oxford University Press.

Lewis, David. 1986b. "Postscripts to 'Causation'." *Philosophical Papers*, Vol. II, 172–213. New York, NY: Oxford University Press.

Mackie, John. 1974. *The Cement of the Universe*. Gloucestershire: Clarendon Press.

May, Larry. 1992. *Sharing Responsibility*. Chicago, IL: University of Chicago Press.

May, Larry, and Stacey Hoffman. 1991. *Collective Responsibility: Five Decades of Debate in Theoretical and Applied Ethics*. Lanham, MD: Rowman & Littlefield.

McKenna, Michael. 2006. "Collective Responsibility and an Agent Meaning Theory." *Midwest Studies in Philosophy* 30:16–34.

Miller, Seumas. 2006. "Collective Moral Responsibility: An Individualist Account." *Midwest Studies in Philosophy* 30:176–93.

Miller, Seumas. 2007. "Against the Collective Moral Autonomy Thesis." *Journal of Social Philosophy* 38:389–409.

Nagel, Thomas. 1976. "Moral Luck." *Proceedings of the Aristotelian Society, Supplementary Volumes* 50:137–51.

Petersson, Björn. 2004. "The Second Mistake in Moral Mathematics Is Not About the Worth of Mere Participation." *Utilitas* 16:288–315.
Pettit, Philip. 2007. "Responsibility Incorporated." *Ethics* 117:171–201.
Rescher, Nicholas. 1998. "Collective Responsibility." *Journal of Social Philosophy* 29:46–58.
Sadler, Brook Jenkins. 2006. "Shared Intentions and Shared Responsibility." *Midwest Studies in Philosophy* 30:115–44.
Sher, George. 2009. *Who Knew?* New York, NY: Oxford University Press.
Shockley, Kenneth. 2007. "Programming Collective Control." *Journal of Social Philosophy* 38:442–55.
Smiley, Marion. 1992. *Moral Responsibility and the Boundaries of Community: Power and Accountability from a Pragmatic Point of View*. Chicago, IL: University of Chicago Press.
Strawson, Peter F. 1962. "Freedom and Resentment." *Proceedings of the British Academy* 48:1–25.
Sverdlik, Steven. 1987. "Collective Responsibility." *Philosophical Studies* 51:61–76.
Sytsma, Justin, Jonathan Livengood, and David Rose. 2010. "Two Types of Typicality: Rethinking the Role of Statistical Typicality in Ordinary Causal Attributions." [Preprint] URL: http://philsci-archive.pitt.edu/id/eprint/5372 (accessed 2011-07-07).
Tännsjö, Torbjörn. 2007. "The Myth of Innocence: On Collective Responsibility and Collective Punishment." *Philosophical Papers* 36:295–314.
Wright, Richard W. 1988. "Causation, Responsibility, Risk, Probability, Naked Statistics, and Proof: Pruning the Bramble Bush by Clarifying the Concepts." *Iowa Law Review* 73:1001–77.
Zimmerman, Michael. 1985. "Sharing Responsibility." *American Philosophical Quarterly* 22: 115–22.
Zimmerman, Michael J. 2008. *Living with Uncertainty*. Cambridge: Cambridge University Press.

Chapter 12
Climate Change and Collective Responsibility

Steve Vanderheiden

Abstract Can persons be held morally responsible for harmful consequences that result from the acts or omissions of their nation or society, even if they conscientiously avoid contributing toward those consequences *qua* individuals? What if those acts and omissions, together with a great many other similar ones committed against the backdrop of social norms that tolerate and even encourage such harmful behavior, contribute to a global environmental problem that gives rise to valid claims for compensation on the part of those harmed by it, but where discrete instances of harm cannot be attributed to any specific persons as directly causally responsible? Such is the case with global climate change, which results in part from social norms that are permissive of polluting activities and which often frustrate efforts to avoid them, rather than being caused by culpable individual choices alone, in which case individual fault and responsibility could more plausibly be assigned. Furthermore, the harm associated with climate change is caused by aggregated greenhouse pollution from a great many untraceable point sources rather than being the direct result of discrete emissions of heat-trapping gases by particular persons, undermining standard accounts of individual moral responsibility and thus giving rise to claims for assigning responsibility collectively instead. But holding nations and peoples collectively responsible for climate change raises objections from the perspective of individual moral responsibility, at least insofar as some persons may be implicated *qua* members of groups when they are faultless as individuals.

12.1 Introduction

Policy responses to climate change challenge conventional accounts of moral responsibility in various ways, and the normative concept of responsibility serves as the theoretical linchpin of climate justice (Vanderheiden 2011). But what does it mean to be responsible in the context of global climate change? Consider first the purely causal sense of responsibility, in which person P is responsible for outcome X insofar as P's actions bring about, intensify, or increase the probability

S. Vanderheiden (✉)
Department of Political Science, University of Colorado, Boulder, CO, USA

Centre for Applied Philosophy and Public Ethics (CAPPE), Canberra, ACT, Australia
e-mail: vanders@colorado.edu

N.A Vincent et al. (eds.), *Moral Responsibility*, Library of Ethics and Applied Philosophy 27, DOI 10.1007/978-94-007-1878-4_12,

of X occurring, as applied to unmitigated anthropogenic climate change. Through their greenhouse gas (GHG) emissions, all persons contribute toward climate change in some way since all emit carbon dioxide through respiration, but the wide variation among individual emission rates entails equally wide variation in *causal responsibility* for the various harms associated with climate change. Moreover, those harms are not expected to be evenly distributed across persons or peoples, with the least advantaged suffering disproportionately from climatic disturbances.[1] Without yet invoking any normative account of responsibility, this causal analysis reveals that those expected to suffer the most damaging effects of global climate change are among the least responsible for causing it. Such an empirical observation invites obvious normative evaluation, constructing an account of moral responsibility from its causal counterpart. It might be unobjectionable if persons were to suffer climate-related harm in exact proportion to their causal contributions to the problem, measured in terms of their GHG emissions (or, as I've argued, their luxury emissions).[2] If this was the case, greenhouse pollution could be seen as imprudent but not unjust, as persons soiled their own nests but imposed no externality costs upon others. In so doing, they would bear one kind of responsibility for their actions and resulting outcomes (i.e. warranting the harm that they impose upon themselves), but this responsibility would invite only prudential rather than moral critique. But in fact many suffer climate-related harm for which they are minimally causally responsible or not responsible at all (by a fault-based standard), while others causally contribute far more than their share to the problem but escape most of its insidious effects.

It is through such an analysis of causal responsibility for climate change that judgments concerning moral responsibility for its mitigation and adaptation can be made. One might endorse the following principle of responsibility in reference to the causes and consequences of climate change: No person should be made to suffer harm (or bear responsibility) from environmental problems beyond their share in having caused such problems, and those responsible for causing those problems should bear liability for ensuring that this is so, in proportion to that responsibility. To speak of *liability responsibility*[3] is to focus on the assignment of remedial costs necessary for ensuring that all and only those causing climate change bear its

[1] According to the Intergovernmental Panel on Climate Change, "the impacts of climate change will fall disproportionately upon developing countries and poor persons within all countries, and thereby exacerbate inequities in health status and access to adequate food, clean water, and other resources." (Intergovernmental Panel on Climate Change 2001).

[2] In contrast to the survival emissions that persons cannot avoid producing in the process of meeting basic needs and for which persons cannot be faulted, luxury emissions are associated with activities that are not necessary for survival and thus form the basis for fault-based liability for climate-related harm. See Vanderheiden (2008:especially chap. 5).

[3] I take this term from Hart (1968), but use it in a slightly different way. Hart argued that the "primary sense" of responsibility concerned charges that, if established, entail "liability to punishment or blame or other adverse treatment", but focus especially on remedial or compensatory orders that issue from assessments of liability. That is, my focus is on how determinations of liability responsibility inform who should pay for resulting harm.

costs, and in proportion to those costs. The assignment of liability costs could be used to fund *mitigation* efforts, which reduce the anthropogenic drivers of climate change by either reducing GHG emissions or sequestering those gases after they are released, or efforts at *adaptation*, which seeks to shield humans from climate-related harm once changes to the climate system are underway. In some cases, liability can be justifiably assigned in the absence of moral fault or even causal responsibility, as when potential rescuers are assigned the liability for saving famine victims by virtue of their capacity and proximity alone. This kind of capacity-based responsibility does not involve assessments of vicarious responsibility, where some are held morally responsible for the actions of others, since it can be assigned even when none are at fault for some potentially bad outcome such as a famine and does not necessarily involve blame or moral disapprobation. Typically, however, capacity-based liability is not employed when fault-based moral responsibility is available, as remedial burdens to avoid or compensate for bad outcomes are thought to accrue to faulty parties first when such parties can be identified, and only fall to capable but faultless parties when they cannot.

My interest here lies in the justification for assigning climate-related remedial liability to some apparently faultless parties when faulty parties can be readily identified, and in failing to assess such liability proportionate to either causal responsibility or fault. Both present problems from the perspective of responsibility theory, since each involves some outcome that is inconsistent with the imperative to hold persons responsible for all and only their personal contributions toward common environmental hazards, and each is complicated by assessments of collective national responsibility for climate change. In the first instance, as Paul Harris has argued, holding entire nations liability responsible for climate change obscures the wide disparity among individual GHG emission rates within both industrialized and developing countries (Harris 2009). In practice, assessments of national responsibility for climate change typically depend on average per capita emissions, making no distinctions between those well above and those well below those averages in finding citizens to be responsible for their national emissions. When entire nations engage in mitigation activities that are financed through tax revenues, such as transportation infrastructure upgrades, tax assessments that finance such activities are typically not indexed to the GHG emissions of taxpayers that are in effect held liable for those mitigation efforts. When nations commit funds toward adaptation projects, they typically also do so through general tax revenues, ignoring distinctions between the relative causal contributions made by various taxpayers. In practice, national liability for climate-related mitigation and adaptation efforts is assigned to persons on the basis of the income categories to which tax rates are indexed but to neither causal responsibility nor fault for climate change. Given that some residents of high-polluting nations take great pains to minimize their personal carbon footprints, often at considerable expense to themselves, this blanket assignment of liability seems initially to be objectionable from the perspective of individual moral responsibility.

One might view the assignment of collective responsibility to entire nations for climate change mitigation and adaptation as a mere administrative convenience,

delegating individual liability assessments to national governments, to be made on the basis of individual causation and fault. For example, under the Kyoto Protocol the United States incurred a mitigation burden on the basis of its GHG emissions that would require significant reductions in national emissions as well as costly offsets – had the U.S. ratified the protocol – but this liability assignment merely delegated authority for assessing fault and responsibility to the national government and took no position on how it was to be domestically allocated among persons, groups, and industry sectors. What looks like an assignment of collective responsibility entailing vicarious liability for at least some persons at the international level need not entail any vicarious liability at the national or subnational level. Domestically, liability could be assigned to individual persons in proportion to their contributory fault for climate change, as for example through some form of carbon tax. None would need to be held vicariously liable for climate-related harm toward which they did not personally contribute, so typical objections to collective responsibility would not necessarily follow from international burden-allocation schemes. Viewing the assessment of national responsibility for climate-related harm in this way, however, belies the important sense in which the benefits of historical patterns of greenhouse pollution accrue even to those residents of high-emissions countries that take significant pains to minimize their personal carbon footprints and the manner in which the costs of significant national mitigation and adaptation efforts must be borne by entire societies, even if those costs are equally distributed among all of its members. Part of the collective liability for climate change that is assigned to entire nations can be reduced to individual liability, but part cannot, and this latter aspect of responsibility for climate change makes opting out or absolving oneself of responsibility for climate-related harm impossible, but justifies the blanket assessments of national liability that have been part and parcel of international climate policy development.

Indeed, I shall argue that the assignment of collective liability to nations for climate change mitigation and adaptation is not a mere administrative convenience, nor does it impose morally objectionable forms of vicarious liability upon persons that cannot validly be implicated in their home country's role in causing climate change and related remedial responsibility to minimize the harm that it causes to others. Rather, it rests partly upon a kind of moral responsibility for climate-related harm from which none in industrialized nations can completely extricate themselves and in which many residents of developing countries are also complicit. As Christopher Kutz notes, "the notion of participation rather than causation is at the heart of both complicity and collective action" (Kutz 2000:138), and persons cannot help but participate in the systems of advantage and disadvantage that have been shaped by national GHG emissions patterns and against the backdrop of social norms that structure individual emission patterns. Since climate change is not caused exclusively by the isolated acts of atomistic individuals, but is also a product of collective forces like culture, public policy and social norms, entire societies can validly be viewed as collectively causing significant proportions of their national emissions, for which they must be held collectively responsible. By participating in these forces – and persons cannot help but participate in them even if

they also aim to resist and reform those social forces – persons acquire at least a minimal complicity in the harm that results such that assessments of national responsibility for climate change need not be seen as violating principles of individual moral responsibility. Unlike forms of collective responsibility that rest of the causation and fault of only part of the larger collectivity, holding some vicariously liable for harm toward which they are in no way individually responsible, the form of responsibility that best captures the sort of national responsibility on display in climate change is what Joel Feinberg terms *contributory group-fault: collective and distributive*. This model illuminates the important link between individual acts and the broader social context in which they are embedded, and offers a reply to objections lodged from the perspective of individualistic conceptions of responsibility that began this chapter. It holds that all residents of nations held liable for climate change mitigation and adaptation are responsible for climate change in at least some significant sense, and therefore that we must all take steps to mitigate our collective contributions to the problem as well as assist those who are threatened by it.

12.2 Fault, Responsibility, and International Climate Policy

In assessing national liability for climate-related harm, the 1992 United Nations Framework Convention on Climate Change (UNFCCC) declares that mitigation, adaptation, and compensation costs should be allocated among the world's nations according to their "common but differentiated responsibilities" for the problem (United Nations 1992). This judgment follows from the recognition that all nations are to some extent responsible in that persons everywhere emit some of the heat-trapping gases that cause the phenomenon, but considerable variation exists among nations in terms of their per capita emissions, levels of economic development, and past and ongoing proactive efforts to abate those hazardous emissions originating within their borders. As I have argued elsewhere (Vanderheiden 2008), this standard is best understood as invoking fault-based rather than strict liability, where parties are assigned remedial burdens based upon their relative causal contributions to the problem combined with some assessment of moral fault. Indeed, debates over the meaning of the "differentiated responsibilities" language and the burden-allocation formula that it entails have focused upon the manner in which such fault can be defensibly assessed. Strict liability (in which fault plays no role, as parties are held liable only for their causal contributions to harm) would unjustifiably jettison the morally relevant difference between the survival emissions that persons and peoples cannot avoid generating through basic activities associated with biological needs and the luxury emissions that cause the avoidable harm of anthropogenic climate change. According to the analysis that faults parties for their luxury but not survival emissions, none can be faulted for acts that are necessary for survival (as *ought* implies *can*), but assessments of fault may legitimately be applied to those activities that generate harmful emissions above the survival threshold, and agents producing

these harmful emissions may defensibly be held liable for redressing the harm that results.[4]

But fault-based liability requires complicated normative judgments that are unnecessary under assessments of strict liability, even when applied to individual persons in relatively simple cases of harm. Judgments of fault require more than determinations of causal responsibility for climate change, which can readily be quantified from existing data on historical greenhouse gas emission patterns. Fault relies on judgments of moral responsibility rather than mere causation, and is most commonly understood in terms that defy its straightforward application to collective entities like nations. Individual persons can be faulted for actions that result in harm to no one and can be faultless despite causing harm to others,[5] with the attribution of fault and assignment of liability turning on mental states that have no parallel in collective entities like nations or cultures. Given the apparent dependence of judgments of fault on cognitive capacities and forms of agency that only individual persons have and exercise, we must ask: Can collective entities like nations be held responsible for harm through fault-based liability at all? Must a remedial global climate regime instead seek out those individuals that are morally responsible for the problem, seeking to assess liability for climate-related harm through billions of separate determinations? If collective responsibility cannot coherently rest on judgments of national fault, the enterprise of assessing national liability for climate-related harm may be an indefensible one.

In allocating climate-related costs to nations rather than persons, nations are assumed to exercise a kind of collective agency that is not fully reducible to individual agency, and some persons are bound to be held responsible for the faulty acts of others. Rather than assigning remedial responsibility to individuals in proportion to their historical emissions (as in an ex post carbon tax), this approach relies upon a kind of collective responsibility where societies are held to be at fault in a way that does not reduce to faulty individual acts or decisions. At least part of my responsibility *qua* American is based not on my past individual emissions, but on the effects of national affluence on my life prospects and global climate, for which I am also responsible, even if I cannot be faulted for these advantages. Another part is based in the harmful social norms in which I have participated and/or not adequately challenged, and which condition the greenhouse-polluting acts that contribute toward high per capita rates of national emissions. I owe some compensation to the victims

[4] Some claim that this sort of backward-looking attribution of responsibility is untenable in cases where individual persons lack non-polluting options or the resources to employ them instead of polluting ones, suggesting that responsibility for climate change be instead assessed in terms of forward-looking obligations to remedy. See, for example, Fahlquist (2009). My concern here is both backward-looking at causal responsibility and moral fault as well as forward-looking toward remedies, using the former to inform the latter. To the extent that better options are not available, as where persons have no mass transit options for commuting to work and so must drive personal automobiles, individual causation is at least partly the product of collective fault in failing to make more sustainable options available.

[5] Perhaps the best account of the disjuncture between assessments of moral responsibility and the consequences of an action can be found in Nagel (1979).

of climate change because of the state policies and social norms that are complicit in causing the problem and undermining potential solutions to it, from which I have benefitted in the past and continue to do so, however reluctantly. Even if I avoid contributing to climate change directly, it might be argued, I can be faulted for my indirect contributions to the problem.

However, this judgment about the sources of my responsibility for compensating those harmed by collective activities in which I participate appears to violate the standard conditions for assessing liability, which requires contributory fault. As Joel Feinberg notes:

> First, it must be true that the responsible individual did the harmful thing in question, or at least that his action or omission made a substantial causal contribution to it. Second, the causally contributory conduct must have been in some way faulty. Finally, if the harmful conduct was truly "his fault", the requisite causal connection must have been directly between the faulty aspect of his conduct and the outcome. It is not sufficient to have caused harm and to have been at fault if the fault was irrelevant to the causing. (Feinberg 1970:222)

From this individualistic conception of causal agency, collective liability in the climate case inevitably but unjustly holds some persons responsible for harm that is in no way their fault. It does this by imposing upon entire nations a liability burden, which not only declares all its citizens to be at fault in producing a global environmental hazard but also would presumably be born by the nation at large through general taxation, rather than mandating that individuals be held liable for their personal contributions to the problem. In reply to this objection, I shall consider how citizenship in a democratic society might affect one's responsibility for this global environmental problem, beyond whatever individual responsibility one might have as greenhouse polluter. My claim is that the justification for holding an entire nation responsible for climate change depends on whether its cause is seen as aggregated individual emissions only, or whether its causes are also (and, in my view, properly) regarded as being a function of citizenship, membership in a culture, and participation in networks of social norms. If the latter, it becomes considerably more difficult (if not impossible) for Americans to extricate themselves from responsibility for the problem, regardless of their individual emissions or personal preferences, and the use of collective responsibility in climate policy becomes less problematic.

12.3 Democracy and Collective Responsibility

To what extent can persons be implicated for the polluting actions of their fellow citizens, even when they themselves conscientiously aim to minimize their individual greenhouse footprints? Does democratic citizenship diffuse responsibility for climate change among an entire populace, even when considerable variation exists among individual pollution patterns? Do citizens assume responsibility for the greenhouse pollution rates of their fellow citizens, when these are conditioned by the social norms and public policies (or lack thereof) for which they are collectively responsible? These questions aim to link democratic citizenship with collective

responsibility for those adverse consequences caused by one's fellow citizens, treating citizenship as the source of responsibilities and well as privileges and regarding the relationships of social solidarity that define citizenship as a potential source of liability for the harmful actions of others.

Climate-related harm displays some of the characteristics of a collectively produced hazard for which responsibility cannot be fully ascribed to individual citizens. Although the GHG-emitting actions and choices of individual citizens can be identified as among the causes of a nation's aggregate emissions, so too can public policies, social norms, and public infrastructure be seen as causally responsible for these collectively generated harms. The citizen driving long distances to and from work may be the proximate cause of the emissions that she produces through her automobile's petroleum combustion, but she may rightly claim that the lack of more efficient personal automobiles (itself a product of a lax regulatory state that fails to encourage automobile fuel efficiency combined with social norms that attach status to fuel-inefficient vehicles) is also partly to blame for the greenhouse pollution from her commute, as is the lack of an adequate mass transit option or affordable housing located closer to her place of employment. Such factors play a causal role in structuring her choice, making difficult or impossible more sustainable individual actions, and yet are themselves not obviously caused by identifiable individuals that could be held responsible for them. When democratic societies fail to enact adequate anti-pollution regulations, develop norms of affluent consumption that equate polluting with higher social status, and build cities and towns without a sustainable transit infrastructure or decent housing that is proximate to jobs that make such housing affordable, these failures are the fault of the group itself, even if no individual member can be faulted for them. When whole societies are held responsible for the collected but evidently faultless acts of individual members, as in holding an entire nation of reluctant car commuters responsible for their aggregate greenhouse emissions, it would seem that collective responsibility is being applied where individual responsibility would be un-warranted.

Such cases tempt us to exonerate collectively-produced harm when no group members can be held individually and fully responsible for causing it, but such exoneration would raise its own set of problems for individual responsibility. Writing about national responsibility, David Miller identifies this problem as one of ensuring that persons are held responsible for their own acts and choices, but not those of others, which he takes to comprise the normative core of individual responsibility. Linking responsibility and justice, he articulates the two-sided nature of individual responsibility, describing its imperative as holding that "as far as possible we want people to be able to control what benefits and burdens they receive, but we also want to protect them against the side effects, intended or unintended, of other people's actions" (Miller 2004:245). In cases where groups make collective decisions or engage in collective actions, even where some group members oppose those decisions or abjure those actions, it is sometimes impossible to assign responsibility to discrete individuals. In such cases, the two parts of this justice aim conflict: either we can hold entire groups responsible for consequences that are beyond the control of some members, or we fail to protect others against the harmful effects of group

actions. From the perspective of group members, it may seem entirely unjustified to hold reluctant participants in collective actions or omissions responsible for the consequences that result from those actions, particularly when some members actively oppose them, but from the perspective of the victims of group actions it is preferable to hold some group members vicariously responsible than to exonerate entire groups when culpable individual parties cannot be identified. Unless the group is held liable for its collective action, the victims of that action may be forced to bear the costs of harm for which they are not responsible, but this group liability may have to be borne by some faultless individual members if faultless external victims are to be adequately compensated for the harm that they are made to suffer.

In an example analogous to problems of assessing national responsibility for climate-related harm, Miller considers groups displaying "cooperative practices" characteristics such as a polluting firm in which a dissenting minority of its employees opposes that pollution, favoring instead the purchase of some costly anti-pollution controls in order to avoid it. Because a numerical minority, these conscientious employees are overruled by those for whom additional private costs on behalf of avoiding a public nuisance are seen as imprudent. Although opposing the group's final decision by voting against it, can they still be held responsible for the resulting pollution-related harm? Miller argues that they can be held responsible, under some circumstances, if "they are the beneficiaries of a common practice in which participants are treated fairly – they get the income and other benefits that go with the job, and they have a fair chance to influence the firm's decisions – and so they must be prepared to carry their share of the costs, and in this case the costs that stem from the external impact of the practice" (Miller 2004:253). Insofar as members have fair and meaningful opportunities to influence group actions – decisions are not made by a small clique of elites against the will of the majority, for example – the mere fact that some oppose the group's final decision cannot exonerate them from responsibility, so long as they benefit from the cooperative endeavor. As Miller claims of such groups, "participating in the practice and sharing in the benefits may be sufficient to create responsibility" (Miller 2004:253). Thus, he suggests, the more open and democratic the group, the more each member must be held responsible for its decisions, whether or not they personally supported them.

This sort of collective responsibility is essential for ensuring group accountability and preventing individuals from becoming moral free riders, harmlessly dissenting from group decisions when possible in order to create benefits for the entire group at some external costs to others and then invoking this ineffective dissent as a grounds for deflecting responsibility. If available as a means for escaping responsibility for group actions, citizens might seek to avoid the burdens and duties of citizenship *en masse*, transferring their democratic agency to unrepentant polluters (in the climate case) that can provide cover for their ongoing harm on the pretext that they would have preferred to have been legally prohibited from polluting but didn't have anyone palatable to vote for in the last election. If merely registering some opposition to harmful group actions was sufficient to exonerate individual members from fault and liability for them, when those same members could enjoy the private benefits of their reluctant public nuisance nonetheless, then dissent would cease to be

sincere or effective and could become a cynical means of obtaining the benefits of membership without accepting its burdens. Dissenters would merely be shirking their responsibility, and might be unfavorably compared to those voting in favor of harmful group actions that at least in principle accept responsibility for the public nuisance from which they derive private benefits. As Miller suggests, refusing the benefits of harmful group actions would be the only way to demonstrate the sincerity of one's opposition to them, and this sort of principled dissent would be the only way of altering this insidious incentive structure.

But forfeiting the benefits of membership in affluent industrialized democratic society is not easily accomplished, and may be altogether impossible. Some benefits are public goods from which none can be excluded, however reluctantly citizens may participate in such consequences of social affluence as democratic governance, political stability, and economic opportunity. By nature, such goods accrue to all, regardless of whether or not citizens voluntarily accept them. Insofar as democratic citizenship constitutes what Miller terms a cooperative practice, is it possible for citizens to escape from this sort of collective responsibility for climate change, short of exercising their exit option from society? Must they go beyond standard avenues of democratic participation before their opposition to some harmful policy can be regarded as adequately sincere, and would such measures release them from responsibility even if ineffective? The illogic of wishing that one's nation or residence had avoided past greenhouse pollution is especially evident. Can one tenably embrace post-materialist environmental values in a pre-industrial society, or regret the economic bases upon which many of one's inherited advantages were forged without undermining the very advantages which make such regret possible? Indeed, a complete opting out of the advantages of residing within nations whose affluence depended on high rates of greenhouse pollution may not be possible at all, but it may be possible to reduce one's personal share of responsibility by acting in ways that tend toward minimizing future bad social conduct or refusing advantages that stem from past bad conduct. The issue concerns the shares of individual responsibility for collective decisions in a democracy, including those to allow ongoing GHG pollution, and to this problem we now turn.

In considering whether fault and liability for social failures to enact sufficient climate policies can be applied to citizens themselves, including those actively encouraging the adoption of such policies, we might consider examples of vicarious fault and liability from other domains of theory. Persons are held vicariously liable for the acts of others when they specifically authorize those acts, as is paradigmatically seen in the principle-agent relationship within military hierarchies. In just war theory, soldiers in the field are obligated to follow orders without question, within reasonable limits, so that while they may be at fault for wartime atrocities, the moral blame and legal liability is typically attributed to commanding officers issuing the orders or failing to control the conduct of those under their command. But can fault and liability similarly transfer in other such principal-agent relationships? Decisions about whether to wage wars are typically made by civilian leaders rather than military commanders, so vicarious liability may likewise be transferred from military commanders to political authorities, and perhaps in turn to those citizens

of democratic states in whose name and presumably with whose consent the war is waged. Since democratic governments function as agents that are authorized by principals within the electorate, citizens are in this sense responsible for the actions of their government, even if they personally oppose them, as Miller claims. But are all citizens equally responsible for the harmful actions of their states or governments, by the mere fact of the principal-agent relationship that defines democratic governance? If so, this form of vicarious liability stretches principal-agent causality much further than does just war theory, and arguably by conflating ineffective resistance to power with acquiescence to and support of it.

Michael Walzer, in considering the case for reparations for victims of aggressive wars, notes that such reparations are generally paid for through taxation of all a nation's citizens (a form of liability), not just the active supporters of the war, and over time such that many who had nothing to do with the decision to wage war continue to bear collective responsibility for it (Walzer 1977:297). This does not, he thinks, pose a particularly difficult philosophical problem for moral responsibility so long as they are only held liable and not guilty for the war's atrocities.

Attributions of liability (as in reparations) are not necessarily attributions of legal or moral guilt, he suggests, but are rather judgments based upon the existence of harm, the finding of fault, and the demand of justice to compensate victims for their injuries. Making such responsibility collective rather than individual, even if this implicates a war's opponents along with its supporters, acknowledges the causal role of citizenship in a state's decision to wage an aggressive war. In the context of global climate policy, where the citizens of causally responsible nations may be held liable for mitigation and adaptation burdens even if they exercised no control over national climate policies, Walzer's parsing of liability and guilt may be attractive. Insofar as national responsibility for climate-related harm is translated into individual citizen responsibility for paying shares of those national liability burdens, climate-related liability resembles reparations for unjust war in that both involve culpable collective actions but questionable individual culpability for harmful state actions or omissions, and both hold individual citizens liable for this collective culpability as the only way in which the collective itself can discharge its remedial obligations. Responsibility for climate change in nondemocratic states mirrors Walzer's description of responsibility for unjust wars in those same states, as both incur obligations to compensate victims for the harm that they are made to suffer regardless of citizen control of relevant policies, and both discharge this remedial responsibility through individual assignments of shares of this collective liability. Yet, Walzer's analysis holds persons liable for decisions over which they as citizens of nondemocratic states have no control, seemingly violating the core tenet of moral responsibility, which insists that individuals be at fault if they are to be held liable for some harm.

Of course, citizenship confers far greater responsibility in democratic states than it does in authoritarian ones, and Walzer also considers the case of a state opting to wage war from open and democratic processes, arguing that more widely dispersed decision-making power in democracy connotes similarly dispersed responsibility

for bad state decisions, basing culpability on a sliding scale according to the extent to which each citizen wields their various powers of resistance. Who, he asks, should be held responsible for the decision to wage unjust war? Those "who voted for it and who cooperated in planning, initiating, and waging it" must be held most responsible for its atrocity, he argues, including those soldiers who, in their capacities as citizens though not in their capacities as soldiers, shared in the decision to wage the war. Those who voted against the war, he provisionally suggests, cannot be morally faulted for it, although they may later be held liable for harm that results. But what about those citizens who didn't vote? Walzer suggests that they are blameworthy for their "indifference and inaction" in failing to do what they could have done to oppose an unjust policy, "though they are not guilty of aggressive war." Here, though, attributions of fault-based liability would not be inappropriate.

The moral language of guilt and blame is invoked against the apathetic citizen, suggesting that omissions can be faulted alongside actions when either results in some avoidable bad outcome and that fault turns on an individual's capacity to affect group actions. This control condition mirrors that of standard accounts of individual moral responsibility, as citizens are held accountable not only for what they personally do but also for what they fail to do in politics. Even if one was to abstain from personally emitting unsustainable levels of greenhouse gases, one's failure to exercise political responsibility on behalf of sustainable climate policy confers fault and triggers liability for climate-related harm without the need for vicarious responsibility. As Larry May argues, "the degree of individual responsibility of each member of a putative group for the harm should vary based on the role each member could, counterfactually, have played in preventing the inaction" (May 1992:106). Suppose that the anti-war minority could have won the decision had they staged marches and demonstrations rather than merely voting, but they opted not to. Would they then bear responsibility? Walzer thinks so, "though to a lesser degree than those slothful citizens who did not even bother to go to the assembly", since their more active but incomplete resistance is less faulty than the predictably ineffective acquiescence of the nonvoter. Fault among citizens in democratic regimes is thus assigned in proportion to their missed opportunities to wield their various powers of citizenship in defense of justice and against injustice. Given the magnitude of the injustice of aggressive war, the democratic citizen is obligated, he argues, to "do all he can, short of frightening risks, to prevent or stop the war" (Walzer 1977:300–01).

As Walzer suggests, democracy can be regarded as "a way of distributing responsibility", and insofar as citizens have some control over their collective decisions they must also be held responsible for them. Those with more control, whether by virtue of their office or influence in democratic societies or their being in a better position to resist collective decisions outside of standard political processes, may be held assigned greater responsibility for collectively-produced harm than may those with less control, and even if all citizens are to some degree responsible for what they do together. Collective responsibility in wartime and its aftermath therefore sometimes extends even to those citizens that opposed the war at the ballot box or public forum, insofar as they did not do all they could reasonably have done to stop it. Here, citizens are the principals that collectively bear responsibility for the

decisions of the state, which acts as their agent. Collective responsibility thus serves a valuable social role in expressing and strengthening the solidarity of groups that share mutual interests or bonds of affection, strengthening norms and encouraging cooperation. But it also raises objections from the principle of responsibility, since the group's fault does not readily reduce to the faults of all individual members held liable for group actions and decisions. The same is true of collectivized responsibility for climate change, as fault that is widely disparate among fellow citizens is obscured by blanket assignments of group liability.

12.4 Social Norms and Collective Fault

Climate change may be caused by individual actions, but significant contributing causes of those actions are state policies and social norms, and in the contemporary United States neither prohibits individual emissions at levels well above those which are globally sustainable. Despite its several democratic deficits, the US government remains answerable to its citizens during periodic elections and through inter-election pressure groups, so the American citizenry must shoulder some share of responsibility for the failure of its government to make adequate domestic climate change mitigation policy, and perhaps also for its continued obstruction of global climate policy efforts, given its widespread passive support for its government's active opposition to global efforts to reduce emissions. But the government's failure to adequately address global climate change is not merely an institutional shortcoming, since social norms are too permissive of pollution to generate genuinely democratic support for taking the necessary policy steps to avoid dangerously high greenhouse gas concentrations from accumulating, much less to achieve those aims in the absence of coercive policies. Part of the problem is a public culture constructed around the personal automobile, large living spaces, high resource consumption, and little regard for the consequences of these upon the world's less fortunate. Democratic decisions ultimately reflect this culture, and the shared values and common identity it fosters create the necessary conditions for attributing collective responsibility as well as generating the preferences for which such attributions are necessary. Prior to those political decisions lies a culture that is inimical to meaningful action to reduce emissions, and that culture can only be the product of society taken as a collectivity, and irreducible to individuals.

The key to linking group fault and liability with individual acts and choices lies within the roles played by social norms and practices and the culture in which they are bound. Describing this role, Howard McGary finds individuals to be culpable for social practices in which they acquiesce, even if they don't personally support or participate in them. A practice, he writes, is "a common accepted course of action that may be over time habitual in nature; a course of action that specifies certain forms of behavior as permissible and others as impermissible with rewards and penalties assigned accordingly" (McGary 1991:79). According to McGary, individual fault is based partially on personal control over group actions, but persons

can escape responsibility for group actions without stopping those actions if they are powerless to affect group outcomes and they refuse to accept unjust enrichment from collectively-produced harm. Here, fault-negating acts of "disassociation can involve publicly denouncing a practice, but only if that is all that one can do, and a refusal to accept any enrichment that occurs as a result of the faulty practice" (McGary 1991:83). Given the inescapable benefits that accrue to members of affluent industrial societies that are primarily responsible for causing climate change, one might infer that citizens of such societies might mitigate their personal fault by publicly opposing the harmful polluting practices and those social norms in which they are embedded, along with taking care not to personally contribute to collectively-produced harm by reducing their own greenhouse emissions to sustainable levels, but that they cannot escape fault and liability altogether. Insofar as climate-related harm is at least partially caused by norms and practices, which provide the context in which individuals make choices and societies set policy, none are held vicariously responsible for environmental harm for which they are not at least at some fault.

Similarly, Kutz finds the concept of "collective intention" to be the key to understanding collective responsibility, wherein individual persons can be complicit in harmful outcomes that they cannot cause by themselves. According to his Complicity Principle:

> (Basis) I am accountable for what others do when I intentionally participate in the wrong they do or the harm they cause. (Object) I am accountable for the harm or wrong we do together, independently of the actual difference I make. (Kutz 2000:122)

What matters for holding individuals morally responsible for collective actions and their outcomes is not the control that each member exercises over group actions or the difference that each makes on their own in producing or avoiding the bad outcome, but it is their "intentional participation in a collective endeavor directly links them to the consequences of that endeavor" (Kutz 2000:138). Social norms may structure our interactions with others and condition our priorities, but we cannot be excused from culpability for contributing toward harmful outcomes merely because our actions are not expressly condemned by those norms. Rather, persons reinforce norms by participating in them and by not challenging or resisting them, but norms themselves are a collective rather than an individual product. To the extent that they are implicated in the causal processes that produce harmful acts, people can be held collectively responsible for these norms and thus the behavior they encourage. As Kutz argues, "it is both a reasonable and a necessary expectation upon agents inhabiting a crowded social landscape that they be prepared to deal with the costs imposed upon others by their freely chosen projects" (Kutz 2000:154). To the extent that persons fail to resist or challenge harmful norms, they freely endorse them and thus are complicit in the outcomes that result.

Likewise, May describes this relationship between individual and group, mediated by culture and based in group identity, as a form of metaphysical guilt, which "arises out of each person's shared identity, out of the fact that people share membership in various groups that shape who these people are, and that each person is at

least somewhat implicated in what any member of the group does" (May 1991:240). Like McGary, May argues that the "moral taint" of metaphysical guilt "arises from the fact that nothing is done to prevent the harms or at least to indicate that one disapproves of them. Due to these failures, the individual does nothing to disconnect himself or herself from those fellow group members who perpetrate harms" (May 1991:240–41). For May, however, this form of responsibility is existential rather than causal, in that by "condemning or disavowing what one's community has done", one "changes that part of one's self which is based on how one chooses to regard oneself" (May 1991:247). If we define ourselves by our choices, we acquire this taint by our choice not to disassociate from the harm that groups to which we belong cause through actions in which some but not all group members participate. Our identity is bound up in what the group does, May suggests, and we are responsible as individuals to avoid personal associations with harmful group actions even if we do not ourselves commit them.

Although May is concerned with the appropriateness of what Bernard Williams terms "agent regret" rather than legal liability (Williams 1981), his diagnosis of the link between individual failures and group fault is instructive for climate change. Individuals become tainted, according to May, from their solidaristic relationships with others in a culture that encourages or allows harmful action, and their willing participation in harmful social norms, where "cultures are both the product of individual actions and attitude, and also the producers of new actions and attitudes in the world" (May 1991:246). Because individual citizens can be faulted for acquiescing to harmful social norms that provide the context for harmful actions by others, they are not held vicariously liable under a climate policy that assigns mitigation and adaptation burdens to them, since they are responsible for one set of causes (the social norms that condition polluting behavior by others) if not for another (that behavior). What we do conditions what others see as permissible and impermissible, and May's account of the mediating role of culture recognizes this link between individual and collective agency. The permissive culture of industrialized nations like the United States implicates those who fail to sufficiently challenge the norms by which high rates of greenhouse emissions are produced, even if they do not produce those emissions personally. Because persons can be more or less faulty in their participation in this culture, fault and liability can be greater or lesser depending upon the efforts by which persons challenge this culture. Since May argues that collective responsibility cannot vary among group members, he terms this form of group-based taint shared responsibility, implying the presence of individual responsibility alongside that assigned to groups.

Others endorse similar versions of collective responsibility but deny that it must be equally shared by all group members. Feinberg, for example, considers the deeply ingrained racism practiced by whites in the post-bellum American South, where only some group members took part in acts of violence against blacks but where "99 percent of them, having been shaped by the prevailing mores, whole-heartedly approved of them" (Feinberg 1968:686). Although the vast majority actively or passively reinforced a hostile environment for blacks – faulty acts for which they may be held responsible – what about that one percent that disapproved? According to

Feinberg, the extent to which they could be implicated in the group's fault – with the community's passive supporters guilty of abetting those actually undertaking violent attacks – depends upon the pains they took to distance themselves from the acts of the majority; acts that appear to go beyond mere voice and appear closer to exit options. One might plausibly oppose this racism, he suggests, but to do so would "totally alienate" a person from the white Southern community, and such total alienation would be "unlikely to be widely found in a community that leaves its exit doors open". Commenting on the same example, Miller argues that one cannot escape collective responsibility merely by speaking out or voting against such practices, but rather "must take all reasonable steps to prevent the outcome occurring" (Miller 2004:255).

Here, the more demanding standard for extricating oneself from responsibility for harm caused by group actions – separately endorsed by Walzer, Feinberg, Miller, McGary, Kutz, and May – is more plausible, requiring democratic citizens to take all reasonable and prudent steps to avoid individually contributing to a problem, whether through individual emissions or through acquiescence with harmful social norms or solidarity with polluting activities. Merely voting against some candidate or policy is insufficient, since such passive opposition to something that finds support not only from other citizens but also through prevailing social norms amounts to too meager an attempt to avoid personally contributing to the problem. Given their vast historical and ongoing responsibility for climate change, Americans cannot merely vote for a losing candidate or ballot measure, return to their oversized homes, park the SUV in their three-car garages, and reasonably expect to be exonerated from liability for the harm associated with climate change. To do so would not only be to personally contribute toward the harm in question, but is also to fully participate in the harmful culture and reinforce the harmful norms on which the group's culpability rests. Exercising political responsibility requires more than low-cost and ineffective action. One must, as Miller argues, "take all reasonable steps" to prevent climate change from occurring – not only at the ballot box or public forum but also in everyday consumer decisions and the manifold ways in which persons may reinforce or challenge prevailing social norms; which all contribute, albeit in different ways, to the problem. And even then, the impossibility of forgoing all unjust enrichment from residing within an historical greenhouse polluter only allows citizens of industrialized nations to mitigate rather than negate their personal responsibility and thus liability for climate-related harm.

Thus, the sort of collective responsibility involved in anthropogenic climate change most closely resembles what Feinberg terms *contributory group-fault: collective and distributive*, as there is contributory fault on the part of all group members, so no one's fault is vicarious and (nearly) all are somehow at fault, if unequally so (Feinberg 1968:683). Hence, as May suggests, this sort of group liability need not run afoul of individual moral responsibility, as all are at fault for climate change, whether directly or indirectly. Fault need not be distributed equally among group members – more liability may be attributed those who are more causally responsible – but all members are responsible in some way for the harm in question, and so can be held liable for it.

12.5 Conclusion

From such examples, a preliminary picture emerges concerning each person's share of the collective responsibility that attaches to citizenship in those nations most responsible for anthropogenic climate change. Even though national per capita averages obscure a wide range within individual emissions, the aggregate rate of fossil fuel combustion within the United States is plainly too high to avoid collective (if distributive) fault, yet those patterns of behavior that generate such high emissions are supported by social norms in the same way that white racism in the post-bellum South was the product of such norms. As is the case in Feinberg's racism example, some may be more responsible than others for contributing to climate change and so might be assessed greater liability for compensating those harmed by it, but none escape some fault altogether, at least insofar as all benefit from group activities that result in greenhouse pollution, regardless of whether or not they personally support those activities. Unlike Feinberg's drowning example, however, none can be released from responsibility by the acts of others, as national GHG mitigation cannot be accomplished by a single rescuer. Such a conclusion need not dismay those pressing their governments to take action to abate national emissions as well as personally reducing their own carbon footprints, for such collective responsibility is part and parcel of democratic citizenship. Justice requires that, insofar as culpable parties owe compensation to the victims of climate change, it also requires those at greater fault to pay more than those at lesser fault. Nothing in the general claims of collective responsibility diminishes the sense of individual responsibility discussed above. More importantly, nothing in the conception of either individual or collective responsibility absolves democratic citizens of their duty to ensure that their government and society does all that it can to avoid harming others. In this sense, the sort of collective responsibility invoked in global climate policy is wholly consistent with the individualistic conceptions of responsibility upon which it has been premised.

References

Fahlquist, Jessica. 2009. "Moral Responsibility for Environmental Problems – Individual or Institutional?" *Journal of Agricultural and Environmental Ethics* 22(2):109–24.

Feinberg, Joel. 1968. "Collective Responsibility." *The Journal of Philosophy* 65:674–88.

Feinberg, Joel. 1970. *Doing and Deserving*. Princeton, NJ: Princeton University Press.

Harris, Paul. 2009. *World Ethics and Climate Change*. Edinburgh: Edinburgh University Press.

Hart, H.L.A. 1968. *Punishment and Responsibility*. New York, NY: Oxford University Press.

Intergovernmental Panel on Climate Change, Climate Change. 2001. *A Synthesis Report. A Contribution of Working Groups I, II, and III to the Third Assessment Report of the IPCC*, edited by R.T. Watson and the Core Writing Team, 12. Cambridge: Cambridge University Press.

Kutz, Christopher. 2000. *Complicity: Ethics and Law for a Collective Age*. New York, NY: Cambridge University Press.

May, Larry. 1991. "Metaphysical Guilt and Moral Taint." In *Collective Responsibility: Five Decades of Debate in Theoretical and Applied Ethics*, edited by Larry May and Stacey Hoffman, 239–54. Savage, MD: Rowman & Littlefield.

May, Larry. 1992. *Sharing Responsibility*. Chicago, IL: The University of Chicago Press.

McGary, Howard. 1991. "Morality and Collective Liability." In *Collective Responsibility: Five Decades of Debate in Theoretical and Applied Ethics*, edited by Larry May, and Stacey Hoffman, 77–87. Savage, MD: Rowman & Littlefield.

Miller, David. 2004. "Holding Nations Responsible." *Ethics* 114:240–68.

Nagel, Thomas. 1979. "Moral Luck." In *Nagel Mortal Questions*, 24–38. New York, NY: Cambridge University Press.

United Nations. 1992. *United Nations Framework Convention on Climate Change*. http://unfccc.int/essential_background/convention/background/items/2853.php

Vanderheiden, Steve. 2008. *Atmospheric Justice: A Political Theory of Climate Change*. New York, NY: Oxford University Press.

Vanderheiden, Steve. 2011. "Climate Justice as Globalized Responsibility." In *Cosmopolitan Conceptions of Climate Change*, edited by Paul Harris. Cheltenham: Edward Elgar.

Walzer, Michael. 1977. *Just and Unjust Wars*. New York, NY: Basic Books.

Williams, Bernard. 1981. *Moral Luck*. New York, NY: Cambridge University Press.

Chapter 13
Collective Responsibility, Epistemic Action and Climate Change

Seumas Miller

Abstract This article undertakes four tasks: (1) outline a theory of joint action, including multi-layered structures of joint action characteristic of organizational action; (2) utilize this theory to elaborate an account of joint epistemic action – joint action directed to the acquisition of knowledge, e.g. a team of scientists seeking to discover the cause of climate change; (3) outline an account of collective moral responsibility based on the theory of joint action (including the account of joint epistemic action); (4) apply the account of collective moral responsibility to the issue of human-induced, harmful, climate change with a view to illuminating both retrospective responsibility for causing the harm and also prospective responsibility for addressing the problem in terms of mitigation and/or adaptation.

13.1 Introduction

In this paper I set myself four connected tasks and the paper is in four sections corresponding to these tasks. In Section 13.2 I elaborate the notion of Joint Action (JA) in its various aspects, including certain forms of institutional action, e.g. the actions of a government (Miller 2001:chap. 2). This notion of joint action underpins my concept of joint epistemic action developed in Section 13.3 and my notion of collective responsibility outlined in Section 13.4. In Section 13.2 I introduce various technical notions, such as that of a multi-layered structure of JA, which are necessary in order for it to be seen that the forms of institutional action in question are at bottom the joint actions of individual human beings and, therefore, individual human beings can be held morally responsible for them.

I analyse and argue for a particular species of joint action, namely, joint epistemic action (Miller 2008). In the case of the latter, but not necessarily the former, agents have epistemic goals, e.g. scientists seeking to acquire knowledge of climate

S. Miller (✉)
Centre for Applied Philosophy and Public Ethics, Australian National University, Canberra, ACT, Australia

Charles Sturt University, Canberra, NSW, Australia

Department of Philosophy, Delft University of Technology, Delft, The Netherlands
e-mail: seumas.miller@anu.edu.au

N.A Vincent et al. (eds.), *Moral Responsibility*, Library of Ethics and Applied Philosophy 27, DOI 10.1007/978-94-007-1878-4_13,

change. The notion of joint epistemic action is a novel notion and the idea that coming to have beliefs (and therefore coming to have knowledge) could be action has been disputed. Moreover, in the theoretical literature on collective moral responsibility, indeed on moral responsibility more generally, epistemic responsibility, i.e. responsibility (including collective moral responsibility) for epistemic action has not been previously identified, nor theorized as a concept in its own right. While joint epistemic action can be morally significant, it differs from non-epistemic forms of responsibility in important respects, e.g. typically, epistemic failures are in themselves less morally serious than (so to speak) behavioural moral failures and tend to derive their moral significance from the behaviour they enable.

In Section 13.4 I outline a general notion of collective moral responsibility (Miller 2010:chap. 4). This notion includes the notion of collective moral epistemic responsibility. It is important to understand the notion of collective moral epistemic responsibility and how it relates to collective moral responsibility in general, since typically morally significant complex joint actions, including institutional action and the joint actions of large groups, have as a key element morally significant joint epistemic action for which agents can be held morally responsibility. Moreover, joint epistemic action is deeply implicated in climate change debates – the issue I address in Section 13.5 – as the influential stand taken by climate change skeptics makes abundantly clear.

In Section 13.5 I put to work the theoretical machinery developed in Sections 13.2, 13.3, and 13.4 and address the question of the moral responsibility for harmful, climate change caused by human action.

The issue of moral responsibility for human-induced, harmful climate change is an intellectually difficult and controversial one. Common sense might lead us to ascribe some moral responsibility to each of us for the harm caused. However, this has been disputed by theorists. For example Sinnott-Armstrong argues that individual citizens (who, for example, drive gas-guzzling SUVs) are not morally responsible for the harm done by human-induced climate change (Sinnott-Armstrong 2005). Sinnott-Armstrong goes on to hold governments morally responsible. Aside from illustrating the potential discrepancy between common sense and applied philosophical theory, it also raises the issue of the moral responsibility of institutions and institutional actors and the relationship of institutional moral responsibility to that of individual human beings. Specifically, is institutional action reducible to the action of individual human agents? If not, then we could have a situation in which an institution, e.g. a government, was morally responsible for harmful climate change but no individual human being (including individual members of that government) was responsible. Indeed, this is precisely the view of many influential contemporary theorists, including Margaret Gilbert (Gilbert 1992; Miller and Makela 2005) and David Copp (Copp 2007; Miller 2007, 2010). If so, then what is the analysis of institutional action that makes this possible; not, I submit, the analyses provided by atomistic individualist accounts (Narveson 2002).

A central purpose of this paper is to display the continuity between individual and institutional moral responsibility for actions: the continuity between, for example, the moral responsibility for harmful climate change of individual citizens and

that of governments. Moreover, as noted above, in the available theoretical discussions of collective moral responsibility, indeed moral responsibility more generally, epistemic responsibility, e.g. of climate scientists, has not been distinguished and differentiated from behavioural responsibility. Accordingly, my broad aim in this paper is to provide a much needed theoretical framework for this highly complex, interconnected set of issues, and to illustrate some of the ways in which this framework can be applied to illuminate the various kinds and levels of moral responsibility in play in the current climate change debates. In so doing I take myself to be demonstrating the explanatory power, fecundity and utility of my novel individualist, joint-action-based account of collective moral responsibility, on the one hand, and to be taking the initial steps toward mapping the nature and strength of some of the key moral responsibilities in play in responding to harmful, human-induced climate change, on the other.

For convenience, I will sometimes refer to joint action that is not joint epistemic action as joint behavioural action – thereby signaling the presence of bodily behaviour – notwithstanding that joint epistemic action and joint behavioural action (in this restricted sense) do not exhaust the types of joint action and also that joint epistemic action typically involves some form of observable behaviour.

13.2 Joint Action

Joint actions are actions involving a number of agents performing interdependent actions in order to realise some common goal (Miller 2001:chap. 2). Examples of joint action are: two people dancing together, a number of tradesmen building a house and a team of researchers conducting an attitudinal survey. Joint action is to be distinguished from individual action on the one hand, and from the "actions" of corporate bodies on the other. Thus an individual walking down the road or shooting at a target are instances of individual action.

The concept of joint action can be construed very narrowly or more broadly. On the most narrow construal we have what I will call, *basic* joint action. Basic joint action involves two co-present agents each of whom performs one basic individual action, and does so simultaneously with the other agent, and in relation to a collective end that is to be realised within the temporal and spatial horizons of the immediate face-to-face experience of the agents. A basic individual action is an action an agent can do at will without recourse to instruments other than his or her own body. An example of a basic individual action is raising one's arm; an example of a basic joint action is two people holding hands.

If we construe joint action more broadly we can identify a myriad of other closely related examples of joint action. Many of these involve intentions and ends directed to outcomes outside the temporal horizon of the immediate experience of the agents, e.g., two people engaging in a two hour long conversation or three people deciding to build a garden wall over the summer break. Others involve intentions and ends directed to outcomes that will exist outside the spatial horizon of the immediate experience of the agents, and involve instruments other than the agent's bodies. Thus two people might jointly fire a rocket into the extremities of the earth's atmosphere.

Still other joint actions involve very large numbers of agents, e.g., a large contingent of soldiers fighting a battle.

13.2.1 Joint Procedures (Conventions)

Basic joint actions can also be distinguished from, what I will call, joint procedures. An agent has a joint procedure to x, if he x-s in a recurring situation, and does so on condition that other agents x. (Procedures are different from repetitions of the same action in a single situation, e.g., rowing or skipping.) Thus Australians have a procedure to drive on the left hand side of the road. Each Australian drives on the left whenever he drives, and he drives on the left on condition the other agents drive on the left. Moreover, joint procedures are followed in order to achieve collective goals, e.g., to avoid car collisions. Joint procedures are in fact conventions (Miller 2001:chap. 3).

It is important to distinguish conventions from social norms. Social norms are regularities in action involving interdependence of action among members of a group, but regularities in action that are governed by a moral purpose or principle (Miller 2001:chap. 4). For example, avoiding telling lies is a social norm. Some regularities in action are both conventions and social norms, e.g., driving on the left hand side of the road. Conventions and social norms are necessary elements of institutional action and it is important for my purposes in this paper that they can be understood in individualist terms.

13.2.2 Joint Institutional Mechanisms

We can also distinguish between joint procedures (in the above sense) and, what I will call, joint mechanisms (Miller 2001:chap. 5; 2010:chap. 1). Examples of joint mechanisms are the device of tossing a coin to resolve a dispute and voting to elect a candidate to office. Joint mechanisms are typically – but not necessarily – constitutive elements of social institutions (Miller 2010) and when they are they are joint institutional mechanisms.

In some cases, that these joint mechanisms are used might be a matter of having a procedure in my earlier sense. Thus, if we decided that (within some specified range of disputes) we would always have recourse to tossing the coin, then we would have adopted a procedure in my earlier sense. Accordingly, I will call such joint mechanisms, joint procedural mechanisms.

Joint mechanisms (and, therefore, joint procedural mechanisms) consist of: (a) a complex of differentiated but interlocking actions (the input to the mechanism); (b) the result of the performance of those actions (the output of the mechanism), and; (c) the mechanism itself. Thus a given agent might vote for a candidate. He will do so only if others also vote. But further to this, there is the action of the candidates, namely, that they present themselves as candidates. That they present themselves as candidates is (in part) constitutive of the input to the voting mechanism. Voters vote

for candidates. So there is interlocking and differentiated action (the input). Further there is some result (as opposed to consequence) of the joint action; the joint action consisting of the actions of putting oneself forward as a candidate and of the actions of voting. The result is that some candidate, say, Jones is voted in (the output). That there is a result is (in part) constitutive of the mechanism. That to receive the most number of votes is to be voted in, is (in part) constitutive of the voting mechanism. Moreover, that Jones is voted in is not a collective end of all the voters. (Although it is a collective end of those who voted for Jones.) However, that the one who gets the most votes – whoever that happens to be – is voted in, is a collective end of all (or nearly all) of the voters.

Joint mechanisms play a central role in institutional action and as is the case with conventions and social norms, it is important for my purposes in this paper that they can be understood in purely individualist terms and by recourse to my core notion of joint action.

13.2.3 Organisations, Institutions and Multi-Layered Structures of Joint Action

Organizations consist of an (embodied) formal *structure* of interlocking roles (Miller 2001:chap. 5; Miller 2010:chaps. 1 and 2). An organizational role can be defined in terms of the agent (whoever it is) who performs certain tasks, the tasks themselves, procedures (in the above sense) and conventions. Moreover, unlike social groups, organizations are individuated by the kind of activity that they undertake, and also by their characteristic ends. Many organisations are also social institutions. Social institutions are organisations with a moral dimension by virtue of, for example, the authority relations they involve and the fact that their collective ends are also collective goods (Miller 2010:chap. 2). Thus governments have as a collective end the regulation of other institutions (a collective good), universities the end of discovering knowledge (a collective good), and so on.

A further defining feature of organizations is that organizational action typically consists in, what I have elsewhere termed, *a multi-layered structure of joint actions* (Miller 2001:chap. 5; Miller 2010:chaps. 1 and 2). One illustration of the notion of a layered structure of joint actions is an armed force fighting a battle. Suppose at an organizational level a number of joint actions ("actions") are severally necessary[1] and jointly sufficient to achieve some collective end. Thus the "action" of the mortar squad destroying enemy gun emplacements, the "action" of the flight of military planes providing air cover, and the "action" of the infantry platoon taking and holding the ground might be severally necessary and jointly sufficient to achieve

[1] Here there is simplification for the sake of clarity. For what is said here is not strictly correct, at least in the case of many actions performed by members of organizations. Rather, typically some threshold set of actions is necessary to achieve the end; moreover the boundaries of this set are vague.

the collective end of defeating the enemy; as such, these "actions" taken together constitute a joint action.

At the first level there are individual actions directed to three distinct collective ends: the collective ends of (respectively) destroying gun emplacements, providing air cover, and talking and holding ground. So at this level there are three joint actions, namely those of the members of the mortar squad destroying gun emplacements, the members of the flight of planes providing air cover, and the members of the infantry taking and holding ground. However, taken together these three joint actions constitute a single joint action. The collective end of this second level joint action is to defeat the enemy; and from the perspective of this second level joint action, and its collective end, these (first level joint) constitutive actions are (second level) individual actions. I note that typically in organisations not just the nature but the quantum of the individual contributions made to the collective end will differ from one agent to another.

Obviously, given the crucial role of institutions and institutional actions in harmful climate change, it is important for my purposes in this paper that organisations that are institutions can be understood in purely individualist terms and by recourse to my core notion of joint action; hence the significance of the technical notion of a multi-layered structure of joint action.

13.2.4 Collective Ends

Joint actions are interdependent actions directed toward a common goal or end. But what is such an end? This notion of a common goal or, as I shall refer to it, a collective end, is a construction out of the prior notion of an individual end. Roughly speaking, a collective end is an individual end more than one agent has, and which is such that, if it is realised, it is realised by all, or most, of the actions of the agents involved; the individual action of any given agent is only part of the means by which the end is realised. The realisation of the collective end is the bringing into existence of a state of affairs. Each agent has this state of affairs as an individual end. (It is also a state of affairs aimed at under more or less the same description by each agent.) So a collective end is a species of individual end (Miller 2001:chap. 2; Miller 2010:chap. 1).

This completes my general theory of joint action, the individualist theory which underpins my account of collective moral responsibility, including the moral responsibilities of institutions and institutional actors. However, although I have provided a theoretical account of joint action and its various permutations, I have not elaborated an account of a key species of joint action that, as climate change skepticism illustrates, is deeply implicated in harmful climate change, namely, joint epistemic action. In what follows my task is to analyse joint epistemic action in terms of my general theory of joint action, and thereby enable the provision of a theory of collective moral responsibility which accommodates both joint behavioural action and joint epistemic action. I begin with an account of epistemic action in general, since this notion is far from transparent and its existence has largely been ignored (Goldman 1999; Miller 2008).

13.3 Epistemic Action

As noted above, epistemic action is action directed to an epistemic goal, especially knowledge. I take it that the key notion of knowledge in play here is propositional knowledge, i.e. "knowing that", as opposed to, for example, "knowing how". Moreover, I will assume in what follows that knowledge is justified true belief. Naturally, it could be argued – by recourse to Gettier cases, for example – that justified true belief is not sufficient for knowledge. Here I will assume that such counter-examples can be defeated by, for example, an additional condition to the effect that the justification relied on by a true believer does not itself essentially depend on any falsehoods (Lehrer 1987).

Justified true belief is a serviceable definition of knowledge for our purposes here, albeit by no means is it a complete or definitive one. Let us now turn to the notion of epistemic action (Goldman 1999; Miller 2008).

13.3.1 Knowledge, Belief and Action

The justification of a true belief involves reasoning that provides a good and decisive justification for the believer to truly believe the content of the belief in question. Moreover, this reasoning is in large part theoretical, as opposed to practical, reasoning. For it is reasoning that terminates in a belief (or structure of beliefs), as opposed to an action.

So justification is of two sorts: justification in relation to actions, and justification in relation to beliefs.

An important difference is that actions can often be done at will, e.g., I can raise my right arm now, whereas apparently this is not so for belief acquisition, e.g., if I believe that the world is round I cannot simply decide to believe that it is flat. Moreover, this contrast can lead one to opt for a sharp division between questions of morality and questions of knowledge. Morality, it might be held, pertains only to actions (and habits, including virtues and vices) for which one can be held responsible, whereas belief acquisition – and, therefore, knowledge acquisition – not being under one's control can have no intrinsic moral dimension. To be sure, knowledge acquisition involves the application of principles of rationality, e.g. in relation to good and bad evidence for one's beliefs, but such principles do not include, on this view, any specifically moral principles.

However, it is doubtful that moral properties can be ascribed only to the actions, agential conditions or other states for which someone can be held morally responsible (usually the person who performed the action, is possessed of the condition or caused the state in question). Consider someone brought up in a racist society or a paedophile who was himself routinely subject to sexual abuse as a child and developed paedophilia as a consequence. Indeed, lack of autonomy and, therefore, of the ability to act with moral responsibility is itself a moral deficiency, notwithstanding that one might not be responsible for this lack, e.g. if one was raised as a slave or became a drug addict in one's mother's womb. If it be objected that *someone* is morally responsible for these deficiencies of character, even if not the

person possessed of them, it can be replied that this is not obviously so in some of these examples, such as the racist society which has never had any exposure to non-racism, or the drug addict whose drug addict mother became pregnant as a result of being raped. But in any case there are examples where it is clearly not the case that anyone is morally responsible, such as that of a person whose violent temper is caused by some neurological defect.

Moreover, the contrast between actions and beliefs should not be overstated.[2] First, even if one cannot freely choose between believing that p and believing that not p, one can certainly in many instances freely choose to have neither the belief that p nor the belief that not p; one can do so by refraining from inquiring or otherwise investigating whether or not it is the case that p. And, of course, in such cases typically one can freely choose to investigate whether or not that p, in which case one is in effect choosing to come to have either the belief that p or the belief that not p, depending on the outcome of the investigation. In short, one can often freely choose between an absence of belief and the presence of belief with respect to some matter. For many beliefs are, and can only be, acquired after a process of investigation, e.g. the belief that Sutcliffe is the Yorkshire Ripper (or that he is not) could not have been acquired if detectives had not decided to investigate the murders of Yorkshire prostitutes. That is, without this act of will – to conduct an investigation – the detectives would simply not have had a belief as to the identity of the Yorkshire Ripper; they would have remained in a state of ignorance.

So much for choosing whether or not to have a belief with respect to some matter; but are there cases in which one can freely choose between having the belief that p and having the belief that not p? As long as the notion of freely choosing is understood broadly, then it seems that there are many such cases.

Beliefs are often the terminal point of an act of judgement, and evidence-based acts of judgement are typically freely performed. Consider, for example, an examinee who comes to believe on the basis of a series of calculations that the answer to a complex mathematical problem is zero. The examinee is not absolutely certain that the answer is zero; after all she well knows that she could have made a mistake. However, after checking she is very confident that her judgement is correct. As it turns out the examinee has given the right answer based on valid mathematical reasoning. Surely the inference based judgement that terminated in her belief that the answer was zero was freely performed; certainly she is held responsible for providing this answer and marked, awarded prizes and so on, accordingly. Here I am not simply claiming that she freely chose to try to answer the mathematical problem, although this is also true. Rather I am claiming that although her act of judgement was “compelled by logic”, it was, nevertheless, freely performed. Indeed, in these

[2] I accept the arguments of Montmarquet (Montmarquet 1993: chap. 1) to the conclusion that one can be directly responsible for some of one’s beliefs, i.e. that one’s responsibility for some of one’s beliefs is not dependent on one’s responsibility for some action that led to those beliefs. In short, doxastic responsibility does not reduce to responsibility for actions. However, if I (and Montmarquet) turn out to be wrong in this regard, the basic arguments in this chapter could be recast in terms of a notion of doxastic responsibility as a form of responsibility for actions.

types of case there is no tension between the "compulsion" of logic and the exercise of freedom.

Now consider a second example. The belief that Sutcliffe was the Yorkshire Ripper was formed as a result of the detectives' judgement that on the basis of the evidence gathered he was the Yorkshire Ripper. Naturally, their judgement was not freely made in the sense that the detectives could have made any old judgement that they felt like making, including a judgement that was completely inconsistent with the evidence. But freely performed judgements are not to be identified with capricious or irrational judgements.

In this respect judgements are akin to actions in general; an action that is "compelled" by reason does not thereby cease to be a freely chosen action. Suppose that A desires to go home immediately after work to relax and have dinner, and also that A has promised A's spouse that A will do so; in addition, suppose that A has no other competing desires or obligations, and also that the only available means for A to get home is for A to take the bus. Needless to say, A takes the bus home. For A has good and decisive reasons to take the bus home, and A has no reasons to perform any competing action. So A's taking the bus home is "compelled" by reason; but it is no less a freely performed action for being so. It remains true, of course, that A could have chosen to do otherwise than take the bus home, albeit this might have been somewhat psychologically difficult for A, given A is a rational being.

Now suppose a police officer uses a tazer gun on an armed offender in self defence, and could not have done otherwise if he was to preserve his own life. Here the police officer has acted freely, notwithstanding that "he had no other choice" rationally speaking; that is, his action was fully rationally justified and the alternative (to allow himself to be killed) was without rational justification (assuming tazer guns are a form of non-lethal force). Indeed, it may well be that given the threat to his own life, he would have found it psychologically difficult not to use the tazer gun; nevertheless, his action was freely performed.

So the sharp contrast drawn between belief formation and actions with respect to being freely chosen does not hold up. Moreover, our examples have revealed a further analogy. In both the case of the rational use of the tazer gun and the evidence-based judgement of Sutcliffe's guilt there is a further dimension, namely, a moral dimension. In both cases the "actions" have moral significance. In one case, a life is at stake, in the other case life imprisonment (at least) is at stake. Accordingly, it is doubtful that the epistemic sphere (pertaining to knowledge) exists entirely independently of morality, notwithstanding that knowledge acquisition can be, and often is, without moral significance.

One can be held morally responsible for failing to pursue knowledge or for arriving at false beliefs on the basis of sloppy evidence gathering. Naturally, the knowledge in question must have moral significance. In criminal investigations, for example, the knowledge (or false belief) in question almost always has moral significance, at least potentially, since crimes are typically immoral acts. In scientific work on climate change, at least in contemporary settings, the knowledge acquired has moral significance by virtue of the potential for massive harm that climate change brings with it.

And there is a further point to be made here. Truth is teleological: assertions, judgements, beliefs and so on aim at the truth and are deficient in so far as they fail and are false. So qua maker of judgements, former of beliefs, assertor and so on, one ought to aim at the truth and do so for its own sake. To this extent truth-aiming is an epistemic norm that is intrinsic to the kinds of epistemic acts in question. However, some kinds of knowledge are worth having for their own sake by virtue of being an inherent human good, as opposed to being merely instrumentally valuable or of trivial interest. More specifically, such knowledge is inherently valuable by virtue of being a constitutive element of human understanding. This is an additional property of some epistemic states; it is not a necessary feature of all epistemic states. Accordingly, the norms governing human understanding are not simply the purely epistemic norms governing the narrower practice of knowledge acquisition per se. Moreover, in the light of human understanding being a human good, the norms governing human understanding are moral, or quasi-moral, norms.

13.3.2 Joint Epistemic Action

Thus far we have discussed joint action and epistemic action in general terms. It is now time to bring these two notions together and focus on the resultant notion of joint epistemic action. Joint epistemic action is joint action that has a collective epistemic end, e.g. the acquisition of knowledge.

Naturally, many truth-aiming attitudes or actions, such as beliefs, inferences, perceptual judgements, assertions to one-self, and so on are individual, not joint, actions or attitudes. Moreover, I am not an advocate of irreducibly collective beliefs (Gilbert 1992:chap. 5) or of collective subjects that engage in some form of irreducibly, non-individualist reasoning or communication (Pettit 2001:chap. 5; Miller and Makela 2005). However, much knowledge acquisition is performed by individuals acting jointly to realize a collective epistemic end, e.g. a team of linguistic experts who work together to discover the meaning of some lost manuscript.

Many speech actions, including acts of assertion are joint epistemic acts (Miller 2008). Assertion typically involves a speaker (assertor) and a hearer (audience) and the practice of assertion involves, I suggest, three connected features. Firstly, assertions have a communicative purpose or end; they are acts performed in order to transmit beliefs, true beliefs, and often knowledge (justified true beliefs). Secondly, they are acts at least constrained by considerations of truthfulness. It is central to the practice of assertion that participants (in general) aim at the truth, or at least try to avoid falsity. Thirdly, speakers purport to be, or represent themselves as, or make out that they are, aiming at the truth. Indeed, they make out not only that they are aiming at the truth, but also that they have succeeded in "hitting" the truth.

Assertions are joint actions and, more specifically, joint epistemic actions. Assertion is a joint epistemic action involving a speaker and hearer each having as their end that the hearer gain knowledge and that each has this end is a matter of mutual knowledge, i.e. the end in question is a collective end.

Naturally, speakers can assert insincerely in which case they do not have the collective end of knowledge acquisition by the hearer; hence the definition of assertion involves the weaker condition mentioned above, namely, that speakers purport to be aiming at the truth.

Assertions and like speech acts are governed by epistemic norms. Thus it is held that the speaker epistemically ought to tell the truth and that the hearer epistemically ought to trust the speaker; if these norms are infringed then the collective end is likely not to be realized. Moreover, this collective end itself has a normative character in the context of the speech act; the acquisition of knowledge by the hearer ought to be the aim and result of an assertion. Indeed, I suggest that this normative condition with respect to the collective end of assertions is in part definitional of this type of speech act and explains, in particular, how it is that speakers purport to be aiming at the truth (Miller 2008).

These epistemic norms are often also social norms in my above-mentioned sense, i.e. they are mutually believed by a community (at least one community) to have moral content. Hence there is the community-wide (and surely correct) belief that a speaker morally ought not to mislead a hearer by pretending to comply with the epistemic norm to tell the truth but not do so; in most (all?) linguistic communities lying is (correctly) mutually believed to be morally wrong.

More generally, epistemic action, including the acquisition of knowledge, involves procedures, e.g. verification procedures such as observation, interviewing suspects. Many of these procedures are epistemic norms.

These procedures include joint procedures (i.e. conventions, as described above). Thus each uses the procedure, given others do and they realize a collective end, namely, that we come to have knowledge. For example, testimony is a joint procedure in this sense.

These joint procedures include joint institutional mechanisms (as described above). For example, the replication of experiments by scientists is a joint institutional mechanism in this sense. The collective end here is that the initial experimental outcome is verified or remains unverified (and is possibly disconfirmed); the result is that (say) the initial experimental outcome is verified (as opposed to remaining unverified or being disconfirmed).

As is the case with joint action more generally, some joint epistemic action has no moral significance, e.g. two people jointly solving a crossword puzzle. However some joint epistemic action has profound moral significance, e.g. designing the atom bomb, discovering the cure for cancer. Moreover, some epistemic joint procedures and joint procedural mechanisms have moral significance, e.g. clinical trials for drugs.

Notwithstanding these similarities with respect to moral significance between joint epistemic action and joint action more generally there are important differences. I do not have the space to elaborate these differences here. However, I note that, typically, epistemic actions – including joint epistemic actions – are in themselves less morally significant than behavioural actions. Specifically, epistemic failures are in themselves less morally serious than behavioural failures. This is in

part because epistemic actions (and omissions) tend to derive their moral significance from the behaviour they enable or (more indirectly) from the consequences of the behaviour that they enable. Thus the discovery of the properties of various combinations of chemicals and, specifically, how to make gunpowder (joint epistemic activity) enabled military forces to more effectively destroy buildings in which civilians were housed and to commit other atrocities of war (joint behavioural activity). But this epistemic activity did not cause these morally reprehensible outcomes and those who discovered the chemical properties and processes in question are not responsible for war crimes; rather these civilian deaths were intentionally caused by the military personnel in question and they should be the ones brought before the relevant war crimes tribunal. On the other hand, joint epistemic action in which the epistemic end is qua collective end aimed at as a means to achieve evil, as in the case of scientists engaged in designing and building weapons of mass destruction is a different matter; in such cases joint epistemic action might attract a high degree of (collective) moral responsibility for the final outcome.

As we saw was the case with joint behavioural action, there are some multi-layered structures of joint epistemic action, including joint epistemic actions undertaken by the members of organizations. Consider a major crime squad undertaking a criminal investigation.

At the first level there is joint epistemic action. Here are three instances of this. (1) The victim communicates the crime (assault) and a description of the offender to a police officer; note that the speech act of assertion is a joint epistemic action, as is the process of asking and answering questions. (2) Two detectives interview a suspect to determine his motive and opportunity; this is a joint epistemic action. (3) The forensic team analyses the physical evidence, e.g. DNA; this is a joint epistemic action.

At the second level there is a joint epistemic action constituted by the three first level joint actions but with an additional collective end, namely, that of solving the crime. The crime squad solves the crime; this is a joint epistemic action. Moreover, the crime squad solving the crime is a multi-layered structure of joint epistemic action comprised, as it is, of the three first level joint epistemic actions directed to the second level collective end of solving the crime.

13.4 Collective Moral Responsibility

Let me now outline my account of collective moral responsibility (Miller 2006). I note that this account is underpinned by my analyses of joint action (including joint epistemic action), and by the various technical notions derived from basic joint action (e.g. joint mechanisms, multi-layered structures of joint action) that are required for the understanding of institutional action. I further note the above-mentioned desideratum that my account of collective moral responsibility display the continuity between collective moral responsibility for joint behavioural action and collective moral responsibility for joint epistemic action, notwithstanding the differences between them (in particular, the typically greater stringency of moral

obligations with respect to joint behavioural action by comparison with moral obligations with respect to joint epistemic action).

We need first to distinguish some different senses of responsibility (Miller 2001:chap. 8; Miller 2010:chap. 4). Sometimes to say that someone is responsible for an action is to say that the person had a reason, or reasons, to perform some action, then formed an intention to perform that action (or not to perform it), and finally acted (or refrained from acting) on that intention, and did so on the basis of that reason(s). Note that an important category of reasons for actions are ends, goals or purposes; an agent's reason for performing an action is often that the action realises a goal the agent has. As we have seen, such goals could include epistemic goals in which case the action is an epistemic action. I will dub this sense of being responsible for an action "natural responsibility", i.e. intentionally performing an action and doing so for a reason.

On other occasions what is meant by the term, "being responsible for an action", is that the person in question occupies a certain institutional role, and that the occupant of that role is the person who has the institutionally determined duty to decide what is to be done in relation to certain matters. For example, the computer maintenance person in an office has the responsibility to fix the computers in the office, irrespective of whether or not he does so, or even contemplates doing so. Since fixing a computer will typically involve an epistemic task of finding out the cause of the problem, the computer maintenance person has an institutional responsibility to perform certain epistemic tasks.

A third sense of "being responsible" for an action, is a species of our second sense. If the matters in respect of which the occupant of an institutional role has an institutionally determined duty to decide what is to be done, include ordering other agents to perform, or not to perform, certain actions, then the occupant of the role is responsible for those actions performed by those other agents. We say of such a person that he is responsible for the actions of other persons in virtue of being the person in authority over them. Thus a schoolteacher might be held to be responsible for some of the epistemic failures of her students, e.g. their failure to learn the letters of the alphabet.

The fourth sense of responsibility is, of course, moral responsibility. Roughly speaking, an agent is held to be morally responsible for an action – including an epistemic action – if the agent was responsible for that action in one of our first three senses of "responsible", and that action is morally significant.

Here the notion of a morally significant action is an imprecise term of art designed to cast the net widely rather than a precise definitional term. An action is obviously morally significant if it morally ought to be performed or morally ought not to be performed. Thus the action could be intrinsically morally wrong, as in the case of a human rights violation. Or the action might be the means to a morally good or bad end, or the outcome that it actually had might be morally good or morally bad. However, an action might be morally significant even though it is neither the case that it morally ought to be performed nor the case that it morally ought not to be performed. It might be morally permissible, for example, or it might be one

option in a moral dilemma in which there is nothing to choose, morally speaking, between the available options.

As with behavioural actions, epistemic actions can be morally significant in a number of ways. As we saw above, they can be morally significant by virtue of the ends that they serve or the outcome that they produce, e.g. discovering the cure for cancer and thereby saving lives. Moreover, knowledge and false belief can have moral significance independently of their consequences. Some knowledge (and the corresponding false beliefs) is morally significant simply in virtue of being by definition (so to speak) moral knowledge, e.g. knowledge that the guilty ought not to be punished. Again, some knowledge and corresponding false beliefs have an implicit moral content and, as such, are morally significant. For example, the false moral belief (based on an error in the identification procedure) among police and the community that a suspect is a paedophile.

Some knowledge is such that certain persons have a moral right to it; accordingly, the epistemic acts that realize these rights have moral significance. For example, perhaps citizens have a right to know about climate change. Again, there are rights that others not know certain things, e.g. some privacy rights.

We can now make the following quasi-definitional claim concerning moral responsibility:

1. If an agent is responsible for an action – including an epistemic action – in the first, second or third senses of being responsible, and the action is morally significant, then – other things being equal – the agent is morally responsible for that action, and – other things being equal – can reasonably attract moral praise or blame and (possibly) punishment or reward for the action.

Here the term "action" also refers to omissions and to the intended outcomes of actions. Further, the first "other things being equal" clause is intended to be cashed in terms of the capacity for moral agency; for example, a psychopath might not have the capacity to make moral judgements and thus ought not to be held morally responsible for his actions. The second "other things being equal" clause is to be cashed in terms of justificatory or exculpatory conditions, such as that the agent was not coerced, could not reasonably have foreseen the consequences of his or her action, and so on. It is also important to note that – consistent with this quasi-definition – agents can be held morally responsible for the morally significant foreseeable outcomes of their actions and omissions, notwithstanding that such outcomes were not intended or otherwise aimed at.

Having distinguished four senses of responsibility, including moral responsibility, let me now turn directly to collective responsibility.

As is the case with individual responsibility, we can distinguish four senses of collective responsibility. In the first instance I will do so in relation to joint actions.

Agents who perform a joint action – including a joint epistemic action – are responsible for that action in the first sense of collective responsibility. Accordingly, to say that they are collectively responsible for the action is just to say that they performed the joint action. That is, they each had a collective end, each intentionally

performed their contributory action, and each did so because each mutually believed the other would perform his contributory action, and that therefore the collective end would be realised.

Here it is important to note that each agent is individually (naturally) responsible for performing his contributory action, and responsible by virtue of the fact that he intentionally performed this action, and the action was not intentionally performed by anyone else. Of course the other agents (or agent) believe that he is performing, or is going to perform, the contributory action in question. But mere possession of such a belief is not sufficient for the ascription of responsibility to the *believer* for performing the individual action in question. So what are the agents *collectively* (naturally) responsible for? The agents are *collectively* (naturally) responsible for the realisation of the (collective) end which results from their contributory actions. Consider a team of three detectives trying to solve a burglary; one is conducting an interview of the suspect, one an interview of the witness, and the third is checking fingerprints found at the crime scene. Each is individually (naturally) responsible for completing his task (assuming the tasks are completed, i.e. the fingerprints are found to match those of the suspect, the witness identifies the suspect as the burglar and the suspect confesses). Moreover, the three detectives are collectively (naturally) responsible for bringing it about that the crime is solved.

Again, if the occupants of an institutional role (or roles) have an institutionally determined obligation to perform some joint action – including a joint epistemic action – then those individuals are collectively responsible for its performance, in our second sense of collectively responsibility. This is the case in our detective scenario. Here there is a joint institutional obligation to realise the collective end of the joint epistemic action in question, namely, to solve the crime. In addition, there is a set of derived individual institutional obligations; each of the participating detectives has an individual institutional obligation to perform his or her contributory action. (The derivation of these individual institutional obligations relies on the fact that if each performs his or her contributory action then it is probable that the collective end will be realised.)

There is a third sense of collective responsibility which might be thought to correspond to the third sense of individual responsibility. The third sense of individual responsibility concerns those in authority and is a species of institutional responsibility. Suppose the members of the cabinet of country A (consisting of the prime minister and her cabinet ministers) collectively decide to exercise their institutionally determined right to abandon the country's carbon tax in the light of its unpopularity in the electorate. The cabinet is collectively (institutionally) responsible for this policy change.

There are a number of things to emphasise here. First, the notion of responsibility in question here is, at least in the first instance, institutional – as opposed to moral – responsibility.

Second, the "decisions" of committees, as opposed to the individual decisions of the members of committees, need to be analysed in terms of the notion of a joint institutional mechanism introduced above. So the "decision" of the cabinet can be analysed as follows. At one level each member of the cabinet voted for or against

the carbon tax policy; and let us assume some voted in the affirmative, and others in the negative. But at another level each member of the cabinet agreed to abide by the outcome of the vote; each voted having as a collective end that the outcome with a majority of the votes in its favour would be pursued. Accordingly, the members of the cabinet were jointly institutionally responsible for the policy change, i.e. the cabinet was collectively institutionally responsible for the change.

A corresponding example in relation to joint epistemic action would be a decision on the part of an investigations management committee in a police organisation not to investigate a case of reported fraud on the grounds that in the context of limited investigative resources due to government cutbacks, the reported fraud in question was less serious and less likely to be solved than other competing pending fraud cases.

Third, in so far as an organisation, such as a government comprised of the prime minister, the cabinet ministers and the members of the supporting government bureaucracy, makes and implements a policy then, by virtue of the earlier introduced notion of a multi-layered structure of joint action, the participating organisational actors (i.e. the relevant individual human beings who occupy those organisational positions) can, at least in principle, be held collectively institutionally responsible for that policy being in place. Naturally, there are differential degrees of responsibility attaching to different members of an institution, e.g. some lower echelon institutional actors may have diminished institutional responsibility by virtue of their subordinate role. (There is a further complication here in the case of the traditional Westminster system of government. For under that system there is a convention whereby the cabinet minister is taken to be institutionally responsible for certain failings of the members of his or her bureaucracy. However, this does not affect the general point that all the members of an organisation could be held collectively institutionally responsible for realising some collective end of that organisation on the basis of having contributed to it qua occupant of their organisational role.)

What of the fourth sense of collective responsibility, collective *moral* responsibility? Collective moral responsibility is a species of joint responsibility on the view that I am advocating. Accordingly, each agent is individually morally responsible, but conditionally on the others being individually morally responsible: there is interdependence in respect of moral responsibility. This account of collective moral responsibility arises naturally out of the account of joint actions. It also parallels the account given of individual moral responsibility.

Thus we can make our second quasi-definitional claim about moral responsibility:

2. If agents are collectively responsible for the realisation of a joint action – including a joint epistemic action – in the first, second or third senses of collective responsibility, and if the joint action is morally significant then – other things being equal – the agents are collectively morally responsible for that joint action, and – other things being equal – can reasonably attract moral praise or blame, and (possibly) punishment or reward for performing it.

As is the case with the above corresponding claim in respect of individual responsibility, the term "joint action" refers to joint omissions and also to outcomes that are the collective end of prior joint actions (as in the case of multilayered structures of joint action). Moreover, the first "other things being equal" clause is intended to be cashed in terms of capacity for moral agency, and the second in terms of justificatory or exculpatory conditions. Finally – and consistent with this quasi-definition – agents can be held collectively morally responsible for the morally significant foreseeable outcomes of their joint actions and omissions, notwithstanding that such outcomes were not collective ends of the agents in question.

In accordance with this second definitional claim, collective moral responsibility for epistemic states can legitimately be ascribed to a set of agents if those epistemic states are morally significant and are the collective ends. Moreover, as in effect just mentioned, agents can be held collectively morally responsible for the morally significant foreseeable outcomes of their joint epistemic actions.

Suppose, for example, that a team of scientists working for the military discover how to make the atomic bomb; the scientists are collectively morally responsible for this morally significant knowledge and potentially (to some degree) morally responsible for the intended or foreseeable use to which it is put by the military. However, as noted above, the matter of the degree, if any, of collective moral responsibility for the indirect morally significant outcomes of joint epistemic action is far from clear cut.

Thus far we have elaborated, and displayed the relationships between, various theoretical notions, notably those of joint action, joint epistemic action, joint institutional mechanisms, multilayered structures of joint action, institutional action and collective moral responsibility. Importantly, my notion of collective moral responsibility crucially depends on the prior defined notions of joint action and joint epistemic action; moreover, the notion of institutional action – applicable when it is the collective responsibility of institutions and institutional actors that is in question – crucially depends on the notions of a joint institutional mechanism and of a multilayered structure of joint action. It is now time to apply this theoretical machinery to the question of collective moral responsibility for harmful, climate change caused by human action.

13.5 Climate Change

Evidently the emission into the atmosphere of excessive quantities of GHG or green house gases (importantly carbon, and to a lesser extent methane) produced by human activities (notably the burning of fossil fuels) are causing changes in global climactic conditions (especially global warming), which are in turn likely to have catastrophic consequences for human and other life forms on the planet, if the rate of emissions is not slowed and ultimately stabilised at an acceptable level. The changes in question include the melting of the ice-caps and consequent rising sea levels, variations in seasonal rainfall patterns which impact negatively on food production, and increased levels of natural disasters such as hurricanes, tsunamis and the like. While there is

dispute about the direct empirical evidence for global warming and what, if anything, ought to be done by way of response, there is general agreement in relation to the high and increasing levels of human-induced carbon emissions, in particular, and the reality of the "greenhouse effect" (Gardiner 2004; Vanderheiden 2008).[3] Moreover, it is indisputable that thus far (i.e. since the Industrial Revolution in the late 18th century) it is the developed economies that have contributed the lion's share of human-induced carbon emissions, albeit developing economies, notably China and India, are now major contributors.

In what follows I abstract away from the details, ignore extreme forms of climate skepticism and simply assume that the human race is likely to suffer catastrophe at some point in the future unless it addresses the problem of human-induced climate change and does so quite soon.[4]

Let us first address the issue of collective responsibility with respect to a relatively discrete, self-contained, type of environmental problem, namely, the required response to immediately pending disasters such as tsunamis and focus, in particular, on the collective responsibility for joint epistemic action in such cases. I then turn to the generic, multi-level, underlying and long term problem of collective moral responsibility for harmful climate change caused by humans. In doing so I differentiate the various kinds of collective behavioural and epistemic responsibilities in play and display the relationships between them.

Consider the 2004 Indian Ocean tsunami that devastated areas of Indonesia, Thailand and Sri Lanka. The relevant officials in these countries mutually know that tsunamis can bring death and destruction on a large scale. Accordingly, there is a need (among other things) for early warning systems to enable people to avert disaster, at least to the extent of preventing or minimising loss of life. So there is a collective (behavioural) end which is also a collective (moral) good (Miller 2010:chap. 2), namely, that members of the relevant community (or communities) avert disaster. However, there is in addition a collective epistemic end the realization of which is necessary to achieve the prior collective behavioral end, namely, that there be mutual knowledge of any impending tsunami. Given that the collective epistemic end is a necessary means to achieving a collective moral good, then the collective epistemic end has moral significance and so, by the lights of our above-described account of collective responsibility, there is (other things being equal) a collective moral responsibility to realize this epistemic end.

As we have seen, in the tsunami scenario joint action is required to realize a collective epistemic end, namely, mutual knowledge of any impending tsunami. However, the required "joint action" is actually a multi-layered structure of joint epistemic action (in our above-described sense) with constitutive joint epistemic

[3] The "greenhouse effect" works roughly as follows. GHG simultaneously admit short wave solar radiation while blocking some of the long wave radiation emanating from the earth's surface thereby ensuring that the temperature at the earth's surface is greater than it otherwise would be.

[4] Weaker epistemic assumptions are, of course, consistent with accepting the need to act to avoid catastrophe, e.g. that catastrophe has a 50 percent chance of taking place if we do not act or even that we are not sure of the probability in question.

actions e.g. the design of the early warning system. Note that such large scale, multi-layered structures of joint actions are typically constituted by both joint behavioural actions, e.g. constructing detectors of seismic activity, and joint epistemic actions, e.g. communications with respect to any impending tsunami.

Designing, building and deploying an early warning system for tsunamis in the Indian Ocean is a multi-layered structure of joint action which is also an institution (Miller 2010:chap. 2) in which roles are established, tasks allocated and so on, in order to realize the overarching collective epistemic end. Accordingly, there is no barrier to holding the individuals in question collectively (i.e. jointly) morally responsible for the existence of such a system or, alternatively, for its non-existence (a joint omission). Moreover, in the light of the moral significance of averting disaster and saving lives, the institutional responsibilities of the role occupants to perform joint epistemic tasks such as monitoring undersea volcanic activity that are the necessary means to avert such disasters, are collective moral responsibilities and quite stringent ones (albeit they fall short of being morally responsible for the deaths themselves).

Let us now turn to the generic global issue of collective moral responsibility for harmful, climate change caused by human action.[5]

The Intergovernmental Panel on Climate Change's (IPCC) 1990 Report drew the world's attention to harmful climate change consequent upon (in particular) humanly produced carbon emissions. Accordingly, since 1990 each one of millions of the earth's human inhabitants, especially in the developed world, have not only made a minute causal contribution to current massive environmental damage and consequent large-scale harm to humans, e.g. climate change causing rising sea levels and flooding of Pacific Island villages, they have done so knowingly (in some sense, but see below). Can we conclude from this that the millions in question are collectively morally responsible for the harm already done and the future harm already in train? Naturally, we here rely on the above-described theoretical account of collective moral responsibility, since the meaning in ordinary language of the term, "collective moral responsibility", is more or less indeterminate and (as noted above) if one turns to the theorists one finds an array of competing theoretical accounts with diverse practical implications.

To assist in the identification of the nature, and determination of the strength, of the moral responsibility in play in the climate change scenario, let us contrast that scenario with the following stabbing scenario. Assume that in the stabbing scenario there are just five men who each intentionally stab a sixth man, Smith, having as an end (a collective end) that Smith die as a consequence of his wounds, and having no good reason to kill Smith. Assume further that Smith does in fact die from these stabbings, albeit the stab wounds inflicted by any one of the agents was neither necessary nor sufficient for Smith's death. Clearly, each of the five is fully

[5] I will not concern myself in what follows with the common resource or "sink" issue; roughly, the issue of the injustice arising from the fact that the citizens of developed economies have exhausted the limited capacity of the earth to absorb carbon emissions and, thereby, denied others from their "fair share" of that common resource (Gardiner 2004).

morally responsible for deliberately and unjustifiably killing Smith and, if the facts of the case emerged, each would be found guilty of murder by a competent court of law. I note that the stabbing scenario straightforwardly exemplifies collective moral responsibility as presented in my above account, i.e. there is a morally significant collective end which is aimed at by each and to which each makes an individual intentional causal contribution.

One important difference between the climate change and the stabbing case as described thus far is that in the stabbing case, but not in the climate change case, each had as an end that the harm be done; in the climate change scenario there are foreseen untoward consequences but they are not intended or otherwise aimed at. Let us then adjust the stabbing case so that each stabs Smith in the knowledge that he would die but without having this as an end. I note that each of the men remains morally responsible for killing Smith; a court of law would find them guilty, if not of murder, then of a somewhat lesser, but still very serious, offence such as culpable homicide; moreover, by the lights of our theory there is no collective moral responsibility – because no joint action as such – however, there is aggregate individual moral responsibility (a notion sometimes confused with collective moral responsibility). At any rate, the two scenarios are now apparently relevantly similar, the only difference being the number of participants (millions versus a handful) and the magnitude of the causal contribution that each makes (minute versus substantial). Naturally, these differences are morally important, however I am trying to identify additional moral considerations.

Let us further elaborate the climate change scenario. Each of us unavoidably produces carbon emissions and, therefore, necessarily makes some contribution to the total quantum of carbon emissions produced by human activity; each of us has to do so in order to survive. By contrast, each of the five men does not need to stab or otherwise interfere with Smith in order to survive. Nevertheless, if each of us had reduced our carbon emissions to the level required for us to survive (or even somewhat above that level), i.e. if each of us had foregone luxury emissions, then the harm consequent upon our 1990–2010 emissions would in turn have been reduced to a morally acceptable level.

Assume that the large scale harm caused by this total quantum of luxury emissions was foreseeable. Thus each individual (or most of them) was aware of the likelihood of the harm consequent upon this quantum of luxury emissions. Assume further that each individual, considered on his or her own, could have avoided the production of his or her contributing luxury emissions, e.g. by selling his or her car and any of his appliances which use a large amount of electricity generated by burning coal, installing a solar energy heater in his roof, becoming a vegetarian, and quitting his or her job at a petrol station in favour of going on welfare. Accordingly, each is not only fully, individually, naturally responsible for the minute luxury emissions he or she individually produced, each is also fully, individually, *morally* responsible for those emissions since they have moral significance; they are a causal contribution to the large-scale harm. Is it morally wrong to do something which is in itself morally innocuous, but which you know will make a tiny causal contribution to a massive harm? (Naturally, there are morally

relevant differences in the size of contributions made by individuals and, crucially, differences between the average (and aggregate) contributions of the members of developed nation-states and those of undeveloped and developing nation-states (Pickering, Vanderheiden and Miller 2010).) Surely it is, at least in some cases. If so, then it is presumably a minor wrongdoing. At any rate, I am going to assume that in the climate change scenario each of the millions is fully morally responsible for a minor wrongdoing (in the sense of knowingly, albeit unintentionally, contributing causally to harming others). Similarly, each of the five men in the stabbing scenario is fully, naturally and morally responsible for his knife stabs – it being a further question whether each is fully morally responsible for Smith's death.

As we have seen, the millions considered in aggregate are *causally* responsible for the large-scale harm done by the carbon emissions. (And being causally responsible for harming others is typically a morally relevant consideration, including in relation to climate change, albeit it does not constitute moral responsibility in the sense elaborated above since it does not necessarily involve knowledge that the harm will be caused (Shue 1992).) Similarly the five men are in aggregate causally responsible for the death of Smith.

However, each of the five men is in addition fully morally responsible for (let us say) culpable homicide. By contrast, it would be absurd to claim that each of us is fully morally responsible for the large scale harm caused by the totality of 1990–2010 luxury carbon emissions, e.g. Jones is not fully morally responsible for the loss of habitats and lives consequent upon the climate change in question. Rather each of the millions has at most a radically diminished moral responsibility for the large-scale harm resulting from the 1990–2010 emissions.

Doubtless, the reason for the absurdity of the claim of full individual moral responsibility for the massive harm lies in part in the large numbers involved in the climate change scenario and the fact that each makes a tiny causal contribution to harming (for the most part) future persons. Moreover, in the climate change scenario the action performed by each (his or her carbon emissions) are not harmful per se, but rather in aggregate have harmful effects that are in the distant future, at the end of a long and complex causal chain, and (most of) the persons in harm's way are notional in the sense that they do not yet exist. In such contexts of causal responsibility, moral responsibility is diffuse (and is a species of aggregate individual moral responsibility, as opposed to collective moral responsibility per se). Moreover, the idea of moral responsibility is likely to be somewhat inchoate in the minds of the agents in question, and likely also (relatedly) to lack a strong psychological underpinning.

In these respects, aggregate individual moral responsibility for harmful climate change is different from the aggregate moral responsibility of the five men for the wrongful killing of Smith.

So far so good, but I suggest that we have still not identified all the important moral differences between the two scenarios. What moral consideration is there, in addition to those just mentioned, by virtue of which each of the five men is fully morally responsible for Smith's death, but each of us is not fully morally responsible

for the harm consequent on 1990–2010 luxury carbon emissions? I suggest that a key difference is that practically speaking – as opposed to as a matter of logic – the five men could have acted otherwise and, if they had, Smith would still be alive, but the millions could not have acted so as to avert the harm done by the 1990–2010 emissions (future emissions and the consequent harm are another matter – see below). Let me defend this claim.

The two main positive responses to human-induced, harmful, climate change are mitigation and adaptation measures. Mitigation measures are aimed at reducing carbon emissions and consist of interventions in the causal chain at the point at which human activities cause environmental damage (e.g. by emitting excessive quantities of carbon). Adaptation measures are interventions in the causal chain at the point at which environmental damage, e.g. rising sea levels resulting from global warming, causes harm to humans, e.g. flooding of coastal villages. Thus relocating to higher ground is adaptation. Presumably in the long term mitigation must take priority, since in the long term ever-increasing carbon emissions will make the planet uninhabitable. At any rate, I take it that it is the reshaping of existing institutions, and the development of new technologies, in the service of mitigation and/or adaptation that is the principal means by which to avert the harm to present and future humans caused by environmentally damaging emissions and, specifically, a necessary means if 1990–2010 luxury emissions were to have been reduced to the level at which the consequent harm would in turn not rise above a morally acceptable level.[6]

Accordingly, only if each (or most or a very large percentage) of the millions of the earth's human inhabitants could have, jointly with the others (or most of the others), during the period 1990–2010, formed a collective end to avert the harm consequent upon luxury emissions, and devised and deployed the institutional and technological means to realize this end, e.g. mutual knowledge of required emission reduction targets, "clean" energy organisations, compliance mechanisms, then is it the case that all (or most) of the millions are collectively morally responsible for the harm caused by 1990–2010 luxury emissions. Note the dependence of the realization of a collective behavioural end on joint epistemic action (the collective end of which is mutual knowledge of emission reduction targets).

However, I suggest that between 1990 and 2010 each (or most) of the (relevant) millions could not reasonably be expected to have, jointly with the others, formed the requisite collective end, and designed and implemented the technological and institutional means to realize it. For one thing, and notwithstanding the 1990 IPCC Report, it is not the case that there was sufficiently widespread and adequate mutual knowledge – that is, each not only knows but also knows that most others know etc. – of harmful, humanly produced, luxury carbon emissions among members of

[6] I realize that the latter claim, in particular, is disputable. However, given that those in the developed world were responsible for the lion's share of carbon emissions during this period, and given the dependence of most citizens on current institutions and technologies, it is surely plausible that reshaping institutions and developing new technologies would have been necessary. For example, return to a more primitive economic and technological system is not for modern citizens a feasible option.

the relevant populations; nor was there such mutual knowledge of the necessary institutional and technological means to reduce these emissions.

For another thing, even if the members of these populations had the necessary mutual knowledge they were not in a position themselves to implement such fundamental institutional and technological change. Here it is important to understand that while it might be feasible for each individual member of a large group to do x, it might not be feasible for all or most of the members of the group to do x; to suppose otherwise is to commit a version of the fallacy of composition. Thus while it might be possible for any singe member of a community to go on welfare, it is not possible for everyone to do so; since with everyone out of work eventually there would be no welfare funds to be dispersed. Again, while it might be feasible for one or a minority of people to immediately and simultaneously switch to alternative energy sources, it is not feasible for everyone to do so immediately and simultaneously, since an entire national – indeed, international – system of energy infrastructure based on fossil fuels cannot be replaced overnight but will take decades of well-planned and coordinated institutional redesign and technological development. I conclude that the millions are not collectively morally responsible for the harm in question, and each is certainly not fully morally responsible for that harm (including for the initial reasons given above).

It might be argued in response to this that the members of the relevant governments are collectively morally responsible for the harm in question since (within the 1990–2010 time frame) they could have acted in accordance with the collective end to avert the harm, and devised and implemented the required mitigation and adaptation measures by causing the necessary redesigning and reshaping of relevant institutions. Notwithstanding the collective action problems faced by national governments (e.g. if one nation-state substantially cuts carbon emissions and the others don't then the first will be significantly economically disadvantaged), and the pressure to maintain the status quo applied by powerful corporations (e.g. oil companies) and community interest groups (e.g. mining communities), arguably the members of these governments are collectively morally responsible for failing to put in place policies to avert or substantially ameliorate the harm done (or about to be done) by 1990–2010 luxury carbon emissions. However, the members of the governments in questions are not morally responsible for the harm itself; a few thousand politicians did not produce a quantum of luxury carbon emissions sufficient to cause the massive harm in question.

Thus far we have discussed collective moral responsibility for the harm caused by 1990–2010 luxury carbon emissions; that is we have been concerned with retrospective collective moral responsibility. It is now time to look at things prospectively and to consider the collective moral responsibility to act to avert future harms.

I take it that the potential future harms in question are catastrophic if the increasing level of global luxury carbon emissions is not slowed and then stabilized in the coming decades. Moreover, I take it that for this to occur there needs to be joint action commencing in the very near future on the part of (at least) the USA, the major European nations, Japan and the major emerging economies, especially China and India. Further, it is clear that the joint action in question will not only

need to realize the collective end (or ends) in question, it will entail an allocation of burdens, especially in respect of emission reduction targets, and the burdens should reflect the quantum of past emissions, the capacity to bear the burdens and so on (Pickering, Vanderheiden and Miller 2010). However, I will not address these issues of allocation and distributive (including intergenerational) justice here (Gardiner 2004; Vanderheiden 2008).

As we have seen, there are an array of collective epistemic moral responsibilities in relation to luxury carbon emissions that derive from the collective moral behavioral responsibility to avert the catastrophic future harms consequent upon such emissions.

As we have also seen, these collective, epistemic, moral responsibilities pertain to the means to realise a collective end, namely, averting future large-scale harms, and the means in question include a wide variety of institutional rearrangements and of new technologies.

Moreover, these multiple, collective, epistemic, moral responsibilities attach to the members of a variety of different groups. Crucially, we do not need to go beyond the ascription of moral responsibility to individual human beings. I will focus attention on three of these groups, namely, members of governments, climate scientists and citizens.

In the light of the catastrophic nature of the potential harm in question and the need for knowledge of climate change, climate scientists have an epistemic, moral responsibility to acquire the required knowledge, and members of government (and of the media) have an epistemic, moral responsibility to disseminate this knowledge to their citizens. These responsibilities are also institutional responsibilities; for example, an important acknowledged institutional responsibility of governments is to protect the lives and habitats of their citizens and, therefore, to provide to the citizenry such information as is necessary to achieve this.

However, in the light of our account of collective moral responsibility, we can now understand these responsibilities as collective moral responsibilities and not simply as aggregates of individual responsibilities or as corporate responsibilities that attach to the institutions or groups in question per se. For in the case of the climate scientists, the intellectual work in question is a joint enterprise that ultimately yields mutual knowledge of climate change, as it already to some extent has e.g. with respect to the greenhouse effect (notwithstanding, that at any point in time there will be scientific disputes). In the case of the governments, the collective responsibility in question is an institutional responsibility now to be understood as the joint moral responsibility of the members of each of these governments. Naturally, the nature and strength of these moral responsibilities is relative to the magnitude of the potential harm in question and the period of time available to avert it. But it is also a function of the extent to which these moral responsibilities have been, so to speak, institutionalized, e.g. by the creation of institutions and associated institutional roles, such as a department of climate change headed up by a minister for climate change. As our theory makes clear, collective moral responsibility and institutional responsibility interact, and give direction to, and mutually reinforce, one another.

On the plausible assumption that the reality of human-induced, harmful, luxury carbon emissions is no longer seriously disputed by the world's governments, or that if it is by some, it morally ought not to be, then an important collective moral responsibility of national governments is to see to it that emission reduction policies are designed (joint epistemic action) and implemented (joint behavioral action).[7] This is in effect a collective moral responsibility with respect to the reshaping and/or establishment of institutional arrangements (notably economic ones), i.e. with respect to multi-layered structures of joint action comprised of both joint epistemic and joint behavioral tasks.

The collective moral responsibilities of governments operate at two levels, namely, inter-governmental joint action (where the various separate governments cooperate on carbon emissions policies, e.g. emissions targets agreed (hopefully) at international gatherings such as Kyoto and Copenhagen), and intra-governmental joint action (where the members of a single government cooperate on carbon emissions, e.g. via a national carbon tax).

The collective moral (epistemic and behavioural) responsibility of governments at the intra-governmental or national level reduces to the collective moral (epistemic and behavioural) responsibility of the relevant individual members of the government in question qua member of that government. This is an instance of the collective moral responsibility that can attach to the morally significant decisions of a joint institutional mechanism (as described above).[8]

I also note that the collective moral (epistemic and behavioural) responsibility of governments at the *inter-governmental* level reduces to the collective moral (epistemic and behavioural) responsibility of the relevant individual members of each of the governments in question (in each case qua member of his or her own government). This is an instance of the collective moral responsibility that can attach to the morally significant decisions of a multi-layered structure of joint action and, indeed, a multi-layered structure of joint action that comprises a multi-layered structure of joint institutional mechanisms, one at the international level and a number of others at the national level.

---

[7] Designing a public policy (on my account) is in essence a process of finding out a complex – and hitherto unknown – means to achieve a prior, given collective end e.g. the end of reducing carbon emissions. Moreover, a policy is simply a linguistic structure of propositions describing the means to achieve that end – as such it is essentially an epistemic structure – and if the alleged means are not actually means then it has failed epistemically. (And if the policy maker is not aiming at actually finding a means but simply (say) pretending to find one then this is deception.) Accordingly, many policies are not implemented and remain only as documents in the filing cabinets of bureaucrats. So I am distinguishing between a policy and its implementation. Implementation is principally a behavioural matter. I reiterate that joint epistemic action typically involves joint behavioural action and vice-versa.

[8] And in so far as the notion of a government includes the bureaucracy which implements its policies, this is an instance of the collective moral responsibility that can attach to a multilayered structure of joint action.

Finally, let me turn briefly to the moral responsibilities of citizens, especially citizens of democracies given their greater capacity to influence the policies of their governments. Twenty years after the 1990 ICCP Report and in the wake of numerous, widespread and ongoing mass media reports on climate change and governmental climate change policies, there is mutual knowledge of the climate change issue and of the claims on the part of credible authorities that catastrophe awaits if it is not addressed. Accordingly, there is at the very least an aggregate of individual moral responsibilities on the part of citizens of each of the democracies to inform him or herself (epistemic action) in relation to carbon emissions, government carbon emissions policies and the like. Is there also a collective moral responsibility to vote for a political party (behavioral action via a joint institutional mechanism) that has rational and fair policies to address adequately the problem of human-induced, harmful carbon emission? Presumably, the answer to this is in the affirmative, given the reality and magnitude of the problem and given that the prior aggregate of individual epistemic responsibilities of citizens to inform themselves with respect to carbon emissions etc. have been adequately discharged (or even if they have not, they morally ought to have been).

There are further collective moral responsibilities of citizens that now follow, including to comply with (behavioral action) rational and fair policies on mitigation and adaptation measures in order to realize the morally significant collective ends of these measures, and also to establish conventions and social norms with respect to carbon emission reduction, e.g. to use forms of public transport that reduce per capita carbon emissions. The latter involving as it does attitudinal changes, notably in mutual knowledge, is in large part an exercise in implicit joint epistemic action (Miller 2001:chap. 2).

References

Copp, David. 2007. “The Collective Moral Autonomy Thesis.” *Journal of Social Philosophy* 38(3):369–88.

Gardiner, Stephen. 2004. “Ethics and Global Climate Change.” *Ethics* 114(3):555–600.

Gilbert, Margaret. 1992. *Social Facts*. Princeton, NJ: Princeton University Press.

Goldman, Alvin. 1999. *Knowledge in a Social World*. Oxford: Clarendon Press.

Lehrer, Keith. 1987. *Knowledge*. Oxford: Oxford University Press.

Miller, Seumas. 2001. *Social Action: A Teleological Account*. New York, NY: Cambridge University Press.

Miller, Seumas. 2006. “Collective Responsibility: An Individualist Account.” *Midwest Studies in Philosophy* 30(1):176–93.

Miller, Seumas. 2007. “Against the Moral Autonomy Thesis.” *Journal of Social Philosophy* 38(3):389–409.

Miller, Seumas. 2008. “Collective Responsibility and Information and Communication Technology.” In *Moral Philosophy and Information Technology* edited by J. van den Hoven and J. Weckert. New York, NY: Cambridge University Press. pp. 226–250.

Miller, Seumas. 2010. *The Moral Foundations of Social Institutions: A Philosophical Study*. New York, NY: Cambridge University Press.

Miller, Seumas, and Pekka Makela. 2005. “The Collectivist Approach to Collective Moral Responsibility.” *Metaphilosophy* 36(5):634–51.

Montmarquet, James. 1993. *Epistemic Virtue and Doxastic Responsibility*. Lanham, MD: Rowman and Littlefield.

Narveson, Jan. 2002. "Collective Responsibility." *Journal of Ethics* 6(2):179–98.

Pettit, Philip. 2001. *A Theory of Freedom*. London: Polity.

Pickering, Jonathan, Steve Vanderheiden, and Seumas Miller. 2010. "Ethical Issues in the United Nations Climate Negotiations: A Preliminary Analysis of Parties' Positions." Accessed April 1, 2010. www.cappe.edu.au.

Shue, Henry. 1992. "The Unavoidability of Justice." In *The International Politics of the Environment* edited by A. Hurrell and B. Kingsbury. Oxford: Oxford University Press. pp. 373–397.

Sinnott-Armstrong, Walter. 2005. "Its Not My Fault: Global Warming and Individual Moral Obligations." In *Perspectives on Climate Change* edited by W. Sinnott-Armstrong and R. Howarth. Amsterdam: Elsevier. pp. 285–299.

Vanderheiden, Steve. 2008. *Atmospheric Justice*. Oxford: Oxford University Press.

Index

N.A Vincent et al. (eds.), *Moral Responsibility*, Library of Ethics and Applied Philosophy 27, DOI 10.1007/978-94-007-1878-4,