Hedonistic Act Utilitarianism

Action Guidance and Moral Intuitions

Simon Rosenqvist

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in University Main Building, Room IV, Biskopsgatan 3, Uppsala, Friday, 3 April 2020 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Ben Eggleston (University of Kansas).

Abstract

Rosenqvist, S. 2020. Hedonistic Act Utilitarianism. Action Guidance and Moral Intuitions. 137 pp. Uppsala: Department of Philosophy, Uppsala University. ISBN 978-91-506-2808-1.

According to hedonistic act utilitarianism, an act is morally right if and only if, and because, it produces at least as much pleasure minus pain as any alternative act available to the agent. This dissertation gives a partial defense of utilitarianism against two types of objections: action guidance objections and intuitive objections.

In Chapter 1, the main themes of the dissertation are introduced. The chapter also examines questions of how to understand utilitarianism, including (a) how to best formulate the moral explanatory claim of the theory, (b) how to best interpret the phrase "pleasure minus pain," and (c) how the theory is related to act consequentialism.

The first part (Chapters 2 and 3) deals with action guidance objections to utilitarianism. Chapter 2 defines two kinds of action guidance: doxastic and evidential guidance. It is argued that utilitarianism is evidentially but not doxastically guiding for us. Chapter 3 evaluates various action guidance objections to utilitarianism. These are the objections that utilitarianism, because it is not doxastically guiding, is a bad moral theory, fails to be a moral theory, is an uninteresting and unimportant moral theory, and is a false moral theory.

The second part (Chapters 4, 5 and 6) deals with intuitive objections to utilitarianism. Chapter 4 presents three intuitive objections: Experience Machine, Transplant, and Utility Monster. Three defenses of utilitarianism are subsequently evaluated. Chapter 5 and 6 introduces two alternative defenses of utilitarianism against intuitive objections, both of which concern the role that imagination plays in thought experimentation. In Chapter 5, it is argued that we sometimes unknowingly carry out the wrong thought experiment when we direct intuitive objections against utilitarianism. In many such cases, we elicit moral intuitions that we believe give us reason to reject utilitarianism, but that in fact do not. In Chapter 6, it is argued that using the right kind of sensory imagination when we perform thought experiments will positively affect the epistemic trustworthiness of our moral intuitions. Moreover, it is suggested that doing so renders utilitarianism more plausible.

In Chapter 7, the contents of the dissertation are summarized.

Keywords: hedonistic act utilitarianism, consequentialism, action guidance, moral intuitions, hedonism, thought experiments, ignorance, imagination, morality, normativity, ethics, moral theory, normative ethics

Simon Rosenqvist, Department of Philosophy, Ethics and Social Philosophy, Box 627, Uppsala University, SE-75126 Uppsala, Sweden.

ISBN 978-91-506-2808-1
urn:nbn:se:uu:diva-404266 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-404266)

Acknowledgments

I must first thank the person who, far beyond any call to duty, commented on more drafts, excerpts, and chapters of this book than did everyone else combined: my supervisor Jens Johansson. His detailed and insightful comments on *everything* in this book has shaped my discussion and philosophical beliefs in many ways. I could not have finished this project without his help.

I am also very grateful to my supervisor Folke Tersman. Over the years, Folke has read and commented on all aspects of the chapters in this book, and has provided numerous helpful comments on the project as a whole. His calm demeanor, persistent encouragement, and constructive suggestions helped me stay on track towards finishing my studies.

Other people were also crucial in bringing this project to a close. Special thanks go to David Alm, who acted as my final seminar opponent, as well as Per Algander, Eric Carlson, Sebastian Lutz, and Andrew Reisner for their detailed comments on an early draft of the whole book. Moreover, Elinor Mason, Michael Ridge, Jakob Green Werkmäster, Henrik Andersson, Karl Bergman, Rebecca Wallbank, Olle Risberg, Victor Moberger, Caleb Perl, Andrew Moore, and Dorothee Bleisch kindly read and commented on individual chapters, helping me improve them in numerous ways. Finally, while writing I have often ventured beyond my limited area of expertise – many thanks to Andreas Stokke, Nils Franzén, Henrik Rhydén, and Nicholas Wiltsher for answering naïve questions about philosophical topics I knew nothing (and still know very little) about. Finally, my thanks to Anna Gustafsson for her help in administrative matters.

Many other people also deserve thanks for making the Department of Philosophy at Uppsala University such a wonderful place, and for the inspiring discussions that we have had over the years. Thank you, Emil Andersson, Johan Boberg, Tommaso Braida, Ekrem Çetinkaya, Tomas Ekenberg, Karl Ekendahl, Matti Eklund, Anna Folland, Susanne Gauffin, Elinor Hållén, Magnus Jedenheim-Edling, Kasper Kristensen, Guilherme Marques Pedro, Irene Martinez Marin, Patricia Mindus, Carl Montan, Jeremy Page, Jessica Pepp, Elena Prats, Sebastian Andres Reyes Molina, Pauliina Remes, Elisabeth Schellekens Dammann, Rysiek Sliwinski, Maarten Steenhagen, Hallvard Stette, Alexander Stöpfgeshoff, Maria Svedberg, Oda Tvedt, Ulrika Valdeson, and Annika Wennersten.

Some people were crucial to enabling my Ph.D. studies in the first place. Special mention goes to Jonas Bååth, with whom I spent many hours in libraries, dreaming about future doctoral studies, and to Maria Hansson, with whom I have read and discussed so much philosophy over the years. I am also grateful to the Department of Philosophy at Lund University for the years I spent there during my undergraduate and master's studies – to Wlodek Rabinowicz and Dan Egonsson for their kind tutelage, and to Frits Gåvertsson, Jakob Green Werkmäster, Henrik Andersson, Andrés Garcia, Eric Brandstedt, and Fritz-Anton Fritzon for our highly enjoyable postgraduate seminars and reading groups.

I have presented drafts of the chapters in this book at several places: at nearly a dozen meetings of the Higher Seminar of Practical Philosophy in Uppsala, at the Swedish Conference of Philosophy in 2015 and 2017, at the Higher Seminar in Practical Philosophy at Lund University in 2017, at the PG-WIP Seminar at Edinburgh University in 2016, at the Cambridge-Uppsala Graduate Workshop in 2017, at the Conference of the International Society for Utilitarian Studies in 2016 and 2018, at the Conference of the British Postgraduate Philosophical Association in 2016, at the Rocky Mountain Ethics Congress in 2017, and at the European Congress for Analytic Philosophy in 2017. My thanks for the helpful comments offered by the audiences at those occasions, which helped me improve this manuscript in various ways.

I am honored to have received a number of grants over the years, which helped me fund travels to conferences as well as costs of living. My thanks to Erik and Gurli Hultengren's Foundation, Anna Maria Lundin's Scholarship Fund, Ekmans V's Scholarship Fund, Håkansson's Travel Stipend, Helge Ax:son Johnsson's Foundation, Anders Karitz's Foundation, and Svanberg's Scholarship Fund. I am also very grateful to Edinburgh's Department of Philosophy for letting me visit them for four months in 2016.

Most of all, I owe my family for their enduring support over the past few years. Special thanks to my mother, Jane Asplund Rosenqvist, for unwavering encouragement and support, and for helping me take care of my little son Caspian during the final and most intense period of writing. And most importantly, thanks to Milena Siegrist, my partner and love, for careful and insightful comments on the final manuscript of this book, for crucial emotional support, and for our many intellectually stimulating discussions.

For Milena.

Contents

1. Introduction

1.1 About the Book

Moral theories are concerned with *normative* reality, with what we should do and why we should do it. Scientific theories, on the other hand, are concerned with *non-normative* reality, with questions of what the world is like and why it is like that. Just as we can ask the non-normative questions "Will Tom donate to Amnesty?" and "Why will Tom donate?", we can ask the corresponding normative questions "Should Tom donate money to Amnesty?" and "Why should Tom donate?". In both cases, we expect there to be intelligible and reasonable answers to our questions.

While moral theories are about normative reality, they are not about all of it: they are restricted to the *moral* part of normative reality. There are many normative questions that are not moral ones. For example, if you avoid checkmate by moving your king in a chess game, then there is a sense in which it is true that "you should move your king." Although this is a normative statement, because it is about which piece you *should* move in the game and not about which piece you *did* or *will* move, it is not a *moral* statement. Even if moving the king saves a million children from disease or death, it remains true that you should move it, because this normative claim is limited to what should be done *to win a game of chess.* Such "chess normativity" contrasts with "moral normativity," which is my concern here. A moral theory purports to answer not any normative question, such as what to do to win a game of chess, but only normative questions about what we morally should do. Prototypical moral questions include whether to give to charity, whether to forego flying on this year's vacation, and whether to vote for a particular political party.

Why should we study moral theories? First, they answer fascinating and interesting questions about an important dimension of our lives. We believe of various actions that they are morally right or wrong, and we use these judgments to navigate the world in our daily lives. But it is not obvious *why* certain acts are right or wrong. Why is it right to save a child from drowning, but wrong to let a child drown? What does murder and sexual harassment have in common that *make* such acts wrong? Is there even a common explanation for why such acts are wrong? And why are there so many exceptions to moral rules? For example, why can I harm others in self-defense, but only within reasonable limits? What determines those reasonable limits? Why are we expected to give special priority to our children privately, but not in the capacity of government officials? These questions tickle our curiosity and moral theories promise to answer them – to explain how these aspects of morality hang together.

Second, sometimes we find that two of our moral beliefs conflict, or coexist only uneasily, with each other. In these cases, we want to know whether these beliefs are compatible. If they are not compatible, we want to know which one to give up. For example, most of us are opposed to non-human animals being hurt for the purpose of entertainment, which has given rise to the common notice that "No animals were hurt in the making of this film." However, we care surprisingly little about animals being hurt to make food, even when the food is neither healthy nor nutritious. Is there a morally significant difference between these cases? Moreover, we think that it is deeply immoral to kill an unconscious child, but that it is permissible to kill an unconscious fetus – although the latter act is only permissible if we do it before a particular week in gestation. What could account for these different judgments? Nearly everyone thinks that it is best for a mortally ill and deeply suffering dog to be euthanized, but the idea that a human being in the same situation should even be allowed to take her own life voluntarily is controversial, and we find it clearly wrong to kill her without her explicit permission. If we learn more about what principles underlie moral normative reality, then we can determine whether these apparent tensions can be rationally motivated or whether we need to revise one or more of our moral beliefs.

In this book, I give a partial defense of hedonistic act utilitarianism – a moral theory – against two kinds of objections.[1] The book's topic is therefore theoretical rather than practical, as it aims to contribute to the understanding and justification of moral theories, and not to answer particular moral questions in areas such as healthcare, engineering, food production, or aid distribution. As a moral theory, hedonistic act utilitarianism says that you morally should perform an act if and only if, and because, the act produces more pleasure minus pain than any alternative act available to you. In slogan form, you should *maximize pleasure minus pain*. Utilitarianism is a unified theory of morality, according to which every moral question is determined by facts about the pleasure and pain produced by our acts.

On the face of it, utilitarianism gives us satisfying answers to the first set of questions that I mentioned earlier. Saving a child from drowning lets her experience more pleasure in the future, so utilitarianism tells you to save the child. Murder causes immediate pain and prevents future pleasure, while sexual harassment leads to pain in the form of future trauma: in typical such cases,

[1] Hedonistic act utilitarianism is defended by Blake (1926) and Tännsjö (1998). For general introductions to utilitarian and consequentialist theories, see Quinton ([1973] 1989), Scarre (1996), Shaw (1999), Mulgan (2007), Bykvist (2010), Driver (2012), and Lazari-Radek and Singer (2017). For more in-depth discussions, see Smart (1973), Carlson (1995), Goodin (1995), Bergström (1996), Feldman (1997), and Lazari-Radek and Singer (2014).

actions are available that will produce more pleasure minus pain. Most instances of violence are for similar reasons wrong according to utilitarianism, because they cause unnecessary pain. But self-defense within reasonable limits is an exception, because it prevents more pain than it causes. What constitutes reasonable limits on self-defense is precisely whether the pain that is caused is outweighed by the pain that is prevented. In addition, utilitarianism explains at least part of the moral difference between how we may act as private persons and as government officials. For example, if government officials help their families by using public funds, then few citizens will trust in the government. The result is that less suffering can be avoided, and less pleasure can be brought about – and so we can conclude that such acts are wrong according to utilitarianism. Finally, utilitarianism has the pleasing implication that there are no fundamental moral differences between people of different nationalities, religions, races, sexual orientations, or genders.

To other moral questions, utilitarianism gives more provocative answers. For example, if every episode of pleasure and pain is morally relevant, then it is difficult to justify treating human and non-human animals differently. So if we think that it is right to kill a suffering dog, then we should permit suffering humans to kill themselves as well. Moreover, even if utilitarianism tells public officials to focus on the public good, it need not tell them to focus on the good of their families as private persons. Instead, the theory may advocate giving their surplus wealth to charity. In other cases, utilitarianism is a provocative theory not because of *which* answers it gives, but because of *why* it gives them. For example, when determining whether a woman is morally permitted to perform an abortion, a utilitarian will find it irrelevant whether the fetus is a child (as pro-life activists will emphasize) or whether the woman has a right to her body (as pro-choice activists will emphasize) and will instead focus on what kind of life the fetus would have if it grew up, whether the parents would go on to have another child after an abortion, and how the woman would suffer from being forced to carry an unwanted pregnancy to term. Moreover, if utilitarianism is true, then many of our cultural and ideological preoccupations become less important, including those of sexual morality, patriotism, honor, justice, purity, equality, freedom, cleanliness, respect, and social status. From the utilitarian perspective, these things are at most *indirectly* morally relevant, in virtue of being typical causes of pleasure and pain.

However, and this is the crucial question, is hedonistic act utilitarianism correct? That is, is it a *true* moral theory? If it is not, then we need not consider its implications in particular cases such as those just mentioned. Over the chapters that follow, I discuss two kinds of objections to hedonistic act utilitarianism. First, that it is not action guiding. Second, that it conflicts with our moral intuitions. More details will follow, but allow me to give the general outline of these objections.

To begin with, *action guidance objections* are about utilitarianism's inability to guide our actions in moral decision making. Although we might be able to see the general implications of utilitarianism for morality, it is hard to see what utilitarianism recommends in any *specific* choice situation. For example, we know that global temperatures are rising and that they rise mainly because of human-related emissions of carbon dioxide, methane, and nitrous oxide. We also know that the effects from shifting temperatures on our world are significant and destructive, resulting in extreme weather events and rising sea levels. These non-normative facts are known from a combination of careful measurements, projections from current trends, and our best scientific theories. However, when we learn these climate-related facts, there remains an unanswered normative question: What action should we undertake in response to climate change? Although there *is* an answer to this question according to utilitarianism, that answer is not – in a sense – available to us, because we do not know *which* climate-related action will maximize pleasure minus pain. For utilitarianism to tell us to "maximize pleasure minus pain" in response to global warming is analogous to a professor who tells her student to "write a better paper." As advice goes, this is unhelpful.

Similarly, when we examine more closely the questions of right and wrong that I brought up earlier, it is less clear what utilitarianism prescribes. When we consider all the consequences of an act stretching into the distant future, we realize that we do not actually know the full consequences of a specific act of giving to charity, embezzling money, sexually harassing someone, killing someone, or letting a child drown. What we know are only some short-term consequences of performing these acts – some immediate pain or pleasure resulting from them. But this is not enough to determine whether these actions are right or wrong according to utilitarianism.

In general, there are several reasons to think that utilitarianism cannot guide our actions. To begin with, the theory sets no restraints on *whose* pleasures and pains count for determining what we should do. That is, future generations are as morally important as the present one, and this is true even for generations living a million years into the future. Next, massive numbers of acts are available to an agent at any given moment – we have thousands or tens of thousands of different acts to choose from at any time. Of these, utilitarianism will only recommend one or a few optimal one(s). Finally, our acts have wildly unpredictable "chaotic" effects on the distant future – such that even the most insignificant act will affect the future in highly unexpected and important ways. I discuss these issues and especially that of chaotic effects in Chapter 2. In Chapter 2, I also discuss what action guidance *is*, or at least how it can be defined for the purpose of formulating action guidance objections. I define two notions of action guidance, and I suggest that hedonistic act utilitarianism is not action guiding in one of these ways (it is not "doxastically guiding"), but that it is action guiding in another way (it is "evidentially guiding"). In Chapter 3, I evaluate a number of action guidance objections to utilitarianism. My discussion in this chapter shows that even if it is intuitively attractive to think that utilitarianism “should be action guiding” in some sense, it is not easy to construct a successful action guidance objection to the theory.

The second part of the book moves from the topic of *action guidance* objections to that of *intuitive objections* to hedonistic act utilitarianism. In the case of intuitive objections, the concern is that the utilitarian theory conflicts with strong moral intuitions. For example, if utilitarianism implies that most instances of abortion are wrong, that it is permissible to kill suffering patients against their will, and that it is obligatory to give nearly all of one’s income to charity, then the theory seems to give the wrong results – and so, the argument goes, it is false. In this context, for you to have a moral intuition is for something to *seem* to you in a particular way. For example, to many of us, it seems that early-stage abortion is permissible, that no one should be killed involuntarily, and that we are permitted to keep most of our money for the exclusive benefit of ourselves and our families. These intuitions, one could argue, give us reason to reject utilitarianism, because their content conflicts with the theory’s implications for what is right or wrong. The methodological assumption is that, just as we can test scientific theories against our visual perceptions, we can test moral theories against our moral intuitions.

In this book, I will not attempt to fully evaluate the idea that moral intuitions can be used to test moral theories – although I do consider some arguments for this view in Chapter 4. In that chapter, I also evaluate three responses to intuitive objections that have been made by utilitarians, each of which I suggest is inadequate. In Chapters 5 and 6, I propose my own preferred defenses of utilitarianism against intuitive objections, both of which draw on assumptions about how we imagine and carry out thought experiments.

The conclusions that I draw in this book vary in their implications for hedonistic act utilitarianism. In my discussion of action guidance objections in Chapter 3, I argue that none of the objections that I consider are successful – although my defense of utilitarianism against these objections is still only partial, because I do not discuss a certain kind of meta-ethical objections. My discussion of intuitive objections in Chapters 4-6 is more probative: I show how a number of intuitive objections can be met, or at least weakened, given some plausible assumptions about the role that imagination plays in thought experimentation. The book concludes that the action guidance and intuitive objections to utilitarianism are not as decisive as commonly thought.

Although the two groups of objections that I discuss are independently interesting, they are also related to each other. Action guidance objections rely on how utilitarianism gives us *no* or *inadequate* moral advice, and intuitive objections rely on how it gives the *wrong* moral advice. These objections seem, on the face of it, to require inconsistent assumptions. Surely, any difficulties with *applying* utilitarianism should also lead to difficulties in *testing* utilitarianism against our moral intuitions? As I argue in Chapter 4, this is true with the appropriate qualification: we should conclude that, at least with respect to what is right, permissible, and obligatory, we need to test utilitarianism by means of thought experiments about imagined cases, and not by considering real world cases.

This book is concerned with the first-order *normative* project of determining which moral theory is correct, and not with the second-order *meta-ethical* project of answering fundamental questions about moral thought, language, and reality. For example, I will not attempt to answer questions such as: "What does 'x is morally right' mean?"; "Is there a property of moral rightness?"; "Do acts ever have (instantiate) this property?"; "Are moral facts independent of human activity?"; and "Are we necessarily motivated by our moral judgments?". But neither will stay neutral on these issues – indeed, I assume the philosophically controversial view that there *are* moral facts and that they *can* at least in principle be known.[2] For example, I assume that there are true moral statements of the kind "you should donate a portion of your income to Amnesty" or "you should take care of your parents." As a result, I will not consider threats from wholesale nihilism ("there are no moral facts") or global moral skepticism ("nothing whatsoever can be known about morality"). The picture of morality that I assume is not the least controversial one, but one according to which the plausibility of utilitarianism becomes a live issue.

In what remains of this introductory chapter, I discuss how to more precisely understand hedonistic act utilitarianism. In Section 1.2, I formulate the theory more carefully. In Section 1.3, I discuss how to understand the moral explanatory "because" claim of the theory. In Section 1.4, I discuss different interpretations of the phrase "pleasure minus pain." Finally, in Section 1.5, I explain how and why hedonistic act utilitarianism is different from act consequentialism.

1.2 Formulating Utilitarianism

Intellectuals have accepted utilitarian and consequentialist flavored views for a long time. An early example is the Chinese philosopher Mo Tzu (470-391 BC) who opposed Confucian partialism and advocated "universal love," holding that "men should [...] love the members of other families and states in the same way that they love the members of their own family and state."[3] Later proto-utilitarians include Richard Cumberland (1631-1718), Anthony Ashley Cooper, the 3rd Earl of Shaftesbury (1671-1713), Francis Hutcheson (1694-1746), John Gay (1699-1745), David Hume (1711-1776), and William Paley

[2] For a defense of such a view, see Huemer (2008).

[3] Watson (1963), p. 9. See also Scarre (1996), pp. 27-33.

(1743-1805).[4] The first card-carrying hedonistic act utilitarian is Jeremy Bentham (1748-1832), who is known for promoting and applying the theory to legal and social institutions.[5] Other early hedonistic act utilitarians followed Bentham, most famously John Stuart Mill (1806-1873) and Henry Sidgwick (1838-1900).[6] Today the theory is accompanied by numerous and often less radical "cousin" theories – utilitarian and consequentialist theories diverging from it in one or more respects. However, one can still find well-known contemporary hedonistic act utilitarians, such as Peter Singer and Torbjörn Tännsjö.[7]

I will now give a full formulation of hedonistic act utilitarianism, which will be useful for the discussion that follows. While I earlier formulated the theory merely in terms of what we *should* do, I will here make use of a wider range of deontic terms, including "obligatory," "must," "should," "right," "permissible," "wrong," and "ought." That a term is "deontic" means simply that it is normative or prescriptive, rather than an evaluative term such as "good" or "bad." The meanings of these deontic terms are closely related: e.g., something is obligatory if and only if I should do it, something is wrong if and only if I should not do it, and something is permissible if and only if it is right. Nevertheless, I include all the terms in the formulation of hedonistic act utilitarianism for the sake of completeness.

Consider then:

> HEDONISTIC ACT UTILITARIANISM
>
> An act is morally obligatory (must/should/ought to be performed) if and only if, and because, it produces more pleasure minus pain than any alternative act available to the agent.
>
> An act is morally right (permissible) if and only if, and because, it produces at least as much pleasure minus pain as any alternative act available to the agent.
>
> An act is morally wrong (must not/should not/ought not to be performed) if and only if, and because, it produces less pleasure minus pain than some alternative act available to the agent.

I frequently refer to this theory as just "utilitarianism," unless there is a risk that it will be conflated with other utilitarian theories, in which case I always use its full name.

[4] Quinton ([1973] 1989), pp. 11-26; Driver (2014).

[5] See Bentham ([1780] 2008).

[6] Sidgwick ([1907] 1984); Mill ([1871] 2007).

[7] See Tännsjö (1998) and Lazari-Radek and Singer (2014). Notably, Singer is no longer a preference utilitarian, as he was in Singer (1993).

In what follows, I sometimes sacrifice precision for readability. To begin with, I employ only the terms “right” and “wrong” when formulating other moral theories than hedonistic act utilitarianism, while in other cases I switch between deontic terms as is convenient. I usually avoid the qualifier “morally” and I often use “maximize pleasure minus pain” as a shorthand for “producing at least as much pleasure minus pain as any alternative act available to the agent.” I also express myself imprecisely with regard to what *thing* utilitarianism is. Utilitarianism is, I assume, essentially a proposition. Propositions are the primary bearer of truth value and the object of propositional attitudes like desires and beliefs. Moreover, they are neither psychological entities like thoughts nor linguistic entities like sentences. For example, “Jag älskar godis” and “I love candy” both express the same proposition: the proposition that *I love candy*. On the standard view, propositions exist timelessly and independently of human thought, language, and action. For this reason, I doubt that utilitarianism can be invented, formulated, revised, stated, put forward, thought, or written down – all commonly used phrases in philosophical discussions of moral theories. In addition, because utilitarianism is a proposition, I suspect it is not correctly described as a view, perspective, position, claim, or idea. For the sake of convenience, however, I employ the conventional way of talking about moral theories and hope that the intended meaning is clear enough.

The theory that I call “hedonistic act utilitarianism” is only one of several hedonistic act utilitarian theories. Importantly, it is an *actual utility* hedonistic act utilitarian theory, which must be distinguished from an *expected utility* hedonistic act utilitarian theory, such as:

> EXPECTED UTILITY HEDONISTIC ACT UTILITARIANISM
>
> An act is morally right if and only if, and because, it produces at least as much *expected* pleasure minus pain as any alternative act available to the agent.
>
> An act is morally wrong if and only if, and because, it produces less *expected* pleasure minus pain than some alternative act available to the agent.

The expected pleasure minus pain of an act (its “expected utility”) can be calculated by following five steps.[8] First, list every possible outcome of the act. Second, assign to each outcome the amount of pleasure minus pain that you

[8] Cf. Bykvist (2010), pp. 85-87.

believe the outcome contains.[9] Third, list for every outcome your level of confidence – the subjective probability – that the act, if performed, will bring about this outcome. Fourth, multiply the assigned value of each outcome with its assigned subjective probability. Fifth, sum these products together.

To see how expected utility hedonistic act utilitarianism compares to actual utility hedonistic act utilitarianism (i.e., to the theory that I refer to simply as hedonistic act utilitarianism), we may consider the following example. You must choose between giving candy or a toy to a young child, Fiona, on her birthday. The potential outcomes, together with the assigned amounts of pleasure minus pain and levels of confidence, are listed in Table 1:

Table 1. Example of calculating expected pleasure minus pain

	GIVE HER THE CANDY		GIVE HER THE TOY	
	Outcome A1	Outcome A2	Outcome B1	Outcome B2
Pleasure minus pain contained in each outcome	10 – *Fiona is happy*	-5 – *Fiona is somewhat unhappy*	15 – *Fiona is very happy*	-30 – *Fiona is miserable*
Subjective probabilities that the act will result in the outcome	.5	.5	.7	.3
Expected pleasure minus pain	(.5*10) + (.5*-5) = 2.5		(.7*15) + (.3*-30) = 1.5	

Let us stipulate that if you give Fiona the toy, this results in outcome B1 (i.e., she becomes very happy) and that if you give her the candy, this results in outcome A2 (i.e., she becomes somewhat unhappy). It follows that to give Fiona the toy is right according to actual utility hedonistic act utilitarianism, since this will maximize actual pleasure minus pain. In contrast, giving Fiona the toy is wrong according to expected utility hedonistic act utilitarianism, since this will produce less expected pleasure minus pain than giving her the candy.

Why should we go for the actual utility version rather than the expected utility version of hedonistic act utilitarianism? The expected utility version might be thought to be more practicable and plausible, so we might expect that accepting this view, rather than the actual utility one, will let us respond more effectively to action guidance and intuitive objections. However, the

[9] Various alternative formulations are possible here, such as considering the amount you *justifiably believe* it contains, or that you *know* it contains.

case for the expected view is generally overstated. To begin with, the actual utility version has more plausible implications in a range of cases, including the above one: intuitively, of course, we should make Fiona very happy rather than somewhat unhappy. Moreover, as has been noted in the literature, the expected utility version has no obvious advantages over the actual utility version with respect to being action guiding, even though it is usually marketed as the more practicable approach.[10] Indeed, the expected utility version is at least as difficult to apply as the actual utility one. For example, applying the expected utility version of hedonistic act utilitarianism involves listing all the potential outcomes of all alternative acts, a seemingly impossible task. It also requires us to assign to each outcome a subjective probability, something that we may not have access to, or fail to be in a position to correctly ascertain. Moreover, it tells us to assign to each outcome an associated amount of pleasure minus pain, something that we may have no beliefs about.

Apart from actual utility and expected utility act utilitarian theories, there are other hedonistic act utilitarian theories that we could adopt as well. For example, we might hold a *satisficing* rather than *maximizing* view; that is, we might hold that acts are right if and only if, and because, they produce at least up to a threshold level of pleasure minus pain, where this threshold is less than the maximal amount that can be produced.[11] Moreover, some utilitarian theories reject either the hedonistic focus on pleasure and pain, or the focus on the evaluation of acts, and therefore are not hedonistic act utilitarian theories.[12] I will not evaluate these various views in this book here, but will simply focus on the maximizing actual utility form of hedonistic act utilitarianism. That said, much of what I say should be relevant to the evaluation of other utilitarian and consequentialist theories as well.

1.3 Moral Explanation

Hedonistic act utilitarianism tells us not only *which* acts are right or wrong, but also *why* they are right or wrong. That is, utilitarianism gives us a *moral explanation* of rightness and wrongness. In this section, I contrast hedonistic act utilitarianism with a theory that I call *exclusivist hedonistic act utilitarianism*. These theories differ in how the moral explanatory claim is spelled out. While the explanatory claim of ordinary hedonistic act utilitarianism is formulated merely as “because,” the explanatory claim of exclusivist hedonistic act utilitarianism is formulated as “because and only because.” In the course

[10] For discussion of the impracticality of expected utility versions of utilitarianism, see Feldman (2006).
[11] See Slote and Pettit (1984). For discussion of satisficing consequentialist views, see Carlson (1995), pp. 13-19, and Bradley (2006).
[12] For example, see preference utilitarian views such as that advocated by Hare (1981), and rule utilitarian views such as those advocated by Brandt (1963), Barrow (1991), and Hooker (2000).

of presenting this alternative "exclusivist" theory, I will also argue that it is less promising than ordinary hedonistic act utilitarianism.

At the outset, two clarifications are in order. First, the claim that pleasure minus pain maximization *explains* why an act is right is stronger than the claim that it *implies* that an act is right. The difference between implication and explanation is the difference between a theory telling us merely which acts are right or wrong, and the theory also telling us why they are right or wrong. For implication, it is enough that every time an act has the property of maximizing pleasure minus pain, then it also has the property of being right. For explanation, we need an additional component: that the act has the property of maximizing pleasure minus pain must *explain why* the act has the property of being right. To give some non-moral examples of the difference between implication and explanation, note that "*S* knows that *p*" implies that *p* is true, but that it does not explain *why p* is true. Similarly, that a creature is human implies that it has human DNA, but it does not explain *why* it has human DNA.

Second, note that the term "because" in the formulation of a moral theory can refer to either of two explanatory relations.[13] The first is an *epistemic* relation, which is about *what makes sense* of an act being right or wrong. The epistemic relation is about providing us with information, giving reasons, making evidence available, or furthering our understanding. Moreover, the epistemic relation is person-relative: something always and only makes sense of something *for* a particular person. For example, that an act of becoming a vegan maximizes pleasure minus pain makes sense of it being right *for me*, but perhaps not *for you*.

In addition to the epistemic relation, the term "because" can refer to a *metaphysical* relation. The metaphysical relation is about what *makes* an act right or wrong. The metaphysical relation is related to grounding relations, such as how atoms being arranged in a particular way makes it the case that there exists a table; and to causal relations, such as how pressing the gas pedal makes it the case that the car moves. Unlike the epistemic relation, the metaphysical relation is not person-relative. For example, if it is right for you to become a vegan because (in the metaphysical sense) becoming a vegan maximizes pleasure minus pain, then this is not so only relative to me but not you: instead, the relation holds in an "absolute" sense. In general, the epistemic relation appears to be more "subjective," focusing on our relation to the world; while the metaphysical relation seems to be more "objective," focusing on what the world is like.

In what follows, I choose to understand the term "because" of hedonistic act utilitarianism as referring solely to the metaphysical relation, and not to the epistemic relation. There are two reasons for this choice. To begin with, the metaphysical relation is more interesting than the epistemic relation, as it

[13] Cf. Rydéhn (2019), pp. 16-17; Fogal and Risberg (Forthcoming), pp. 5-6.

is about normative reality itself rather than our understanding or access to it. In addition, the metaphysical relation is indirectly epistemically significant, since it will often be because a person believes that pleasure minus pain maximization *makes* an act right that pleasure minus pain maximization *makes sense* of it being right for her. So the metaphysical relation seems, in some sense, more central and fundamental than the epistemic relation.

(At this point, you may object that explanations are always about the epistemic relation and never about the metaphysical one, simply because the term "explanation" suggests something epistemic in nature. Such a dispute appears merely terminological, however. If you have this complaint, I invite you to think of a suitable different phrase to pick out the metaphysical relation, such as "making relation" or "determining relation.")

Let me now introduce the theory that I mentioned earlier: exclusivist hedonistic act utilitarianism. First, note that when two moral theories are formulated in terms of "because," the given explanations are sometimes compatible with each other. In these cases, we can have *explanatory overdetermination*: namely two compatible full explanations of the rightness or wrongness of an act. This is not surprising, as it is similar to how explanations work in other domains. We can look at, for example, overdetermination with respect to causal explanations. Suppose that you and I set fire to a forest, each in a different place but at the same time. Since it is a dry and warm summer, each of our acts is *sufficient* to cause the forest to burn down, so each of our acts *fully* causes the forest to burn down. That the explanation is full – and not merely partial – is evidenced by how, to give a full account of why the forest burned down, it suffices to tell about one of our pyromaniac acts: you do not need to cite both acts in the explanation. Since it is explanatorily sufficient to note that one of us set fire to the forest, each act explanatorily overdetermines that the forest burns down. Similarly, consider grounding explanations. That a tree has a trunk, branches, leaves, and so on fully makes it a tree: it is a tree fully in virtue of it having these things. But that a tree has a large number of atoms arranged tree-wise *also* fully makes it a tree: it is a tree fully in virtue of it having its atoms arranged tree-wise. So the fact that a tree has branches, trunk, and leaves *and* the fact that the tree has atoms arranged tree-wise each serves to fully make it a tree. Here, too, we have a case of explanatory overdetermination.

Importantly for my discussion, just as there can be two full causal explanations for why the forest burns down, and there can be two full grounding explanations for why something is a tree, there can also be two full moral explanations for why an act is right or wrong. To see this, let us briefly consider the following non-utilitarian moral theory:

> THE DIVINE COMMAND THEORY
>
> An act is morally right if and only if, and because, God allows it.

> An act is morally wrong if and only if, and because, God forbids it.

Suppose for the sake of the argument that God exists and that She is a hedonistic act utilitarian. Consequently, God always and only allows us to perform acts that maximize pleasure minus pain. It follows that utilitarianism and the divine command theory have the same implications for action. The theories are, to use the appropriate technical terminology, *necessarily extensionally equivalent* with respect to rightness and wrongness. That two theories are necessarily extensionally equivalent means that they have the same implications for rightness and wrongness in every metaphysically possible world (they have these implications by "metaphysical necessity"). As a result, necessarily, when and only when an act maximizes pleasure minus pain, God allows it; and when and only when an act does not maximize pleasure minus pain, God forbids it. Importantly, although hedonistic act utilitarianism and the divine command theory are under these assumptions extensionally equivalent, they will still give different *explanations* of rightness and wrongness. Utilitarianism gives its explanation in terms of maximizing pleasure minus pain, and the divine command theory gives us its explanation in terms of the commands of God. For this reason, one of these theories could be true even if the other is false – because one explanation could be correct, while the other is incorrect. But it could also turn out that the theories are *both true*, because it could be true that acts are right *both* because they maximize pleasure minus pain *and* because they are allowed by God. In such a case, we have an example of moral explanatory overdetermination: we have two compatible and full moral explanations of an act's rightness. This point may come as a surprise to some normative ethicists, since we are accustomed to thinking that necessarily extensionally equivalent moral theories will typically rule each other out by giving different moral explanations of rightness or wrongness. However, as the above case demonstrates, different full moral explanations can at least in principle co-exist. This possibility of moral explanatory overdetermination raises questions about how to best formulate utilitarianism, to which I now turn.

Consider how I formulated hedonistic act utilitarianism earlier:

> An act is morally right if and only if, and because, it produces at least as much pleasure minus pain as any alternative act available to the agent.
>
> An act is morally wrong if and only if, and because, it produces less pleasure minus pain than some alternative act available to the agent.

As I just noted, hedonistic act utilitarianism is in principle compatible with the divine command theory. But if we want to, we can produce a theory that is *even in principle* incompatible with the divine command theory. We only need to state the explanatory claim as "because and only because" instead of as "because." This view is attractive for those who want their utilitarian view to stand in clear opposition to non-utilitarian moral theories, such as those who want to maintain that utilitarianism and the divine command theory rule each other out even in principle. Consider then the following moral theory:

> EXCLUSIVIST HEDONISTIC ACT UTILITARIANISM
>
> An act is morally right if and only if, and *because and only because*, it produces at least as much pleasure minus pain as any alternative act available to the agent.
>
> An act is morally wrong if and only if, and *because and only because*, it produces less pleasure minus pain than some alternative act available to the agent.

If exclusivist hedonistic act utilitarianism is true, then ordinary hedonistic act utilitarianism is true as well, as "because and only because" implies "because." However, if exclusivist hedonistic act utilitarianism is true, then the divine command theory is false, and vice versa, as these two theories are incompatible.

Should someone who is sympathetic to utilitarian views want to defend exclusivist hedonistic act utilitarianism, rather than ordinary hedonistic act utilitarianism? I do not think so, because even if the exclusivist theory makes sense of how we normally view moral theories (i.e., as clear-cut competitors or alternatives to each other), it is significantly less plausible than ordinary hedonistic act utilitarianism. The problem is that the exclusivist view excludes too many explanations from being true. For example, it prevents not only the divine command theory, but also the following two views from being true:

(1) An act is morally right if and only if, and because, it produces at least as much happiness as any alternative act available to the agent.

(2) An act is morally right if and only if, and because, it produces at least as good consequences as any alternative act available to the agent.

Why does exclusivist utilitarianism prevent (1) and (2) from being true? Recall that according to exclusivist utilitarianism, acts are right *because and only because* they maximize pleasure minus pain. It follows that acts are *never* right because they produce at least as much happiness, or as good consequences, as any other act. But if exclusivist hedonistic act utilitarianism is incompatible with (1) and (2) it is highly implausible. Surely, *if* acts are right because they maximize pleasure minus pain, *then* (1) and (2) are also true. Even many non-consequentialists will agree with this conditional claim, even as they reject the antecedent.[14] At the very least, the possibility of (1) and (2) being true should be left open by the utilitarian theory, even if it is not implied by it – but exclusivist hedonistic act utilitarianism unacceptably rules them out from the start.

Here is another problem for exclusivist hedonistic act utilitarianism. Let B refer to those brain states that give rise to pleasures and pains, whatever these states are. Clearly, there are some such brain states, even if we do not know which they are or how to individuate them. Next, consider the claim that:

(3) An act maximizes pleasure minus pain because it maximizes the occurrence of B.

It seems plausible to think that there is *some* interpretation of B which makes (3) come out as true. Let us assume this interpretation. As we have seen, both ordinary and exclusivist hedonistic act utilitarianism agree that:

(4) An act is right because it maximizes pleasure minus pain.

However, if (3) and (4) are true, then it is attractive to conclude that:

(5) An act is right because it maximizes the occurrence of B.

The problem is that on the exclusivist view, acts are *never* right because they maximize the occurrence of B, since acts are always and only right because they maximize pleasure minus pain. As a result, the exclusivist hedonistic act utilitarian is faced with a dilemma: she must either reject the move from claims (3) and (4) to claim (5), or she must reject claim (3). As claim (3) seems very plausible, this leaves her with rejecting the move from claims (3) and (4) to claim (5).

To begin with, the exclusivist utilitarian could claim that the above argument presupposes that the because-relation is *transitive*, and then attempt to argue that this relation is not in fact transitive.[15] A simple example of a transitive relation is "taller than." If Jordan is taller than Pete, and Pete is taller than Amanda, then Jordan must also be taller than Amanda. Similarly, if the because-relation is transitive, then if x is A because x is B, and x is B because x is C, it follows that x is A because x is C. Therefore, if the because-relation is transitive, then (5) follows from (3) and (4). But, the exclusivist utilitarian could argue, we should reject such a claim of transitivity. For one thing, a very strong kind of transitivity would be needed, since it would need to operate

[14] Some would disagree. Consider, for example, Haybron (2010), pp. 63-77, who denies that happiness is a simple function of pleasure minus pain; and Barrow (1991), pp. 65-90, who defends an account of happiness as a state of contentment, or of being at one with the world. See also the discussion of Haybron by Lazari-Radek and Singer (2014), pp. 249-252.

[15]

over several different because-relations – note that the because-relation involved in (3) is different than those involved in (4) and (5), with the latter (but not the former) being moral metaphysical because-relations. Since transitivity does not hold, the move from (3) and (4) to (5) fails.

However, the above defense of the exclusivist view does not work, as nothing in my argument *presupposes* transitivity. For sure, transitivity would be good news for my argument against the exclusivist. But even if transitivity does not hold, (5) remains *plausible* or *reasonable* in the light of (3) and (4). And that is all I need for my argument against the exclusivist hedonistic act utilitarian to go through.

At this point in the argument, the exclusivist hedonistic act utilitarian may try to deny that the move from (3) and (4) to (5) is in fact reasonable. She could say that even if acts are right because they maximize pleasure minus pain, and even if they maximize pleasure minus pain because they maximize the occurrence of B, it is not thereby reasonable to think that they are right because they maximize the occurrence of B. In support of this claim, she may point out that it sounds odd or strange to say that an act is right because it maximizes the occurrence of a certain type of brain state – "what does facts about our frontal lobe have to do with rightness?", she could ask. However, here the distinction between metaphysical and epistemic because-relations becomes important. I agree that acts maximizing the occurrence of B *makes little sense,* given our lack of knowledge about B, of why they are right. However, it seems implausible to also deny that the metaphysical relation holds – that the acts are *made* right by the occurrence of B; that they *depend* on the occurrence of B for being right.

Finally, an important reason to favor the ordinary version of hedonistic act utilitarianism over the exclusivist one is that the former lets us accept non-utilitarian "local" explanations for rightness and wrongness without having to give up our utilitarian theory. For example, as long as an act maximizes pleasure minus pain, hedonistic act utilitarianism does not rule out that it is right because it is respectful, honors one's ancestors, or exemplifies a good character.

1.4 Pleasure minus Pain

Hedonistic act utilitarianism explains rightness and wrongness in terms of the *pleasure minus pain* produced by acts. This raises the issue of how to best understand the phrase "pleasure minus pain" when formulating hedonistic act utilitarianism. When answering this question, we need not give real definitions of "pleasure," "pain," and "pleasure minus pain." It is enough to find an interpretation (even if it turns out to be a partially stipulative one) on which hedonistic act utilitarianism becomes as plausible as possible, while it remains recognizably hedonistic in character. Moreover, we need not discuss which theory of hedonism is most plausible – be it a theory of moral value, prudential value, the good life, happiness, well-being, or welfare. Ultimately, what is valuable or what constitutes a good life need not be what our favorite hedonistic act utilitarian theory tells us to maximize. Accordingly, in the following discussion I bracket both the question of how to define pleasure and pain, and the question of which hedonistic theories are most plausible.

Consider first an interpretation of "pleasure minus pain" in terms of *pleasure and pain sensations*:

> THE SIMPLE INTERPRETATION
>
> The pleasure minus pain in an episode E is equal to the duration times the intensity of all the pleasure sensations in E minus the duration times the intensity of all the pain sensations in E.

Both the *intensity* and the *duration* of pleasure and pain sensations are represented in the simple interpretation. This ensures that a second of torture comes out as worse than a second of slight headache (the intensity matters), and that an hour of continuous enjoyment comes out as better than a minute of continuous enjoyment (the duration matters). The term "episode" is not meant to carry any particular significance in this context, but is used only as a catch-all term applicable for a period of whatever we are interested in.

The problem with the simple interpretation of "pleasure minus pain" is that some sensations are hedonically relevant, even as they fail to qualify as pleasure or pain sensations.[16] For example, sensations of calm and intense concentration are not typically thought to be pleasure sensations, unlike sensations of eating good food or having sex. Similarly, sensations of anxiety and restlessness are not typically thought to be pain sensations, unlike sensations of being cut with a sharp object or of having a headache. Just as it would be strange for someone who is calm to say that she thereby feels pleasure, it would be strange for someone who is restless to say that she thereby feels pain. Nonetheless, all of the above sensations are still hedonically relevant in some way, and we clearly want to include them when formulating our theory of hedonistic act utilitarianism. So the simple interpretation is too restrictive with respect to *which* sensations count.

Instead of interpreting pleasure minus pain as *sensory* pleasure minus pain, we can interpret it as *attitudinal* pleasure minus pain.[17] Attitudinal pleasures and pains are not sensations, but propositional attitudes. Propositional attitudes are intentional mental states that take propositions as their objects. As

[16] Cf. Labukt (2012), pp. 173-174.

[17] For a discussion of attitudinal pleasures and pains, see Feldman (2004), pp. 55-78. For more recent work on his attitudinal hedonistic theory, see Feldman (2019).

such they have a built in "directedness" towards these propositions, or perhaps to what propositions are about. Attitudinal pleasures and pains are thereby similar to other propositional attitudes such as beliefs ("I believe *that* I own a car") and desires ("I desire *that* I have more money"). As an example of an attitudinal pleasure, consider your taking pleasure in the proposition "I will sleep for two more hours the next morning." Moreover, as an example of an attitudinal pain, consider your taking pain in the proposition "I will work during my upcoming vacation." In ordinary conversation, we express having attitudinal pleasure or pain by phrases such as "I enjoy working out" and "the suffering of innocent animals pains me." As should be clear, pleasure and pain *attitudes* are different from pleasure and pain *sensations* in several ways. Most importantly, a sensation has no object or built-in directedness. Although you can feel the pleasure sensation of tasting chocolate, and you can feel the pain sensation of a headache, you do not thereby feel these sensations *in* something, and they are never directed *at* anything. Moreover, attitudinal and sensory pleasures and pains need not go together. For example, you can take attitudinal pleasure in your having sensory pain, like how a masochist takes attitudinal pleasure in his having a painful sensation.

By referring to attitudinal instead of sensory pleasure and pain, we get the following interpretation of "pleasure minus pain":

> THE ATTITUDINAL INTERPRETATION
>
> The pleasure minus pain in an episode E is equal to the duration times the strength of the attitudinal pleasures in E minus the duration times the strength of the attitudinal pains in E.

The attitudinal interpretation has an important advantage over the simple sensory interpretation: it lets us assign positive moral weight to sensations of calm and intense concentration, and lets us assign negative moral weight to sensations of anxiety and restlessness, as long as we take pleasure and pain in having these sensations. However, the attitudinal interpretation is also unacceptably restrictive, although in a different way: it demands too much in the way of advanced cognitive capacities.[18] For example, suppose that we learn that creatures like garden snails have sensory pleasures and pains, but that they lack the capacity to form and sustain propositional attitudes.[19] Surely, we still want our utilitarian theory to give moral weight to these creatures. But that is

[18] Cf. Aydede (2014), pp. 125-126.

[19] According to Fred Feldman, this is impossible, because sensory pleasures are those and only those sensations that we take attitudinal pleasure in having. See Feldman (2004), pp. 57, 79-81. However, I do not think Feldman's view can be correct: it is simply implausible that a creature cannot have pleasure sensations just because it cannot take pleasure in sensations.

not possible if we formulate hedonistic act utilitarianism according to the attitudinal interpretation of "pleasure minus pain," as these creatures will lack the cognitive capacity required for attitudinal pleasure and pain.

We face a similar problem if we try to single out sensations not by the criterion of their being pleasure or pain sensations, but by the criterion of their being *liked* or *disliked* sensations.[20] Such a view shares the advantages of the attitudinal interpretation, as we like having the sensations of having sex and feeling calm, and we dislike having the sensations of being cut by a sharp object and of being anxious. But again, some creatures may be unable to like or dislike anything, even if they can have pleasure and pain sensations, yet intuitively should still count for moral purposes. Moreover, while the attitudinal proposal is clearly hedonistic in character, since it is about pleasure and pain, it is less clear that the liking/disliking proposal is so. It seems more "preferentialist" than hedonistic, since it is our liking or disliking certain sensations that matters fundamentally for the purpose of moral evaluation.

Pleasure and pain sensations and pleasure and pain attitudes are only two potential hedonic dimensions that the hedonistic act utilitarian can appeal to. In addition, she can also look to the degree to which sensation are *pleasant and unpleasant*.[21] Consider then:

> THE PLEASANTNESS INTERPRETATION
>
> The pleasure minus pain in an episode E is equal to the duration times the intensity of the pleasantness of all sensations in E minus the duration times the intensity of the unpleasantness of all sensations in E.

The pleasantness interpretation is superior to both the simple and attitudinal interpretations. First, sensations of calm and intense concentration are pleasant, and sensations of anxiety and restlessness are unpleasant, so all these sensations are properly counted. Second, no advanced cognitive capacities are required from creatures, so their sensations of pleasantness and unpleasantness are counted as well. Third, the interpretation is properly hedonistic, being fundamentally about pleasantness and unpleasantness, rather than for example liking or desiring.

Some comments are in order at this point. To begin with, consider the difference between the simple and the pleasantness interpretations of "pleasure minus pain." For a sensation to be a pleasure its degree of pleasantness must cross a *threshold level* – this, I would argue, is why sensations of calm and intense concentration can be pleasant even if they do not count as pleasure sensations, and so as *pleasures*. That is, just like to be tall requires crossing a contextually defined threshold of tallness, being a pleasure requires crossing

[20] Cf. Kagan (1992), pp. 173-174; Aydede (2014), pp. 128-129.

[21] Cf. Aydede (2014).

a contextually defined threshold of pleasantness. In contrast, for a sensation to be a pain, it must be *paradigmatically* unpleasant, rather than to cross a threshold level of unpleasantness. This, I would argue, is why anxiety and restlessness do not count as pain sensations, and so as pains, while cuts from sharp objects and headaches do.

Next, note that the pleasantness interpretation of "pleasure minus pain" is inspired by hedonic tone views of pleasure and pain – these are views focusing on the degree to which sensations feel good or bad.[22] However, these views are typically about what pleasure and pain *are*, while I am mainly interested in how to best interpret utilitarianism. And as should be clear from my above discussion, I do not endorse the view that pleasures and pains can be understood simply as those sensations which have an overall positive or negative hedonic tone.[23] My view is that there are at least three different hedonically relevant phenomena in play here, none of which are in any obvious ways reducible to each other: there are pleasure and pain sensations, pleasure and pain attitudes, and the degree to which sensations are pleasant and unpleasant.

Finally, one could complain that the resulting utilitarian view is strictly speaking no longer about "pleasure minus pain," but rather about "pleasantness minus unpleasantness." That is true, but I suspect that most hedonistic act utilitarians have always had in mind something like "pleasantness minus unpleasantness." Moreover, reference to "pleasure minus pain" in the formulation of hedonistic act utilitarianism is so commonplace that it is easier to simply stipulate that "pleasure minus pain" in this context refers to pleasantness minus unpleasantness, than to try to change the formulation itself.

How does adopting the pleasantness interpretation matter for the plausibility of hedonistic act utilitarianism? First, note that hedonistic act utilitarians have always been faced with a stubborn group of intuitive objections: objections to the effect that their theory prescribes *debauchery* – enjoying an endless series of food, drugs, and sex. Such objections are nowadays less striking than in the 19th century, but even many contemporaries will feel that pleasures do not represent all that is good in life. When we adopt the pleasantness interpretation of pleasure minus pain, we get a more nuanced and reasonable view: we see that many sensations besides pleasures count as well. On the pleasantness interpretation of pleasure minus pain, a life of meditation and calm can contain as much pleasure minus pain as a life filled with good food and sex; and anxiety might constitute a greater problem than physical pain. Second, we can explain why "tiring" on certain pleasures makes these sensations less valuable to the utilitarian, without having to argue that the sensations are less

[22] See e.g. Kagan (1992), pp. 172-173; Smuts (2011), pp. 254-256; Labukt (2012). My account is similar to that of Tännsjö (2007), pp. 81-87, although he focuses on the hedonic tone of a person's total experiential state rather than the pleasantness and unpleasantness of her sensations.

[23] But in so far as hedonic tone views are about how to define *pleasantness and unpleasantness*, and not about how to define *pleasures and pains*, I have no objections against them.

intense. On the pleasantness view, we can simply say that when we have too much good food, we still have the same gastronomical pleasure sensations of the same intensity, but these sensations are now less pleasant to us. Third and finally, the resulting utilitarian theory avoids having to explain the value of things like liberty and equality solely in terms of their propensity to cause pleasures: the utilitarian can note that liberty and equality give rise to sensations of *being free* and of *being treated as an equal* – sensations that can be pleasant, even if they are not intense enough to count as pleasures. To sum up, when we adopt the pleasantness interpretation of hedonistic act utilitarianism, we have available to us a more extensive toolbox for defending our theory against intuitive objections.

1.5 Act Consequentialism

Several of the objections and arguments that I discuss in this book could be rephrased as about a different moral theory, namely:

> ACT CONSEQUENTIALISM
>
> An act is right if and only if, and because, it produces at least as good consequences as any alternative act available to the agent.
>
> An act is wrong if and only if, and because, it produces less good consequences than some alternative act available to the agent.

If you want, you can read this book as a book about act consequentialism instead of hedonistic act utilitarianism, as many of the arguments that I discuss will be equally relevant to evaluating act consequentialism. But even if my choice to focus on hedonistic act utilitarianism makes only a limited practical difference, you might still find it unmotivated or odd. Perhaps this is because over the past few decades there has been a move from talking about *utilitarian* theories to talking about *consequentialist* ones, where all utilitarian theories are consequentialist ones.[24] Would it not be more natural for a book of this kind to follow this trend and discuss act consequentialism instead of hedonistic act utilitarianism?

Let me first point out a common misconception: some believe that hedonistic act utilitarianism *implies* act consequentialism, so that whenever the former theory is true, the latter is true as well. This, one may think, gives us

[24] To give some examples of this trend, note some titles of recent work in the area: *The Demands of Consequentialism* by Tim Mulgan (2001); *Beyond Consequentialism* by Paul Hurley (2011); *Consequentialism* by Julia Driver (2012); *The Dimensions of Consequentialism* by Peterson (2013); *Commonsense Consequentialism* by Douglas Portmore (2014); and *The Case Against Consequentialism Reconsidered* by Nikil Mukerji (2016).

reason to focus on act consequentialism instead of hedonistic act utilitarianism in a study of this kind. But even if hedonistic act utilitarianism *did* imply act consequentialism, that would not necessarily give us more reason to focus on the latter. An advantage that hedonistic act utilitarianism has over act consequentialism is that it is a more informative theory: the latter view is in principle open to all sorts of consequences being good, beyond pleasure minus pain. With respect to hedonistic act utilitarianism, it is therefore especially easy to see why it cannot guide our actions, and how exactly it conflicts with our moral intuitions.

More importantly, it is simply not true that hedonistic act utilitarianism implies act consequentialism. To begin with, if there is no moral goodness or moral value in the world, then act consequentialism will tell us that *every* act is right, since every act produces at least as good consequences as any alternative available to the agent. In that case, however, the theory is clearly false, since at least some acts are wrong. In contrast, hedonistic act utilitarianism will in such a situation have different and more plausible implications than act consequentialism: because even if nothing is morally good or valuable, some acts will still produce less pleasure minus pain than others, and so be wrong. In addition, hedonistic act utilitarianism and act consequentialism give different explanations of *why* acts are right – one appeals to the goodness of the consequences being produced by an act, and the other appeals to the pleasure minus pain being produced by an act. We might find either of these explanations more plausible than the other – indeed, hedonistic act utilitarianism might be true (because the explanation that it gives is correct) even if act consequentialism is false (because the explanation that it gives is incorrect). Of course, we could argue in response that goodness *is identical to* pleasure minus pain, and so that these two theories actually give the same explanation. But this reply raises the kind of worries that over the past decades have been directed, to devastating effect, against naturalist theories in meta-ethics, arguments that I will not rehearse here.

A consequence of the above claims is that hedonistic act utilitarianism might be more plausible than act consequentialism. For my own part, I think that hedonistic act utilitarianism *is* more plausible. Although I consider myself both a hedonistic act utilitarian and an act consequentialist, I am more certain that the former view is correct. Pleasure and pain are obvious candidates for explaining the rightness and wrongness of acts and much more so than are properties such as goodness and badness. For example, if I learn tomorrow that pleasure is bad and that pain is good, I will nevertheless be horrified to learn that act consequentialists want to maximize pain minus pleasure. I will not be consoled to learn from them that the pain is good, and that the pleasure is bad. If a person tries to maximize pain minus pleasure in such circumstances, that person needs to be stopped, and whether the pain and pleasure are good, bad, or neutral does not matter. For the above reasons, I hesitate to attempt a defense of act consequentialism rather than of hedonistic act utilitarianism. Because the latter view is both more informative and more plausible, it makes sense to make it the centerpiece of the present inquiry.

As a final remark, it is important to distinguish between *theories* and *categories of theories*. The phrase "act consequentialist theories" refers *not* to a moral theory, such as act consequentialism, but instead to a *category* of theories. Both hedonistic act utilitarianism and act consequentialism are act consequentialist theories: they belong to this category. But they are still *different* theories, since they give different explanations and have potentially different implications. So what exactly do act consequentialist theories have in common, such that hedonistic act utilitarianism and act consequentialism both qualify as act consequentialist theories? Presumably, that they both explain rightness and wrongness solely in terms of the consequences of acts. What distinguishes them is that one theory does so in terms of the pleasure minus pain of the consequences, and the other theory does so in terms of the goodness of the consequences. Again, the phrase "consequentialist theories" refers to a category of theories, and categories are clearly not theories, just as the category of apples is not itself an apple. Indeed, a category *cannot* be a theory – since theories are either propositions or sentences, and categories are neither propositions nor sentences.

1.6 The Plan of the Book

I hope that at this point it is reasonably clear how to understand the theory of hedonistic act utilitarianism. In the chapters that follow, I discuss the two groups of objections to utilitarianism that I outlined earlier: the action guidance objections and the intuitive objections.

The first part of the book (Chapters 2 and 3) is devoted to action guidance objections. In Chapter 2, I examine how to define action guidance. In the first part of the chapter, I suggest two ways in which a moral theory can be action guiding: it can be evidentially action guiding and it can be doxastically action guiding. In the second part of the chapter, I consider in which of these two ways that utilitarianism is and is not action guiding: I show that utilitarianism is evidentially guiding, but that it is not doxastically guiding. In Chapter 3, I move on to discuss action guidance objections to utilitarianism. These objections differ from each other in their central premises, which draw on claims about our ability to gain knowledge, the principle that "ought implies can," the function of moral theories, and more. The objections also differ in their conclusions – that utilitarianism is a bad moral theory, not a moral theory at all, an uninteresting or unimportant moral theory, or a false moral theory.

The second part of the book (Chapters 4, 5, and 6) focuses on intuitive objections to utilitarianism. In Chapter 4, I make some introductory remarks. Specifically, I say more about what moral intuitions are, and I explain why they are potentially epistemically significant and trustworthy, both terms that I define more closely. I also consider three existing attempts to defend utilitarianism against intuitive objections, all of which I argue are problematic. In Chapter 5, I argue that the utilitarian can instead defend her theory by showing, not that our moral intuitions fail to be epistemically trustworthy, but that we sometimes fail to carry out the correct thought experiment. In Chapter 6, I consider another approach to defending utilitarianism against intuitive objections, which depends on how sensory imagination plays an important role in making our moral intuitions more trustworthy. Finally, in Chapter 7, I summarize the book's contents.

2. Action Guidance

In this chapter, I discuss what action guidance is, how to define it, and in which ways that utilitarianism is action guiding. I distinguish between two types of action guidance, *doxastic* and *evidential*. I argue that utilitarianism is not doxastically guiding, but that it is evidentially guiding. The main reason to believe that utilitarianism fails to be doxastically guiding is that our acts have wildly unpredictable "chaotic" effects on the future. At the same time, utilitarianism is evidentially guiding because we can use it to obtain reason to believe that an act is right according to the theory. In the next chapter, these two definitions of action guidance are put to use when I discuss action guidance objections to utilitarianism.

2.1 Action Guidance: A First Take

Suppose that you visit me in Uppsala. As you arrive to the airport, you take out a map I sent you that contains instructions for travelling from the airport to my apartment. But the map is barely readable, and my squiggly hand drawn arrows and unconventional abbreviations of street names are nearly incomprehensible. So you call me up and tell me that the map is useless, and then you ask me where I live. I cheerfully tell you to not worry and to "find my apartment" and that "my home is right here." Clearly, you will not be satisfied with my answer. While the map may lead to my home and my utterance that "the apartment is here" may be true, you want an additional quality in the map and the instruction: you want them to be *helpful* to you. If the map and instructions are not helpful to you, then something is wrong with them. Action guidance objections to utilitarianism are closely related to such complaints – they could just as well be called "unhelpfulness objections." Although utilitarianism tells us to maximize pleasure minus pain, to learn this is not helpful. It is like being given the barely readable map or to be instructed to "find my apartment": it leaves us almost wholly in the dark about what to do. And if utilitarianism is not helpful to us, then there is something wrong with it, just as with the map and the instruction to "find my apartment." This is the underlying idea behind action guidance objections to utilitarianism, although it remains to be seen how exactly these objections are to be understood, which is the aim of the next chapter. The present chapter prepares the ground for this upcoming discussion by considering what action guidance is and whether utilitarianism is or is not action guiding.

The main objective of this chapter is to define two kinds of action guidance, so let me begin with some clarificatory remarks about this project. First, when I propose definitions of these kinds of action guidance, I do not claim that the phrase "action guidance" is *ambiguous*, like how "bank" can refer either to the edge of a river or to a place to deposit one's money. With expressions like "ways of guiding," "kinds of guidance," and "types of guidance," I mean only that there are different *versions* of action guidance. This is like how items of food are tasty in different ways even if the term "tasty" is not ambiguous, or how cars belong to different brands even if the term "car" is not ambiguous.

Second, the project of this chapter is *explicative*: it is partly descriptive and partly stipulative. For example, I do not claim that there are *only* two kinds of action guidance, or that philosophers have always or even typically referred to these kinds of action guidance. What is important is that the two kinds of action guidance that I propose can be defined clearly, capture a large part of what we intuitively feel are ways of being action guiding, make sense of the objections that I discuss in the next chapter, and play constructive roles in evaluating these objections.

While the precise definitions of action guidance will follow later in this chapter, let me start by providing two simplified definitions, so you get a rough idea of what I am after. First, a moral theory is *doxastically guiding* for an agent if and only if, by thinking about what it says, the agent *can come to know that an act is right* according to the theory. This is similar to how a map works: how you can come to know where I live by reading the map and thinking about what it says. In contrast, a moral theory is *evidentially guiding* for an agent if and only if, by thinking about what the theory says, the agent *can come to have reason to believe that an act is right* according to the theory.

What is the relation between doxastic and evidential guidance? A demand for a theory to be doxastically guiding is stronger than a demand for it to be evidentially guiding, because if we assume that knowing *p* entails being justified in believing *p*, a theory being doxastically guiding implies it being evidentially guiding, but a theory being evidentially guiding does not imply it being doxastically guiding. For example, by thinking about what a theory *T* says, one might come to have reason to believe that an act is right according to *T*, even as (i) these reasons do not suffice for knowing that the act is right according to *T*, (ii) one fails to believe that the act is right according to *T*, and (iii) the act is not in fact right according to *T*.

As should be clear, the definition of doxastic guidance tries to capture a very demanding kind of action guidance – one of "being able to teach us what to do" – while the definition of evidential guidance tries to capture a minimal kind of action guidance – one of "being practically relevant to us." This naturally raises the issue of whether there is a definition that lies somewhere between these two extremes: one on which utilitarianism is not action guiding, but which is nevertheless less demanding than doxastic guidance. I address this concern at the end of Section 2.4. Finally, for those acquainted with the literature on how to define action guidance, I should note that my definition of doxastic guidance will be similar in some respects to definitions proposed by Holly M. Smith and Eric Carlson.[25] Evidential guidance, in contrast, has not been previously defined in the literature.[26]

2.2 Action Guiding for Us

To begin with, we need to distinguish between an agent *being guided by* and *being able to be guided by* a moral theory.[27] For example, imagine an animal activist who spends his life taking care of rescue chickens, who live out their full lives on his sanctuary farm. We can suppose that the activist is *able* to be guided by utilitarianism – perhaps he believes this theory is true – even if he never *in fact* is guided by it. For example, perhaps he does not use utilitarianism to guide his actions because that would require him to not only rescue chickens, which he finds deeply rewarding, but also to raise millions of additional chickens for the sake of maximizing pleasure minus pain, which he finds repugnant. In what follows, whenever an agent is able to be guided by a moral theory, or in a position to be guided by it, I will say that the moral theory is *action guiding for* the agent. Another way to explain the difference between being guided by and being able to be guided by a moral theory is this: to be guided by utilitarianism includes performing an action and to do so for distinctly utilitarian reasons. In other words, being guided by a moral theory *ends* in an action; you cannot be guided without acting. But to *be able* to be guided by a moral theory need not end in an action, as one can be able to be guided by utilitarianism even if no action is performed.

When formulating action guidance objections to utilitarianism, the relevant issue is whether the theory is action guiding for us: that is, whether we have the ability to guide ourselves by means of the theory. The relevant issue is *not* whether we are ever *actually* guided by utilitarianism. To continue the analogy with the map, suppose that had you consulted the map, you would have come to know exactly where I live. But as it is, you do not consult the map, take a wild guess, and end up in the wrong part of Uppsala. In that case, the problem clearly resides with you and your behavior, not with the map. Similarly, if the

[25] More precisely, doxastic guidance is similar to what Smith calls external guidance (1988), pp. 91-95, broad guidance (2012), pp. 370-378, and guidance in the extended sense (2018), pp. 11-32, and to what Carlson calls AG1 (2002), pp. 73-76.

[26] Smith defines a weaker notion of action guidance that she calls internal guidance (1988), pp. 91-92, narrow guidance (2012), p. 374, and guidance in the core sense (2018), pp. 12-21. These definitions of action guidance are very different from that of evidential guidance. Moreover, while utilitarianism is evidentially guiding, it is arguably not action guiding on Smith's definitions.

[27] This distinction is made by Carlson (2002), pp. 73-74.

animal activist does not guide his behavior by means of utilitarianism, then the problem resides with him and not with the theory.

When we consider whether a moral theory should be action guiding for agents, we first need to decide *for which* agents it should be action guiding. Perhaps we do not want to require *universal guidance* – i.e. that a moral theory can guide *all* moral agents. There is most likely no moral theory that is universally guiding, and the action guidance objector may not want to put forward such a sweeping criticism of moral theories. Of course, to require only *partial guidance* raises the issue of where precisely to draw the line. For example, is a moral theory problematic if it cannot guide non-human animal or extraterrestrial moral agents? What about the ancient Egyptians or future humans? Does a moral theory need to be guiding for small children and the cognitively impaired? Is it enough if we require that a theory is action guiding for *nearly* all agents, with just a few exceptions? For the objections that I will discuss, we need not answer these questions, as it is enough to focus on a subgroup of all moral agents for which it is uncontroversial that a moral theory should be action guiding: namely the group of nearly all adult contemporary humans with normal human cognitive capacities.[28]

A related issue concerns *in which choice situations* a moral theory should be action guiding for agents. Let us define a choice situation as a range of alternative and mutually exclusive acts available to a particular agent at a particular point in time. Examples of choice situations are deciding whether to buy fries with your meal at McDonald's, or deciding whether to walk, bicycle, or drive to work. Here again the problem of demarcation arises – we can require that a theory should be guiding with respect to all choice situations, or merely to a proper subset of these situations. Some exceptions seem plausible. For example, in choice situations where we lack the necessary time to consult a moral theory, such as when falling off a cliff when hiking in the mountains, we need not require that the theory is action guiding for agents. Other choice situations should be excluded because they are extremely unusual, such as finding yourself in a hostage situation. Finally, some choice situations are such that a normal human adult cannot reasonably be expected to focus on what a moral theory says in the situation, as in the movie *Aliens* when Ripley, pointing a flamethrower, faces down an alien. Therefore, I will limit the relevant choice situations to nearly all those ordinary choice situations where we have sufficient time for deliberation and where we can reasonably be expected to focus on what the moral theory says.

Going forward, it will be useful to have a shorthand expression to sum up the above expectations on a moral theory. Therefore, I will stipulate that a moral theory is "action guiding for us" if and only if it is action guiding for (a) nearly all (b) adult, (c) contemporary (d) humans with (e) normal human cognitive capacities in (f) nearly all (g) ordinary choice situations where there

[28] Cf. Väyrynen (2006), p. 296.

is (h) sufficient time for deliberation and (i) where the agent can reasonably be expected to focus on what the theory says.

2.3 Explicit Cognitive Action Guidance

Doxastic and evidential guidance are examples of *cognitive* action guidance.[29] Cognitive action guidance is action guidance by means of your cognitive abilities. Were you to read the map, think about what it says, and come to learn where I live on this basis, then you would have been cognitively guided by the map. Basically, cognitive guidance involves two main ingredients: cognitively processing what the guiding object "says" (in this case, visually inspecting the map, comprehending the instructions, interpreting the drawings, and so on), and coming to a conclusion on the basis of this processing. Similarly, were you to read the formulation of a moral theory, think about what it says, and come to learn that it tells you to donate to the Against Malaria Foundation on this basis, then you would have been cognitively guided by the theory. This account of cognitive guidance is meant to be agnostic about what cognitive abilities are involved: it could be pattern recognition, linguistic competence, or elementary logic, to name a few examples.

Non-cognitive action guidance is defined negatively, as action guidance that is not cognitive. As an example of non-cognitive guidance, consider a roadblock. A roadblock can guide us in a traffic situation, but not by means of us cognitively processing what the roadblock says. Instead, it guides us in virtue of its shape and form as a physical obstacle. Similarly, a law of nature can guide the research efforts of a scientist, in that the scientist performs her experiments because the law of nature is true and makes these experiments feasible and productive. But the scientist might never have heard of this law of nature, and might never have thought about what it says. While the scientist is guided by it in the sense that it causally influences her behavior, she is not guided by means of using her cognitive abilities to process what it says. To give a final example, utilitarianism has non-cognitively guided my choice of research topic. I decided to write this book because utilitarianism is an interesting theory – therefore, had it not been interesting, I would not have written this book. Even so, I have not been guided in writing this book by *processing information* or *thinking about* what utilitarianism says is right, and so I have not been cognitively guided by the theory. That is, to write this book I never tried to apply utilitarianism. Had I been cognitively guided by utilitarianism

[29] I frequently shorten the names of the different kinds of action guidance to make the text more readable. For example, cognitive action guidance is referred to as cognitive guidance, and doxastic explicit cognitive action guidance is referred to as doxastic guidance.

in my choice of occupation, I would have done something else with my time, such as working a lucrative job to fund effective animal charities.

We can further distinguish between two kinds of cognitive guidance: explicit and implicit.[30] When I am explicitly guided by a rule, principle, or theory, I represent it to myself (i.e., I "have it in mind") and think about what it says. For example, imagine an inexperienced driver. As she arrives at a crossing, she sees that the light is red, represents the rule "stop for red lights," and concludes that she should stop the car. In this case, the driver has been explicitly cognitively guided by the rule "stop for red lights."

In contrast to explicit guidance, implicit guidance involves no explicit representation of a rule, theory, or principle. An experienced driver who sees a red light will not represent the rule "stop for red lights" to herself, but will stop her car without consciously thinking about the rule. Nonetheless, the experienced driver is still cognitively guided by the rule "stop for red lights," since it is because she processes information about the stoplights and the rule that she stops her car. In this case, there is not an absence of cognitive processing; it has merely moved to a sub-conscious or non-conscious level.

Here is another way to explain the difference between explicit and implicit guidance. In the case of explicit guidance, the cognitive work takes place "before the mind's eye," and it is therefore in principle available to introspection. But in the case of implicit guidance, the cognitive work does not take place "before the mind's eye" – the whole process is hidden from view. While the inexperienced driver can follow her thoughts by introspection, the experienced driver may find it less than trivial to account for why she acted as she did. The experienced driver cannot simply introspect her thought processes to find out why she stopped the car.

There is an additional distinction to be drawn between two kinds of explicit cognitive guidance, which follows a commonly drawn division between two systems of thinking: system 1 and system 2 thinking.[31] To begin with, let us consider the difference between these systems. System 1 thinking is automatic, fast, and intuitive, whereas system 2 thinking is manual, slow, and controlled. For example, suppose that you are asked to solve for x in 10*10=x. In this case, the answer "100" just *pops up* without any effort. This is an example of system 1 thinking – you see the answer quickly and effortlessly. On the other hand, suppose that you are asked to solve for x in 12*12=x. In this case, you may need to use some system 2 thinking, and work through the problem step-by-step. For example, you may reason as follows: "10*12 is 120, while 2*12 is 24, so the answer is 120+24=144." This kind of problem solving is an example of system 2 thinking: of manual calculation rather than of direct intuition. In reality, we nearly always use a mix of the two types of thinking to solve a particular problem. For example, even in the latter case I use system 1

[30] This distinction is due to Smith (1988), p. 90; (2018), p. 28.

[31] Kahneman (2013), pp. 19-105.

thinking to some extent, such as to directly see (a) that 10*12 is 120, (b) that 2*12 is 24, and (c) that if I add the products of 10*12 and 2*12, then I will obtain the answer to 12*12.

Corresponding to the distinction between system 1 and system 2 thinking, there is what we may call *automatic* explicit cognitive guidance and *manual* explicit cognitive guidance. For an illustration of these kinds of guidance, consider two inexperienced drivers. The first driver has taken driving courses for some time, while the second driver has never sat in a car before. While the first driver explicitly represents the rule "stop for red lights" to herself, she directly sees that it applies to the situation she is in. She is explicitly cognitively guided by the rule, but in an automatic way. The second driver might instead engage his manual thinking (we are assuming a *very* inexperienced driver). For example, before stopping his car, he may think to himself:

(i) I have reached a crossing and the light is red.
(ii) The rule "stop for red lights" says that I should stop at crossings where the light is red.
(iii) Therefore, I should stop at this crossing according to this rule.
(iv) I should follow the rule "stop for red lights."
(v) Therefore, I should stop at this crossing.

Again, remember that automatic and manual guidance are instances of explicit cognitive guidance. In the above case, each of the drivers represents to herself a proposition by explicitly thinking about what it says, or by "having it in mind." What distinguishes the drivers is how the explicit cognitive guidance plays out, whether it is automatic and fast, or whether it is manual and slow. In fact, most actual instances of explicit cognitive guidance will be automatic and manual not fully, but to certain degrees, such as 80% automatic and 20% manual, for example. In the sections that follow, I assume that action guidance objections are concerned solely with explicit cognitive action guidance, but that whether the guidance is automatic or manual does not matter. Consequently, when I use the expression "thinking about" what moral theories say I mean to be neutral between this thinking being automatic or manual. What is important to keep in mind is that explicit guidance – which essentially involves the representation of principles, rules, and theories – need not only come in the guise of manual guidance, but can be automatic to various degrees.

Figure 1 summarizes the distinctions that have been introduced in this section, as well as their hierarchical order (the precise definitions of doxastic and evidential guidance is the subject of the next section).

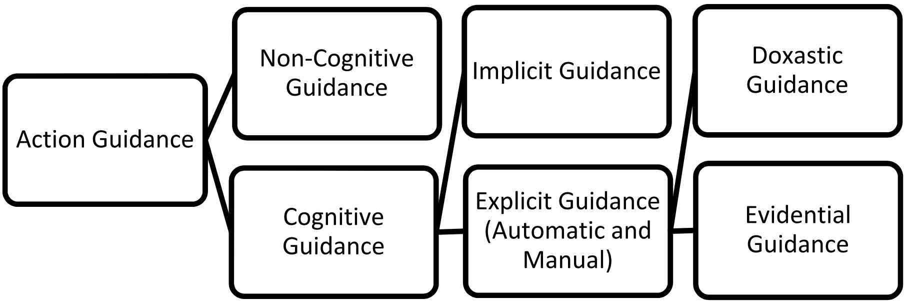

Figure 1. Kinds of action guidance

2.4 Doxastic and Evidential Guidance

How can we define doxastic guidance? As I have noted, doxastic guidance is an instance of explicit cognitive guidance. In Section 2.1 I defined it as follows: a moral theory is doxastically guiding for an agent if and only if, by thinking about what it says, the agent can come to know that an act is right according to the theory. This simplified definition is not ideal for several reasons, which I will elaborate on below. The full definition of doxastic guidance is considerably more complex than the simplified one, so it requires more in the way of explanation. It goes as follows:

> DOXASTIC GUIDANCE
>
> A moral theory *M* is *doxastically guiding* at a time *T* for an agent *S* with respect to a choice situation *C* of the agent if and only if:
> (i) there is an act *A* in *C* that *S* knows at *T* how to perform and that is such that were S both (a) not to know at *T* that *A* is right according to *M* and (b) think about which act is right according to *M*, then *S* would on this basis come to know that *A* is right according to *M*, and
> (ii) *S* has the necessary time and cognitive ability at *T* to think about what is right in *C* according to *M*.

I will now elaborate on each part of this definition. First, note that the definition distinguishes between (a) the time when the theory is guiding for the agent, and (b) the choice situation with respect to which the theory is guiding for the agent.[32] The time and choice situation need not be contemporaneous. For example, a moral theory can be action guiding *for me today* with respect to a choice situation that I will face *tomorrow*.

Second, consider the initial part of condition (i), that "there is an act *A* in *C* that *S* knows at *T* how to perform." The requirement that the agent knows how to perform the act is designed to sidestep a problem raised by Eugene Bales and elaborated on by Smith.[33] A definition of action guidance must prevent a moral theory from being action guiding in an "unhelpful" way, such as in virtue of prescribing the act "maximize pleasure minus pain." Most of us can think about what utilitarianism says and come to know that it tells us to *maximize pleasure minus pain*. But this does not suffice for utilitarianism to be action guiding in any interesting sense. The kind of action guidance we are concerned with is about a theory being *helpful*, and it is not helpful to learn that we should maximize pleasure minus pain. The problem is that we do not know *how* to maximize pleasure minus pain. Therefore, by requiring that we know how to perform the act in question, we exclude acts such as "maximize pleasure minus pain" from consideration; unlike easily performed acts of drinking coffee or taking a walk, which we do know how to perform.[34]

Third, consider the remaining part of condition (i), which consists of the past subjunctive proposition "were S both (a) not to know at *T* that *A* is right according to *M* and (b) think about which act is right according to *M*, then *S* would on this basis come to know that *A* is right according to *M*." This part of the definition is self-explanatory in light of my earlier presentation of explicit cognitive guidance, except for my reason to include conjunct (a) of the antecedent, namely that of were S "not to know at *T* that *A* is right according to *M*." I formulate the definition in this complicated way to avoid the following objection. If an agent *already* knows that an act is right according to a theory, then it is impossible for her to come to know that it is right *on the basis of* thinking about what the moral theory says. This is because already knowing that *p* precludes coming to know that *p*. However, that a person already knows what utilitarianism prescribes in a situation should surely not prevent utilitarianism from being doxastically guiding for her. To return to the map analogy, if you already know the way to my apartment, then you cannot learn where I live by reading a map, simply because you cannot learn what you already know. However, this clearly does not point to any problem with the map. The interesting question here is this: Would you be able to learn where I live by

[32] This has been noted by Smith, who includes the distinction in her definitions of action guidance. See, for example, Smith's definition of having the ability to directly use a moral theory in the extended sense in Smith (2018), p. 18.

[33] Smith (2012), pp. 373-376; (2018), pp. 15-21, and Bales (1971), p. 261.

[34] This is inspired by Smith's solution. In her definitions, Smith employs the concept of epistemic ability, which is due to Goldman (1976), pp. 192-204. See also Smith (2018), pp. 18-21. Another way to escape the problem is to say that the act of maximizing pleasure minus pain is not even available to the agent, and so is not right (or wrong) according to utilitarianism. I discuss such issues later in Section 3.5.

reading the map *if* you did not already know my apartment's location? Similarly, the interesting question with respect to moral theories is this: Would you be able to know what a moral theory says is right by thinking about what it says *if* you did not already know what the theory says is right?

Fourth, consider condition (ii), that "*S* has the necessary time and cognitive ability at *T* to think about what is right in *C* according to *M*." The reason for including this condition in the definition of doxastic guidance is that an agent might lack the time to think about what a theory says, or that the theory could be formulated using extremely complex technical terminology.[35] In such cases, it might still be true that *were* the agent to think about what the theory says, she *would* come to believe that an act is right according to it on this basis. This is because of how, in the closest possible world where the agent thinks about what the theory says, she *does* have the time and cognitive ability to think about what it says. Nevertheless, if she in her actual circumstances lacks such time or cognitive ability, the moral theory is intuitively not action guiding for her.

Let us now consider an example of how to apply the definition of doxastic guidance to a concrete case. Suppose that at 5:00 p.m., I see an unconscious man lying at the wayside 20 meters in front of me. As I close in, I will soon be faced with a choice situation consisting of two mutually exclusive alternative actions: to help the man or to keep walking. Is utilitarianism doxastically guiding for me at 5:00 p.m. with respect to this soon-to-be-realized choice situation? Let us consider the two conditions (i) and (ii), as utilitarianism is doxastically guiding for me if and only if both are satisfied. The satisfaction of the second condition (ii) is easy to determine: I definitively have the time to think about what is right in this choice situation according to utilitarianism; for example, I can simply stop walking for a moment. So condition (ii) is satisfied. What about condition (i)? To begin with, "to help the man" and "to keep walking" are both acts that I know how to perform. But neither act is such that, were I not to know what utilitarianism tells me to do in the situation (which in fact I do not), then thinking about what act it prescribes would make me come to know that the act is right according to utilitarianism. For example, I will not know what consequences that helping the man will have over a million years, even if I think about it for some time. So condition (i) is not satisfied. It follows that utilitarianism is not doxastically guiding for me at 5:00 p.m. with respect to the choice situation of helping the man.

There are three striking consequences of my definition of doxastic guidance, each of which merits further comment.

First, consider a poor, illiterate farmer who is deciding which crop to grow this year. He does not know what utilitarianism says, and he lacks the education and time required to learn it. Nevertheless, on the definition which I have proposed, utilitarianism could be doxastically guiding for the farmer at the

[35] Cf. Carlson (2002), p. 73, on the importance of being able to deliberate.

time just before planting season and with respect to the choice situation of deciding which crop to plant. In particular, condition (ii) of the definition does not rule this out. Even if the farmer does not know about utilitarianism, he still has the time and cognitive ability to think about what it says – we are not requiring that the farmer *can* think about what it says, but merely that he has the *time and cognitive ability* to do so. As it is, I think this is the correct verdict. A map's helpfulness is not diminished by it being hidden in a safe; similarly, a moral theory's helpfulness is not diminished by people not knowing about it. Such "external" factors should not count when it comes to determining whether a theory is action guiding for us.

Second, suppose that you can either give your money to a starving child or use the same money to buy pizza. Suppose that according to utilitarianism, it is right to help the starving child and wrong to buy pizza. Finally, let us assume that, as a matter of fact, you will not think about what utilitarianism says about this choice situation, since you do not care much for this particular moral theory – we can stipulate that you are a staunch Kantian. We now add the following complication to the case: while you deliberate over what to do, your every move is being watched by a terrorist who will bomb a school if and only if you think about what utilitarianism says *and* then help the starving child. In this case, it could be true that utilitarianism is doxastically guiding for you, even if you could never use the theory to find out that it tells you to help the starving child. This is because if you thought (correctly) about which act is right according to utilitarianism, then you would have to conclude that buying the pizza is right – since if you helped the starving child in such a situation, the terrorist would bomb the school. Perhaps surprisingly, this example shows that even if a theory is doxastically guiding with respect to an agent and a choice situation, this does not guarantee that the theory is helpful for letting the agent find out what is *actually* right in that choice situation. The reason is that the act of thinking about what a theory says might have further consequences for what is right according to it. This consequence appears acceptable to me, but others may find it less so.

Third, consider a person who knows which of her available acts maximize pleasure minus pain, and also knows that utilitarianism tells her to maximize pleasure minus pain. It may appear that utilitarianism *must* be action guiding for this person. But utilitarianism can fail to be doxastically guiding for this person. Suppose that the person suffers from a strange mental illness, such that whenever she thinks about what a moral theory prescribes, she is cognitively paralyzed and fails to believe anything about what is right according to the theory. In this case, it will for any of her available acts be false that, were this person not to know that the act is right according to utilitarianism, and think about which act is right according to the theory, then she would on this basis come to know that the act is right according to the theory. So condition (i) is not satisfied for this agent in any choice situation. Is this a problem for my definition of doxastic guidance? I do not think so. We must distinguish between being able to use a theory to guide our behavior (i.e., the question of whether it is action guiding for us) and being able to do what the theory says. The person with the strange mental illness can do what utilitarianism says, but she cannot use utilitarianism to guide her actions.

Having defined doxastic guidance, I will next define the notion of evidential guidance. The differences in formulation from doxastic guidance are relatively minor; rather than referring to coming to know that an act is right, this definition refers to coming to have *reason to believe* that an act is right. The differences between the definitions of doxastic guidance and evidential guidance are put in bold:

> EVIDENTIAL GUIDANCE
>
> A moral theory *M* is ***evidentially guiding*** at a time *T* for an agent *S* with respect to a choice situation *C* of the agent if and only if:
>
> (i) there is an act *A* in *C* that *S* knows at *T* how to perform and that is such that were S both (a) **not to have any reason** at *T* to believe that *A* is right according to *M* and (b) think about which act is right according to *M*, then *S* would on this basis come to **have reason to believe** that *A* is right according to *M,* and
>
> (ii) *S* has the necessary time and cognitive ability at *T* to think about what is right in *C* according to *M*.

To see how this definition can be applied to a particular case, consider again the choice situation where I can either help a man or walk past him. As we saw earlier, condition (i) of the definition of doxastic guidance was not satisfied, since thinking about what utilitarianism says will not provide me with *knowledge* of which act is right to perform according to utilitarianism. However, at least on the face of it, condition (i) of the definition of evidential guidance is satisfied, since by thinking about what utilitarianism says, I will come to have some (although only a little) reason to believe that helping the man is right according to utilitarianism. This is because I have reason to believe that helping him will prevent future pain from illness, or missed pleasures from him being unconscious or dead. Furthermore, condition (ii) of the definition of evidential guidance, being identical to condition (ii) of the definition of doxastic guidance, is again satisfied: I have the time and cognitive ability to think about what utilitarianism prescribes at this time. Therefore, we can conclude, at least tentatively, that utilitarianism is evidentially guiding for me at this time and with respect to this choice situation. I say "tentatively," because it is perhaps not obvious that condition (i) is in fact satisfied in cases like these: I will evaluate more carefully whether utilitarianism is evidentially guiding for us in Section 2.6.

For now, I will return to a concern that I brought up earlier in this chapter: that the definitions of doxastic and evidential guidance are too extreme, and that there exists some more interesting middle ground between these definitions. Could we not formulate a definition that refers, not to us coming to have *knowledge* or *reason for belief*, but to us coming to have *justified belief* that an act is right according to utilitarianism? Would that not give us a more modest definition of action guidance? Consider the following definition:

> A moral theory *M* is action guiding at a time *T* for an agent *S* with respect to a choice situation *C* of the agent if and only if:
>
> (i) there is an act *A* in *C* that *S* knows at *T* how to perform and that is such that were S both (a) **not to have justified belief** at *T* to believe that *A* is right according to *M* and (b) think about which act is right according to *M*, then *S* would on this basis come to **have justified belief** that *A* is right according to *M*, and
>
> (ii) *S* has the necessary time and cognitive ability at *T* to think about what is right in *C* according to *M*.

I admit that the above definition may capture an interesting and "properly demanding" sense of action guidance at least as well, or even better, than does doxastic guidance. Nevertheless, the above definition is unhelpful for evaluating action guidance objections to utilitarianism. The problem is that utilitarianism *is* action guiding on the above definition, and so this definition cannot be used to successfully object to utilitarianism. That utilitarianism is action guiding on this definition can be demonstrated as follows. Consider the following two acts:

A1. Give 23 dollars to the Red Cross today.
A2. Do not give 23 dollars to the Red Cross today.

A2 is what we may call a *negative act* or an *act of avoidance.*[36] By thinking about what utilitarianism says in a choice situation which involves an alternative act such as A2, we will easily acquire justified belief that performing A2 is right according to utilitarianism. The reasoning goes as follows. It is *extremely unlikely* that giving precisely 23 dollars to the Red Cross today will maximize pleasure minus pain among one's available alternatives for purely statistical considerations. There are *so many* alternatives to giving 23 dollars to the Red Cross today, perhaps thousands or even tens of thousands of acts, including many small variations on this act, such as giving 24 dollars, giving 22 dollars, and so on. So we will be justified in believing that A1 – giving to the Red Cross – is wrong according to utilitarianism. And if A1 is wrong according to utilitarianism, then A2 must be right, because the *only* way in which to not perform A1 is to perform A2. Indeed, the above result should be fairly

[36] My thanks to Erik Carlson for drawing my attention to such acts, in the context of theories being action guiding.

unsurprising. It is analogous to how, when I randomly choose a lottery ticket among 1000 tickets and there is only one winning ticket, I am justified in believing that the particular ticket that I bought will not win. In such a case, it is just too *unlikely* that the ticket I picked is the winning one. The above argument generalizes to other choice situations as well. In any choice situation, there will be a highly specific alternative act (such as A1) whose negative counterpart (such as A2) we are justified in believing is right to perform according to utilitarianism.

The same argument cannot be used to show that utilitarianism is doxastically guiding. Even if I am justified in believing that A2 is right, I will not *know* that A2 is right; and even if I am justified in believing that the lottery ticket will not win, I will not *know* that the ticket will not win. In general, knowledge that *p* cannot be obtained by the mere statistical consideration of the probability that *p*, which is not the case with justified belief. There are various explanations for this fact, which I will not evaluate here. In general, however, knowing that *p* seems to require that a certain causal connection holds between one's belief that *p* and *p* itself, for which mere statistical consideration of probabilities do not suffice. Now, for action guidance objections to get off the ground, and for us to get the strongest possible versions of these objections, we need to find a definition of action guidance on which utilitarianism is *not* action guiding. Therefore, we should focus on a definition of action guidance such as doxastic guidance: one which is formulated in terms of gaining knowledge, and not in terms of gaining justified belief, that an act is right according to a moral theory.

Another worry one might have is that doxastic and evidential guidance represent not different *kinds* of guidance, but only *degrees* to which a theory can be action guiding. On this proposal there is only *one* kind of action guidance, and to fulfill the criteria for evidential guidance is to be action guiding in this one way to *some* extent, while fulfilling the criteria for doxastic guidance is to be action guiding in this one way to a *high* extent. However, I believe this concern is unfounded. As I have glossed it, for a theory to be evidentially guiding, it must be able to play a role in deliberation or decision making – i.e., it must be "practically relevant." But if a moral theory is evidentially guiding, it is not just practically relevant to a *low* degree – it is practically relevant to the *maximum* degree.[37] So the definition of evidential guidance does point to something different than doxastic guidance – i.e., a theory's practical relevance. In any case, the difference between these two ways of characterizing the definitions of action guidance will not matter for the discussion that follows in the next chapter. For my purposes in this book, it will do just as well to think about these definitions as being about a single dimension of action

[37] Presumably, there is not even a scale here – either a moral theory is practically relevant, or it is not.

guidance, as it does to think of them (as I have in this chapter) as being about two separate dimensions.

What about *other* definitions of action guidance than those that I have considered in this chapter? Both Smith and Carlson have proposed their own detailed definitions of action guidance, seemingly with the aim to capture a type of demanding action guidance similar to that of doxastic guidance.[38] I have hesitated to use Smith and Carlson's definitions in my discussion for two reasons. To begin with, it seems to me that their definitions require some revisions. Carlson does not address the issue of utilitarianism incorrectly being action guiding for an agent merely in virtue of prescribing that she performs the act "maximize pleasure minus pain." Moreover, Smith's most recent definition of action guidance incorrectly implies that utilitarianism is not action guiding for an agent who lacks a belief about whether an act maximizes pleasure minus pain, even if the agent would have had such a belief (even one amounting to knowledge) had she thought about what utilitarianism says.[39] More importantly, however, my definition of doxastic guidance lends itself well to defining a notion of evidential guidance, and I am not sure of how to best modify Carlson's and Smith's definitions for this purpose.

2.5 Is Utilitarianism Doxastically Guiding?

With our definitions in hand, let us now consider whether utilitarianism is action guiding for us. Recall first the stipulated definition of "action guidance for us" that I introduced earlier:

> A moral theory is "action guiding for us" if and only if it is action guiding for (a) nearly all (b) adult, (c) contemporary (d) humans with (e) normal human cognitive capacities in (f) nearly all (g) ordinary choice situations where there is (h) sufficient time for deliberation and (i) where the agent can reasonably be expected to focus on what the theory says.

[38] Smith (1988), pp. 91-95; (2012), pp. 370-378; (2018), pp. 11-32, Carlson (2002), pp. 73-76. I review Smith (2018) in Rosenqvist (2019). For more work on how to define action guidance, see Väyrynen (2006), pp. 293-294, and Mason (2013), pp. 3-8. With "Carlson's definition" I refer to what he calls AG1. His second definition, AG2, is a substantially more demanding definition of action guiding – impermissibly so, in my opinion – as it includes a motivational requirement.

[39] The problem is that Smith's definition of "ability in the extended sense to use a principle" – one of her more recent definitions of action guidance – requires that the agent *truly believes* that if she performed some act, then that act would (in the case of considering the principle of utilitarianism) maximize pleasure minus pain. But surely utilitarianism can be action guiding for a person without this being true: a person might not form a belief about what utilitarianism prescribes until *after* she has thought about what it says. Nevertheless, utilitarianism can clearly be action guiding for her *before* such thinking takes place.

As I will consider two kinds of action guidance – doxastic guidance and evidential guidance – the relevant questions are whether utilitarianism is doxastically guiding for us versus evidentially guiding for us.

Next, recall the definition of doxastic guidance:

> DOXASTIC GUIDANCE
>
> A moral theory *M* is *doxastically guiding* at a time *T* for an agent *S* with respect to a choice situation *C* of the agent if and only if:
>
> (i) there is an act *A* in *C* that *S* knows at *T* how to perform and that is such that were S both (a) not to know at *T* that *A* is right according to *M* and (b) think about which act is right according to *M*, then *S* would on this basis come to know that *A* is right according to *M*, and
>
> (ii) *S* has the necessary time and cognitive ability at *T* to think about what is right in *C* according to *M*.

What about condition (ii)? When an agent has sufficient time for deliberation, she has the time to think about what utilitarianism says. Moreover, all humans with normal human capacities will have the cognitive ability to think about what utilitarianism says, as it is a fairly simple moral theory. Therefore, condition (ii) is satisfied for us, in the light of my definition of what it means for a theory to be "action guiding for us."

However, condition (i) is not satisfied for us. In many or even all choice situations, there is no act such that: had we not known that the act is right according to utilitarianism, and subsequently thought about which act is right according to the theory, then we would on this basis come to know that the act is right according to utilitarianism. The main problem is that our acts have "chaotic effects," which makes their consequences over even shorter time spans wildly unpredictable. (The term "chaos" comes from the branch of mathematics known as "chaos theory," which deal with deterministic systems highly sensitive to minor variations in initial conditions).[40] Since condition (i) is not satisfied for us, utilitarianism is not doxastically guiding for us.

What does it mean that acts have chaotic effects, and why does it matter for whether utilitarianism is doxastically guiding? That acts have chaotic effects means that their consequences "ripple outward" in wildly unpredictable (that is, chaotic) ways. To illustrate, suppose that one Tuesday afternoon I brew tea instead of coffee before going home from work. Because brewing tea takes a few seconds less time than brewing coffee, I finish work faster and leave for

[40] That consequentialist theories are difficult to apply because they have chaotic effects is noted by Lenman (2000), pp. 344-350. See also Frazier (1994), pp. 46-47. For responses to Lenman, see Mason (2004), Cowen (2006), Lukas (2008), Hare (2011), pp. 199-201, Dorsey (2012), Burch-Brown (2014), and Greaves (2016).

home earlier (we are assuming a flexible workplace schedule). Driving home earlier puts me in a different traffic situation than I would otherwise have been in. Moreover, the exact nature of which traffic situation I am in slightly affects the schedules of other drivers that I interact with. Accordingly, some drivers will be delayed, or arrive earlier, by just a second or two because I brewed tea instead of coffee. Therefore, just as I operate on a slightly different schedule because I brewed tea instead of coffee, other drivers will *also* come to operate on different schedules because of my choice. Moreover, their schedules will affect the schedules of others that they interact with during the rest of the day – partners, children, friends, waiters, and so on. In this way, such small second – or even millisecond – variations in schedules will reverberate through the nearby population and beyond.

At this point in the story, I am considering fairly inconsequential effects of brewing tea: whether someone goes to bed a second or two earlier may not matter for the total amount of pleasure minus pain being produced.[41] But it is easy to see how, sooner or later, the consequences of brewing tea that Tuesday will become much more serious. To see this, it is sufficient to consider acts of procreation.[42] Even very slight variations in how people have pro-creative sex will affect which people are ultimately born, because there are so many possible sperm-ovum combinations, each of which would bring into the world a unique person. And the moment that the effects of my brewing tea spills over into making one person rather than another being born – be it a day or a year from when I performed the act – the ensuing effects will snowball quickly. Different people existing means different lives being lived, which will have more dramatic effects on people's schedules and when still more people are born. And so it continues until every single person on Earth owes her existence to my choice of brewing tea.

The above reasoning shows that a folk theory or folk belief of causation must be false: namely, the view that the causal ramifications of ordinary and trivial acts are both (a) locally restricted and (b) diminish over time. G. E. Moore puts this view (although he does not confidently assert it) as follows:

> It does in fact appear to be the case that, in most cases, whatever action we now adopt, "it will be all the same a hundred year hence," so far as the existence at that time of anything greatly good or bad is concerned: and this might, perhaps, be *shewn* to be true, by an investigation of the manner in which the effects of any particular event become neutralised by lapse of time.[43]

[41] Although even this is not clear: a second of unconsciousness due to sleep will contribute less to total pleasure minus pain than will a second of mild pleasure due to remaining awake.

[42] Cf. Parfit (1984), p. 361.

[43] Moore ([1903] 2004), p. 153. Moore goes on to point out that "[f]ailing such a proof, we can certainly have no rational ground for asserting that one of two alternatives is even probably right and another wrong."

J. C. C. Smart calls this the *ripples in the pond postulate*:

> [W]e do not normally in practice need to consider very remote consequences, as these in the end approximate rapidly to zero like the furthermost ripples on a pond after a stone has been dropped into it.[44]

However, if even trivial acts like brewing tea affects which people will be brought into existence, then few or none of our acts can be such that their consequences will sooner or later "approximate rapidly to zero." Therefore, the ripples in the pond postulate must be false.[45]

While the ripples in the pond postulate is false, a more moderate postulate may be true. In many facets of life, there are "timing correctors," taking the form of norms, habits, desires, traditions, and institutions, which make some events relatively insensitive to the exact timing of other events. For example, if you need to get up at 7:00 a.m. every morning to get to work, then even if you went to bed a few seconds earlier or later due to my brewing tea, you will still get up at 7:00 a.m. Similarly, if I want to go to university, I will not let minor delays stop me, and will find the time to send in my application. If I do not have the time to send it in today, I will do so tomorrow. In general, the existence of timing correctors explain why I can trust everyone to show up at the right time to a meeting, despite the persistent chaotic effects that I just mentioned. Such correcting influences will partly counteract at least some chaotic effects.

Moreover, even when we consider acts that affect which people are born, it is not obvious that which individuals exist will have major long-term consequences on the broad outlines of history. To begin with, many of us are easily replaceable in our current occupations; if a person is not hired to a job, someone else will be hired. In addition, many important technologies would perhaps have been invented even without the actual people who made the inventions simply because "the time was right" for their invention.[46] The above argument amounts to a kind of light-weight technological determinism (not to be conflated with philosophical determinism): although brewing tea rather than coffee might affect the minor details of when *exactly* television or the internet were invented, barring larger catastrophes their invention was, the argument goes, preordained as the result of larger societal developments.

So far so good, but even if what I just said is true, this is not enough to make utilitarianism doxastically guiding for us. To know that an act is right according to utilitarianism, we need to know that it *maximizes* pleasure minus pain. This focus on maximization means that even if an act A produces only

[44] Smart (1973), pp. 33-34.

[45] Several authors have objected to the Ripples in the Pond Postulate. See Frazier (1994), pp. 47-49, Lenman (2000), pp. 350-351, and Greaves (2016), pp. 313-315.

[46] Cf. Wright (2000).

slightly less pleasure minus pain than some alternative act, A is wrong according to utilitarianism. In other words, according to utilitarianism, very little is required to "push" an act from being right to being wrong. For example, suppose that brewing tea instead of coffee only affects the schedules of a few drivers that I interact with during my half hour commute, and does so only for a few hours. After this interval, timing correctors ensure that their schedules are properly "reset." Basically, we are assuming that something like the ripples in the pond postulate is true. In this example, any differences in pleasure minus pain due to brewing tea compared to brewing coffee is rather small. Nevertheless, it seems plausible to think that brewing tea will at least *somewhat* affect which pleasures and pains are produced during that half hour. Even a second more of anxiety on behalf of one driver, or the sudden occurrence of an unpleasant thought to another, might be enough to push brewing tea from being right to being wrong. Moreover, for almost all of our acts it will be extremely difficult to predict even such minor near-term consequences of our acts, because the consequences unfold quickly and mostly out of sight. Finally, note that what we need is *knowledge* that an act maximizes pleasure minus pain, but that the above kind of reasoning gives us at most *justified belief* that an act maximizing pleasure minus pain – the relevant causal connection between this belief and the inner lives of the unknown commuters is lacking.

To make things worse for the utilitarian, note that in the case that I just described, I introduced a major simplification. I assumed that there are only *two* alternative acts available to me in the choice situation – to brew tea or to brew coffee; however, real choice situations include many more alternatives. Of the thousands of acts available to me at any given moment, only a very small number of these will maximize pleasure minus pain. For example, when traveling to work, one might think that there exist only a few alternatives: to travel by car, bus, or bicycle. But there are many alternatives that we usually do not consider, either because they are unusual or because they are inconvenient. I can take a taxi to work, I can walk to work even if it takes several hours, I can walk in the opposite direction for an hour before turning around and taking the bus, and so on. And there are many different routes to travel by car, many different ways to deal with traffic situations, etc. Only one or a few of these acts will actually be right according to utilitarianism, since only one or a few of them will maximize pleasure minus pain. As a result, not only do we have to predict an act's consequences in exact ways in order to know that it maximizes pleasure minus pain, but we also have to do this for the act's thousands of alternatives. This is, to put it mildly, impossible for any normal human being.

Another response to the claim that our acts have chaotic effects is the *cancellation postulate*, which Kagan puts as follows:

> Although we may lack crystal balls, we are not utterly in the dark as to what the effects of our actions are likely to be; we are able to make reasonable, educated guesses. And thus we can – and do – set ourselves goals and choose our acts with an eye toward how we are most likely to promote those goals. Uncertainty need not lead to paralysis. (Of course it remains true that there will always be a very small chance of some totally unforeseen disaster resulting from your act. But it seems equally true that there will be a corresponding very small chance of your act resulting in something fantastically wonderful, although totally unforeseen. If there is indeed no reason to expect either, then the two possibilities will cancel each other out as we try to decide how to act.)[47]

The cancellation postulate might show that we, according to utilitarianism, can know some act to be *rational* to perform in a situation – that for matters of acting rationally, we can ignore unknowable consequences of our acts.[48] However, this is beside the point in the current discussion, as the cancellation postulate is still not useful for defending the claim that we can come to know that an act maximizes pleasure minus pain. And that is what we need for utilitarianism to be doxastically guiding. There is no reason to believe that the consequences of our acts are so well-behaved that unforeseen pleasure resulting from an act will perfectly balance out unforeseen pains resulting from the same act. And even if the unforeseen pleasure and pain produced by an act roughly cancel each other out, this is not enough for the agent to know that the act maximizes pleasure minus pain. As I noted before, very little is required for an act to be wrong instead of right according to utilitarianism; even a one-second episode of anxiety might tip the scale in the other direction. What we need is not *rough* equality between positive and negative unforeseen consequences, but *perfect* equality. It is unlikely that such perfect equality exists, and even if it does exist, we are not in a position to know about it. So we should reject the cancellation postulate.

Finally, let me consider another argument often used to defend utilitarianism against various challenges: that of advancing a decision procedure or set of second-level moral rules for *indirectly* applying utilitarianism.[49] For example, William H. Shaw suggests that:

> secondary rules help utilitarians deal with the future-consequences-are-hard-to-foresee problem. Whatever action we choose to perform, it will be impossible to foresee its full and exact causal ramifications. We will almost always be ignorant of some of the immediate and intermediate consequences of the choices we make. Moreover, those choices have causal effects that continue indefinitely into the future and are exceptionally hard to discern because criss-crossed by other events. Given these uncertainties, it is possible that doing something truly dreadful, like running over a pedestrian who calmly walks in front of my car while I am stopped at a red light, might somehow have good

[47] Kagan (1998), pp. 64-65. For discussion, see Lenman (2000), pp. 351-359, Mason (2004), Dorsey (2012), Elgin (2015), and Greaves (2016), pp. 315-316.

[48] Cf. Norcross (1990) and Mason (2004). For a reply to Mason, see Lang (2008).

[49] Frazier (1994), pp. 49-52. See also Gren (2004).

> results. (Perhaps she will soon die anyway in a yet more terrible way. Perhaps if she lives, she will have evil children who cause the world great harm.) But past human experience teaches that running over the pedestrian is exceedingly unlikely to maximize long-term happiness. This is a precept or "intermediate generalization" on which I can safely rely. I don't need to study the situation further or speculate about remote and unlikely possibilities.[50]

For the purpose of arguing that utilitarianism is doxastically guiding, this strategy does not work. First, we would have to know which decision procedures and moral rules are such that using them for decision making will maximize pleasure minus pain, and this is at least as difficult as coming to know of any *particular* act that it maximizes pleasure minus pain. Second, to look at a rule's existing track record of producing pleasure and pain is of no help to us: because to evaluate its track record, we have to evaluate the consequences of earlier acts conforming to the rule. Yet we cannot do this, since we do not know for any previously performed act conforming to the rule whether it maximizes pleasure minus pain. As a way to see this, note that any act performed throughout history is *still* having significant consequences, and will keep having such consequences for a long time.[51] Of course, none of this suggests that considering rules and decision procedures is of no help to the hedonistic act utilitarian. An appeal to rules and decision procedures will let her answer the (confused) objection that a utilitarian can only decide what to do on the basis of detailed time-consuming calculation. Nevertheless, the appeal does nothing to show that utilitarianism is doxastically guiding for us.

2.6 Is Utilitarianism Evidentially Guiding?

Is utilitarianism evidentially guiding for us? I earlier defined evidential guidance as follows:

> EVIDENTIAL GUIDANCE
>
> A moral theory *M* is *evidentially guiding* at a time *T* for an agent *S* with respect to a choice situation *C* of the agent if and only if:
>
> (i) there is an act *A* in *C* that *S* knows at *T* how to perform and that is such that were S both (a) not to have any reason at *T* to believe that *A* is right according to *M* and (b) think about which act is right according to *M*, then *S* would on this basis come to have reason to believe that *A* is right according to *M,* and
>
> (ii) *S* has the necessary time and cognitive ability at *T* to think about what is right in *C* according to *M*.

[50] Shaw (1999), p. 146.

[51] Frazier (1994), pp. 46-47.

Since condition (ii) of doxastic guidance is satisfied for us, the same goes for condition (ii) of evidential guidance, as these conditions are identical. Therefore, to determine whether utilitarianism is evidentially guiding for us, we need merely to evaluate whether condition (i) is satisfied for us.

In evaluating whether condition (i) is satisfied for us, it is helpful to note that this condition is very weak, in the following two respects. First, having reason to believe that *p* does not imply believing that *p*. Second, having reason to believe that *p* is compatible with *also* having reason – stronger, weaker, or equally strong – to believe that not *p*. That is, I am understanding "reason for belief" as *pro tanto* reason for belief, rather than as *all-things-considered* reason for belief.

Can we come to have some reason to believe that an act maximizes pleasure minus pain? I think we can. For us to have such reason, I would argue, it is enough that among one or more acts that we are considering at a given time, and among the pleasure and pain produced by those acts which we are aware of at that time, we have reason to believe that one of the acts produces at least as much pleasure minus pain as any of the other acts. That is, to obtain a reason to believe that an act maximizes pleasure minus pain, we need not look to *all* our available alternatives in a choice situation, or to *all* the consequences in terms of pleasure and pain. It is sufficient to look to our immediately introspectively accessible, available information at the time – that is, to the considerations that we at have "in mind."

To give an example, suppose that I am sitting down one evening after work. I suddenly become very thirsty, and I consider whether to drink some wine or beer. I am aware of how drinking either of the beverages will give me some pleasure, but also of how drinking wine will give me slightly more pleasure than drinking beer (I like wine more than beer). In this case, I would argue, I have some (very weak) reason to believe that drinking wine maximizes pleasure minus pain.

Of course, if I start to think about additional acts available to me in the choice situation, and additional consequences for pleasure minus pain arising from performing those acts, then I may come to have reason to believe that *another* act maximizes pleasure minus pain, and may consequently cease to have any reason whatsoever to believe that drinking wine maximizes pleasure minus pain. Nevertheless, no fancy calculations are needed at any point in this process. To obtain a reason to believe that an act maximizes pleasure minus pain, I do not need to go beyond my immediately accessible information at the time.

If I am right, it should be easy to see how similar arguments can be made for nearly all human agents, and with respect to nearly all ordinary choice situations that we are faced with. At nearly any time and with respect to nearly any choice situation, we can think about which act is right according to utilitarianism, and on this basis come to have (again, very weak) reason to believe that an act is right according to the theory. Therefore, condition (i) is satisfied, and utilitarianism is evidentially guiding for us

Now, one could object that in the case that I described I actually have *no* reason *whatsoever* to believe that drinking wine maximizes pleasure minus pain – not even a very weak reason in the *pro tanto* sense which I am here considering. But this strikes me as unreasonable. To see this, consider the following analogy. Suppose that you work in a warehouse that stores 1000 barrels of oats and barley. You have no idea of the proportion of oat and barley in each barrel, and you have no relevant background information bearing on this matter, such as how oats and barley are usually distributed among barrels. You now open two barrels – 219 and 517 – and learn that of these two barrels, 219 has more oats minus barley on the thin visible top layer than does 517. Again, you know nothing about the other barrels in the warehouse, and you are still ignorant of what lies beneath the top layer of 219 and 517. At this point, the manager arrives and informs you that he must send the barrel with the most oats minus barley to a customer. He realizes that you do not know which barrel in the warehouse has the most oats minus barley, and that you have not inspected any of the other barrels. But he still asks you to "take your best guess." In this situation, it seems rational for you to guess that barrel number 219 has the most oats minus barley – it would not be acceptable for you to flip a coin. But if it is rational for you to guess that 219 has the most oats minus barley, then you must have *some* reason, however weak, to believe this. Otherwise, it *would* be acceptable for you to flip a coin. Similarly, suppose that I have considered only the pleasure and pain produced by drinking the wine or beer, and that I have to take a guess of which act is right according to utilitarianism. In this situation, it seems rational for me to guess that drinking wine is right according to utilitarianism – again, it would not be acceptable for me to flip a coin. But if it is rational for me to guess that drinking wine is right according to utilitarianism, then I must have *some* reason, however, weak, to believe this. As a result, I think it is reasonable to say that I can come to have some reason to believe that an act maximizes pleasure minus pain, as long as we remember that this reason will only be a very weak *pro tanto* reason for belief.

Let us now take a step back, and return to the overall project of defining two kinds of action guidance. We can now see why for a moral theory to be evidentially guiding for an agent is for it to, in a sense, be *practically relevant* for that agent. When a theory is evidentially guiding for an agent in a choice situation, it must be possible for that agent to identify *some* consideration in the situation which is relevant for decision making according to the theory. Furthermore, we can see that it is not a trivial achievement for a moral theory to be evidentially guiding, and so practically relevant. For example, theories that use vague language or unfamiliar technical terminology might fail to even be evidentially guiding for us. That utilitarianism is evidentially guiding for us might account for why it strikes us as highly relevant to answering various moral questions – and might explain why utilitarians themselves have not shied from engaging in all sorts of practical inquires, such as questions about abortion, euthanasia, and animal ethics. While utilitarianism makes it very difficult to learn what is morally right or wrong according to the theory, the fact that what is morally relevant according to utilitarianism – pleasure and pain – is so easily identifiable makes it very easy to "get started" with reasoning about what is morally right or wrong on the utilitarian theory.

To conclude, I began this chapter by suggesting that we think of a demand for action guidance as a demand for a theory to be *helpful* to us. We can now see that as a moral theory, utilitarianism is in different senses both very unhelpful and very helpful. It is in one sense very unhelpful because *so much* matters according to the theory, too much for us to grasp with our limited cognitive abilities, and as a result utilitarianism is not doxastically guiding for us. In another sense, however, the theory is very helpful because it is easy to identify in every choice situation *something* that matters according to the theory, and as a result utilitarianism is evidentially guiding for us. While utilitarianism cannot teach us what to do, it is nearly always of practical relevance. As we shall see in the next chapter, this matters for what kind of action guidance objections that we can successfully direct against the theory.

3. Action Guidance Objections

With the definitions of doxastic and evidential action guidance in hand, I now turn to evaluating action guidance objections to utilitarianism. In this chapter, I distinguish between several action guidance objections on the basis of their stated conclusions. First, I consider the objection that utilitarianism is a *bad* moral theory. Second, I evaluate the objection that utilitarianism fails to *be* a moral theory. Third, I discuss the objection that utilitarianism is not an *interesting* or *important* moral theory. Fourth, I discuss the objection that utilitarianism is a *false* moral theory, either because it violates the "ought implies can" principle, or because it is incompatible with our ability to gain moral knowledge.

3.1 Setting the Stage

Let me first clarify why and how some arguments fall beyond the scope of this chapter. To begin with, some objections against utilitarianism are not *action guidance* objections, even if they draw upon considerations similar to those that I presented in the previous chapter. In what follows, I will assume that to be an action guidance objection to utilitarianism, an argument must include *either* a premise according to which utilitarianism is not action guiding in some way, *or* a premise that is crucially supported by such a claim. Otherwise, the argument is not ultimately concerned with action guidance. Some objections discussed in the literature fail to satisfy this condition, although they are ostensibly about action guidance. For example, in the following passage James Lenman argues that if utilitarianism is true, then we do not know that the "crimes of Hitler were wrong":

> We have only the feeblest of grounds, from an objective consequentialist perspective, to suppose that the crimes of Hitler were wrong. Here, if anywhere, surely, there is a considered moral judgment at stake that is well-enough entrenched not to be up for grabs in the cut and thrust of reflective equilibrium, a judgment far enough from the periphery of the web of our moral beliefs to furnish a compelling reductio of any theory that might undermine it.[52]

[52] Lenman (2000), p. 349.

Lenman's idea is that if utilitarianism is true, then we are not justified in believing that Hitler's (criminal) acts are wrong; but, the argument goes, we *are* justified in believing that these acts are wrong, and so we can conclude that utilitarianism is false.

A first problem with Lenman's argument is that it concerns *wrongness*. Even if utilitarianism is true, we are still justified in believing that Hitler's (criminal) acts are wrong, because almost *every* act is wrong according to utilitarianism – i.e., only the few optimal acts are right according to the theory. So we have more than "the feeblest of grounds" to believe that Hitler's acts are wrong according to utilitarianism. As a result, Lenman's argument would be more convincing if it was stated in terms of *rightness*. For example, he could argue that we are justified in believing that one of Mother Teresa's charitable acts is right, but that utilitarianism (implausibly) rules such justification out. Alternatively, to adhere more closely to the spirit of his objection, Lenman could state his case in terms of *knowledge*: he could claim that if utilitarianism is true, then we do not *know* that Hitler's (criminal) acts are wrong – that much seems correct.

Suppose that we qualify Lenman's argument in the latter way. Even so, it is still not ultimately concerned with action guidance. To see this, suppose that we extend our definition of doxastic guidance to cover wrongness, and that whenever we think about which act is wrong according to utilitarianism, we will on this basis come to know that one of Hitler's criminal acts is wrong according to the theory. In such a case, the truth of utilitarianism will *still* undermine our *actual* claim to knowing that Hitler's act is wrong. This is, first, because most of us do not have our knowledge about the wrongness of Hitler's acts on the basis of thinking about what utilitarianism says; and, second, because it is implausible that we would, if utilitarianism is true, have this knowledge by other means, such as by direct moral intuition. The latter would (implausibly) require our intuitions to be sensitive to unknown empirical facts about pleasure and pain situated in the far future. So even if utilitarianism is doxastically guiding with respect to wrongness, utilitarianism will still imply that we do not know that Hitler's criminal acts are wrong. Similarly, even if a map is helpful for letting you learn where I live, the map is irrelevant for your *actual* knowledge of where I live *before* you have consulted it. In the light of the above considerations, we should conclude that Lenman's objection is not ultimately concerned with action guidance, but rather with how utilitarianism is incompatible with what we consider to be our actual moral epistemic situation. For this reason, rather than considering Lenman's original argument, I will discuss a closely related objection in section 3.6 that properly counts as an action guidance objection. This is the objection that utilitarianism, because it is not action guiding, is incompatible with our *ability to gain* moral knowledge, rather than with us *having* moral knowledge.

Some objections count properly as action guidance objections, but are still not sufficiently *independent* of other arguments against utilitarianism to merit consideration. For example, Lenman has also argued that our ignorance about the future makes the integrity objection to utilitarianism more pressing; this is the objection that utilitarianism conflicts with how we intuitively are morally permitted to live our lives according to our own ideals and values.[53] Even if Lenman is right, this specific problem relies intimately on the success of the integrity objection; if that objection succeeds, then utilitarianism is done for, and if it fails, then Lenman's objection fails as well. In this text, I focus instead on action guidance objections whose success are not so obviously held hostage to the success of other objections to utilitarianism. In other words, I try to answer this question: Does the fact that utilitarianism is not doxastically guiding for us give us *further* reason to reject the theory, apart from that already provided by other objections?

Another group of objections which is missing from my discussion is that of distinctively *meta-ethical* action guidance objections. For example, perhaps we could argue against utilitarianism by demonstrating that morality is a "human construct" or that it is "determined by humans," and then show how these meta-ethical facts justifies the demand that moral theories should be action guiding.[54] While I will not investigate such arguments in this book, I wish to point out that it is not trivial to make them work – even if we can make the ideas that morality is a "human construct" and is "determined by humans" clear and coherent. For example, intuitively, the idea that morality is "made by humans" seems at most to support that the true moral theory is *practically relevant* to decision making – and for this purpose it is enough that a moral theory is evidentially guiding, which utilitarianism is. We could also argue against utilitarianism by embracing some non-cognitivist meta-ethical theory, according to which moral statements are not even truth-apt – that is, they are neither true nor false. Perhaps such a view can justify a demand that moral theories should be action guiding – for example, it does not seem far-fetched to think that if moral statements are imperatives, then it is the function of moral statements to be action guiding (since it is, we could argue, the function of imperatives in general to "prompt action"), and perhaps we can find a argumentative path from that claim to an objection to utilitarianism. Even so, we might wonder what it even *means* to "object" to a moral theory on non-cognitive meta-ethical views. In particular, we might wonder whether utilitarianism is even a *proposition* on such views, one that can have properties such as being bad, being false, and so on. If utilitarianism cannot be bad, false, etc., how can anything possibly be a "problem" for it? In any case, if action guidance objections are shown to depend on the above kind of controversial meta-ethical assumptions, this makes them weaker as objections to utilitarianism, as they will now inherit any problems that face their underlying meta-ethical views. So

[53] Lenman (2000), pp. 367-370.

[54] Cf. Jonas Gren's discussion of constructivism and the requirement that utilitarianism is action guiding in Gren (2004), pp. 116-136.

even if I do not address the above kind of meta-ethical objections in this book, there will be something gained by learning whether there are successful *non*-meta-ethical action guidance objections to utilitarianism – objections which do not carry any burdensome meta-ethical baggage. If there are no such objections, that is good news for utilitarianism.

3.2 The Badness Objection

In what follows, I distinguish action guidance objections from each other by considering their stated conclusions. Most importantly, I will need to consider objections stating that utilitarianism *is false* – to conclude that a theory is false is the most obvious way of objecting to it. That is the task for Sections 3.5 and 3.6. However, before considering these arguments, I want to examine other potential conclusions of action guidance objections. One such conclusion is that utilitarianism is *worse* as a moral theory in virtue of not being action guiding, whether or not it is also false. That is, we could argue that if a theory is not doxastically guiding, then it is a *bad* moral theory, where the badness is *pro tanto* rather than *all-things-considered* badness. (Of course, a theory could still be *all-things-considered* good in virtue of having various positive qualities, even if it is *pro tanto* bad in virtue of not being action guiding.) The above kind of view is suggested by Pekka Väyrynen, whose "guidance constraint" states that "[o]ther things being at least roughly equal, ethical theories are better to the extent that they provide adequate moral guidance."[55]

A complication for any version of a badness objection to utilitarianism is that "badness" can mean very different things depending on the relevant context. The challenge is therefore to find an interpretation of "bad" that *both* makes it plausible to think that utilitarianism is bad in virtue of not being doxastically guiding *and* that counts as a proper "objection" to the view. In other words, the conclusion of the badness argument has to be properly "problematic" for utilitarianism.

First, that a moral theory is "bad" could mean that there is evidence against the theory, or that we have reason to believe that the theory is false. This interpretation gives us a clear objection to utilitarianism. However, the objection is also subservient to the more fundamental question of whether the theory is true. In this case, we will do better to directly consider objections against the truth of utilitarianism. (Compare: It is more convenient to argue that knowledge is not justified true belief, than to argue that we have reason to believe that knowledge is not justified true belief.) Having supported an objection to the truth of utilitarianism, we will have supported a badness objection against it; conversely, if we cannot support an objection to the truth of utilitarianism, then neither can we support a corresponding badness objection

[55] Väyrynen (2006), p. 292.

against it. On this first interpretation of "badness," we get a proper objection to utilitarianism, but it is not one that we need to evaluate separately from any objections to its truth. Again, I will deal with objections to the truth of utilitarianism in Sections 3.5 and 3.6.

Second, some philosophers have argued that it is the *function* or *point* of moral theories to be action guiding. This might provide us with a version of the badness objection.[56] For example, Elenor Mason notes that "the most important function of a moral theory is to guide action."[57] Lars Bergström writes that "moral norms should have practical relevance" because "[t]hat's what they are for."[58] Terrance McConnell says that "one of the important functions of moral theories is to assess the conduct of others" and that "one of the main points of moral theories is to provide agents with guidance."[59] Peter Singer states that "the whole point of ethical judgements is to guide practice."[60] Robert Goodin writes that "[t]he point of morality is to be action-guiding" and Frank Jackson says that "the passage to action is the very business of ethics."[61] A literal interpretation of these claims gives us the following badness objection to utilitarianism. To begin with, we suppose that moral theories have functions, and that one such function is to be doxastically guiding. Next, we draw a comparison to other entities that have functions, such as human artefacts or biological entities. For example, an umbrella that does not protect against rain fails to fulfill its function, so it is *a bad umbrella*; and a heart that does not pump blood fails to fulfill its function, so it is *a bad heart*. Similarly, we can argue, utilitarianism does not fulfill its function, which is to be doxastically guiding, so it is *a bad moral theory*.

The problem with the above argument is that the sentence "the umbrella is bad" seems to simply *mean* that "the umbrella fails to fulfill its function." But in that case, the corresponding objection against utilitarianism concludes only that utilitarianism fails to fulfill its function: this is not by itself a problem for utilitarianism. In other words, in the context of function-talk, to say that something is "bad" does not express something truly or genuinely evaluative, but gives us merely a circumspect way to talk about the functions of various objects.

Third, we can understand the "badness" as "moral badness." In doing so, we will need to argue that utilitarianism is *extrinsically* or *instrumentally* morally bad (in virtue of not being doxastically guiding) because it is deeply implausible that a moral theory is bad in the sense that, for example, pain is bad – that is, moral theories are never intrinsically or finally morally bad.

[56] Cf. Smith (2018), pp. 53-54.
[57] Mason (2003), p. 327.
[58] Bergström (1996).
[59] McConnell (1989), p. 445; (2018).
[60] Singer (1993), p. 2.
[61] Goodin (2009), p. 3; Jackson (1991), p. 467.

In support of the claim that "if utilitarianism is not doxastically guiding, then it is morally bad," we can appeal to the notion of *autonomy*.[62] Holly M. Smith, who was first to discuss autonomy in the context of moral theories being action guiding, says that a person lacks autonomy if she cannot "translate her moral values into a choice of what to do."[63] Similarly, Pekka Väyrynen argues that if moral theories should be action guiding for us "the best explanation of this fact features certain forms of autonomy and fairness," with the latter referring to "fairness in the provision of the opportunity for morally committed moral agents to act well autonomously."[64] It is not clear what exactly Smith and Väyrynen have in mind here. In an earlier text Smith says that a kind of action guidance is "valuable" because it "makes possible an important form of autonomy," but she does not explicitly state that we are dealing with *moral* value.[65] Väyrynen is more forthcoming on the issue:

> [W]e must also be open to the idea that if ethical theories are better to the extent that they provide adequate moral guidance, then the best explanation of that fact features some morally substantive ideals.[66]

Väyrynen's reference to "morally substantive ideals" suggests that we are dealing with specifically moral value. In any case, I will assume this for the sake of the argument. Clearly, Smith and Väyrynen cannot be concerned with the *truth* of non-guiding theories, since that requires a premise to the effect that we *are* in fact autonomous (which is why the true moral theory must make possible such autonomy), and that seems implausible. Moreover, it is not clear what kind of non-moral badness could be intended.

Let us assume that the concern is about moral value: that utilitarianism is a morally bad theory because it does not allow us to be autonomous, and that it does not allow us to be autonomous because it is not doxastically guiding. Now, there are two interpretations of autonomy that we could have in mind when talking about the "autonomy" of agents. On the first interpretation, for an agent to be autonomous is for her to be able to act according to her *own* values. As I have argued elsewhere, this interpretation does not seem to be problematic for utilitarianism, because most people *are not utilitarians*.[67] Therefore, that these people cannot use utilitarianism to guide their actions will not prevent them from being autonomous and able to act according to their own (non-utilitarian) values.

[62] See Smith (1988), pp. 105-106; (2018), pp. 55-56, 196-202 and Väyrynen (2006). For discussion, see van Someren Greve (2014).
[63] Smith (2018), p. 219.
[64] Väyrynen (2006), pp. 297-301.
[65] Smith (1988), p. 105.
[66] Väyrynen (2006), p. 307.
[67] Rosenqvist (2019), p. 355.

On the second interpretation, for an agent to be autonomous is for her to be able to act according to the *true* values, whatever they are. But this interpretation is not a problem for utilitarianism either. On the one hand, if utilitarianism is true, then its moral badness will depend on how conducive it is to produce pleasure minus pain – something that we, because of reasons familiar at this point, cannot know.[68] On the other hand, if utilitarianism is false, then its inability to guide our actions will not stop anyone from acting according to the "true values" – since in that case utilitarianism will not give us the true values.

Finally, it is far from clear that we cannot live our lives according to utilitarian values – which presumably means to live one's life, in some sense, *according to the utilitarian theory* – just because utilitarianism is not doxastically guiding. It is still possible to let our decision making be *informed* by utilitarianism. Since the theory is evidentially guiding for us, we can think about what it says and come to have reason to believe that an act is right according to it. This may appear sufficient to let us live our lives according to the utilitarian theory.

3.3 The No Moral Theory Objection

Consider next the objection that utilitarianism, because it is not doxastically guiding, is not a *moral theory.* I will examine various versions of this objection, but the intuitive idea is straightforward. Suppose that we ask what happiness is, and that in response we are told a detailed story involving various neurological facts. In this case, and assuming that physicalism about consciousness is false, the response fails to address our question. Even if we learn many or even all neurological facts, we will not learn what happiness is. The responder unacceptably *changes the subject* from the philosophical question of what happiness is, to the biological question of what happens in the brain when we are happy. Likewise, we may suspect that when someone responds to moral questions by proposing a normative view that is not action guiding, she similarly unacceptably "changes the subject." The proposed objection is not that the offered view is *bad* or *false*, although that could still be true for other reasons, but that it is *out of place* or *irrelevant*.

Now, for the no moral theory objection to succeed it is not enough that utilitarianism employs a *different concept* of moral rightness than do other moral theories. In that case, utilitarianism will still be a moral theory, just one that is formulated using a different concept of moral rightness. That is, the

[68] Some people might think that a theory cannot "produce" pain or pleasure, because theories are not (they believe) causally efficacious. I do not share this concern; however, if a moral theory cannot produce pleasure or pain, then that makes the argument under consideration *even worse*. We would have to conclude that, if utilitarianism is true, it is *neither* good nor bad, because it *cannot* in such a case be extrinsically or instrumentally good or bad.

subject matter will not have been unacceptably changed from a moral to a non-moral topic. Consider, for example, Jan Österberg's suggestion that:

> Since consequentialism is not action-guiding […] its deontic notions are not conceptually related to the other moral notions. This means that its use of, for example, "right" and "wrong" are quite different from that of ordinary moral thought. […] But this, in turn, means that consequentialism is not a competitor to common-sense morality even in the latter's role as a theory of right- and wrong-making criteria: their criteria do not concern the same deontic notions.[69]

Österberg's idea is that while utilitarianism employs *one* concept of moral rightness, this is not the *same* concept of moral rightness as those which are employed by other moral theories. More generally, the idea is that theories that are doxastically guiding employ different rightness concepts than moral theories that are not doxastically guiding. However, the above argument is not problematic for utilitarianism, because as long as their theory is about *some* moral deontic concept of rightness, it has not radically changed the subject. Utilitarianism is in such a case still a moral theory; it is still about morality. For example, perhaps there is a distinction between subjective and objective moral rightness, where only moral theories that are doxastically guiding can successfully employ the notion of subjective rightness. In that case, utilitarianism can still be a theory about objective rightness (which presumably does not require a theory to be doxastically guiding in such a way) and thereby qualify as a moral theory.

Another approach is to argue that the utilitarian *proposition*, while it correctly employs the concept of moral rightness, does not deserve the *label* "moral theory."[70] However, even if this claim was plausible, it would make the resulting objection toothless. Utilitarians will respond that even if their view does not deserve the name "moral theory," it is nevertheless a true view about moral rightness. Whether utilitarianism *classifies* as a moral theory is beside the point. What matters is that utilitarianism is about moral rightness and that it is true. Moreover, utilitarians will take issue with the claim that their theory does not deserve the label "moral theory." Surely, they will argue, if a theory gives us an explanation of moral rightness, then it *is* a moral theory, and therefore utilitarianism *does* deserve this label. This illustrates how, to get a proper version of this objection to utilitarianism, the objection must relate to the *content* of utilitarianism, and not just to the mere *labelling* of it.

Consider then what I think is the most interesting version of the no moral theory objection. It goes as follows. Because utilitarianism fails to even employ a *deontic and moral concept*, it radically changes the subject. The rightness concept it employs could, for example, be a purely evaluative concept, such as goodness or badness. For example, Vuko Andrić suggests that:

[69] Österberg (1988), pp. 282-283.

[70] Cf. Österberg (1988), p. 282.

> [M]oral theories must make sense of the deontic vocabulary we use in everyday life. Terms like *right*, *wrong*, and *obligatory* are essentially action-guiding, whereas evaluative notions like *good* or *bad* are not. Evaluative properties can properly be ascribed to all sorts of entities, whereas the properties of rightness and wrongness primarily apply to actions or choices. Conceptual analysis thus reveals the action-guiding function of moral theories.[71]

Why believe that these terms – right, wrong, and obligatory – are essentially action guiding? The following remarks by Österberg are helpful:

> In common-sense morality, deontic notions such as *right* and *wrong* are conceptually related to other moral notions, such as *guilt* and *responsibility*, *praiseworthiness* and *blameworthiness*, *being good* and *being bad*. (For example: you cannot be blameworthy, or reasonably feel guilt, for having performed (what you know is) a right action; you are praiseworthy only if you have performed a right action; if many of your actions are wrong, you are hardly a good man.)[72]

For the sake of the argument, I will grant that moral deontic concepts such as right, wrong, and obligatory are in fact essentially action guiding concepts. Even so, a problem remains for the objector. Let us assume that for the concept of rightness to be essentially action guiding is for the following to be true: in any case where an act is right, we can be guided by means of considering the proposition that it is right. (Although my definitions of action guidance do not apply to concepts but only to theories, I assume that a suitable counterpart to doxastic guidance can be invented here). It follows that, if the concept of rightness is essentially action guiding, then any true proposition of the form "x is right" can be used to guide behavior. For example, it cannot be true that "maximizing pleasure minus pain is right," because that proposition cannot be used to guide one's behavior.

However, and this is my concern with the objection being considered, even if the concepts of rightness and wrongness are essentially action guiding in this way, it does *not* follow that *moral theories* are essentially action guiding. The above argument seems to falsely presuppose – similar to the ought implies can objection which I discuss in Section 3.5 – that utilitarianism tells us to carry out a difficult-to-perform act such as "maximize pleasure minus pain." However, utilitarians can deny that their theory tells us to perform such an act. Instead, they can argue that utilitarianism tells us to perform only very easily performed acts, like "drink the glass of water" and "open the door," where these are acts that as a matter of fact maximize pleasure minus pain. This could be achieved by restricting the notion of an act being "available to be performed" in the formulation of utilitarianism, so that it applies only to easily

[71] Andrić (2016), p. 77.

[72] Österberg (1988), p. 282. Cf. Carlson (2002), pp. 72-73 and Andrić (2017), p. 77.

performed acts. In other words, we can be guided by means of considering that the act "drink a glass of water" is right, even if utilitarianism – that *implies* that we should drink a glass of water – is *not* action guiding for us. As a result, it is unclear how to argue from the claim that "right" is essentially action guiding to the claim that moral theories are essentially action guiding. Moral deontic notions can be essentially action guiding, even if utilitarianism is not.

Now, we can reintroduce talk about the function of moral theories at this point in the discussion, not in support of the claim that utilitarianism is a bad moral theory as I discussed earlier, but in support of the view that moral theories are essentially action guiding. For example, we could argue that because the function or role of moral theories is to be action guiding, moral theories (and not just moral concepts of rightness or wrongness) are essentially action guiding. However, even if we manage to drum up support for the claim that moral theories have as their function not only to be *evidentially* guiding, but also to be *doxastically* guiding, this is not a promising strategy. It is easy to see why. In general, just because something has as its function to be in a certain way, it is not thereby *essentially* in such a way. For example, a heart has as its function to pump blood, but hearts are not essentially pumping blood, since a heart that fails to pump blood is still a heart. Similarly, a moral theory that fails to fulfill its purpose to be doxastically guiding can still be a moral theory.

3.4 The Importance and Interestingness Objection

A moral theory can be true, good, and about a concept of moral rightness, yet fail to be *important* or *interesting*. Questions about interestingness and importance often arise in philosophical discussions, although they seldom take center stage. For example, we commonly think that what is right according to the rules of etiquette is less important and interesting than what is morally right. Or suppose that we define a technical term "right*" to *mean* "maximize pleasure minus pain." It would for such a rightness concept be trivially true that an act is right* if and only if it maximizes pleasure minus pain. But such a utilitarian theory is both unimportant and uninteresting, and therefore not worthy of further consideration. Such a theory does not tie into or promise to answer any of our central philosophical concerns. In similar ways, we may think that utilitarianism is not important or interesting (i.e., not worth caring about) if it is not doxastically guiding.

What *is* importance and interestingness? These concepts are fairly undertheorized within philosophy, but at least we can say this much: Whether a proposition *p* is important does not depend on us *finding* it important. For example, even if no one thought that a moral question about euthanasia was important, it would still retain its importance. Importance is "provided by the world" in some sense – it is forced on us by the way things are. In contrast, whether a proposition *p* is interesting is constituted or determined at least partly by our subjective responses towards *p*, so that if something is interesting, it is interesting *because we find it interesting*. Matters of interestingness are thus "up to us." This does not mean that we can change what is interesting merely by changing our wants and preferences, but that, unlike matters of importance, we can change what is interesting by changing our perspective on the world.

Is it true that, because utilitarianism is not doxastically guiding – or action guiding in a similarly demanding sense – it is not important or interesting? James Lenman suggests this much (my emphasis):

> If consequentialism is to be a theory of any real normative *interest*, it must at least furnish us with a regulative ideal to guide our choices either of actions or decision procedures; it must offer such choices a consequentialist rationale.[73]

Similar, consider remarks to the effect that moral theories should be "useful" or fulfill some of our desires or wants. For instance, Krister Bykvist says that a "moral theory seems useless if it can never guide agents when they deliberate about what to do."[74] Mark Timmons writes that "the practical aim of moral theory has to do with the desire to have some method to follow when, for example, we reason about what is right or wrong."[75] Finally, Andrić notes that:

> [W]e want to apply moral theories; we want them to tell us what is the moral thing to do in real-life cases of moral conflict, dispute, and uncertainty, and we want to implement moral theories by doing what they require. Life poses moral challenges; and moral theories are supposed to yield convincing deontic judgements in order to equip us with knowledge about what ought to be done on which, ideally, we can base our decisions.[76]

A plausible background assumption here is that, if a moral theory like utilitarianism is not useful or fulfills our desire to solve moral problems and challenges, which is arguably the case because it is not doxastically guiding, then it is not important or interesting.

However, there is a problem for constructing an action guidance objection to utilitarianism along the above lines. Intuitively, a theory only needs to be important and interesting to *some* degree to be *acceptably* important and interesting. In other words, a theory can be less than *maximally* important and interesting without being *unacceptably* low in interestingness and importance. If this was not so, very few theories in philosophy would be acceptably important and interesting, since few theories are maximally important and inter-

[73] Lenman (2000), p. 360.
[74] Bykvist (2010), p. 14.
[75] Timmons (2013), p. 3.
[76] Andrić (2017), p. 77.

esting. Clearly, however, most philosophical theories *are* acceptably important and interesting. And although utilitarianism is not maximally important and interesting, because it is not doxastically guiding, I will now argue that it is still interesting and important to a very high degree, or at least to an acceptable degree. I will present four reasons to think that utilitarianism is both important and interesting, even though it is not doxastically guiding.

First, even if utilitarianism is not doxastically guiding, the truth of utilitarianism rules out the truth of other moral theories, including moral theories that *are* doxastically guiding. Accordingly, merely by knowing that utilitarianism is true, we can know that a large number of moral theories are false. This qualifies utilitarianism as both important and interesting to a significant degree. That is, many non-utilitarians will have to worry about utilitarianism being true even if the theory is not doxastically guiding, because the truth of utilitarianism rules out the truth of their own favored views.

Second, utilitarianism is both important and interesting in virtue of being *theoretically* important and interesting – that is, in constituting a plausible explanation of rightness and wrongness, which as such needs to be taken seriously. Therefore, utilitarianism's lack of doxastic guidance can be partly compensated by its theoretical plausibility. Moreover, utilitarianism is not theoretically interesting just to academic philosophers, but to anyone trying to seriously explain why an act is right or wrong – a person asks such questions also in her day-to-day life, like when asking: "Why do I have to visit that relative?" or "Why should I give money to the Red Cross?"

Third, utilitarianism is a *provocative* theory, which makes it more interesting, although not more important. Utilitarianism hints that moral reality may be very different than what common-sense tells us, and this kindles our interest. Of course, utilitarianism provokes us mainly by having us consider thought experiments where we stipulate the levels of pleasure minus pain produced by acts. However, by judging from people's reactions to the theory's implications in such cases, this does not seem to make utilitarianism any less provocative.

Fourth, as we have seen, utilitarianism is evidentially guiding as it is practically relevant for decision making. If utilitarianism is true, then in nearly all choice situations you can learn *something* about what reason you have to believe that an act is right by thinking about what utilitarianism says. This makes the theory more important and interesting to us.

Taken together, the above four considerations suggest that while utilitarianism is not maximally important and interesting, it is at least acceptably so. We should therefore reject the importance and interestingness objection.

3.5 The Ought Implies Can Objection

The final two objections that I will discuss in this chapter are objections to utilitarianism in the classic sense: they are arguments that utilitarianism is *false*. To begin with, Frances Howard-Snyder argues that utilitarianism is incompatible with the principle that "ought implies can."[77] Her argument, reconstructed with premises and a conclusion, goes as follows:

> THE OUGHT IMPLIES CAN OBJECTION
>
> (P1) You cannot maximize pleasure minus pain.
> (P2) If you cannot maximize pleasure minus pain, then it is not the case that you ought to maximize pleasure minus pain.
> (P3) If it is not the case that you ought to maximize pleasure minus pain, then utilitarianism is false.
> (C) Utilitarianism is false.

The conclusion of the argument follows from the premises, so let us consider each premise in turn. To begin with, premise (P1) is supported by a comparison to acts that we intuitively cannot perform. For example, you cannot defeat the Hungarian grandmaster Judith Polgár in a game of chess or open a safe to which you lack the code. In both of these cases, you lack the ability to perform the act – i.e., you cannot do it. Various explanations are available for why one suffers an inability to perform an act – although, naturally, giving such an explanation is not necessary to *justify* that one has the inability, since this matter can be judged on intuitive grounds. Some examples of potential explanations of inability include: that you cannot perform an act because you do not *know how* to perform it, because you *would fail to perform it if you tried*, or because it is *too difficult* for you to perform it. Whichever explanation is correct, Howard-Snyder is right in saying that if you cannot defeat Polgár or open the safe, then you cannot maximize pleasure minus pain. Any explanation that applies to the former claims must also apply to the latter. For example, you do not know how to maximize pleasure minus pain, you will fail to maximize pleasure minus pain if you try, and maximizing pleasure minus pain is too difficult for you. That is, if you cannot defeat a chess grandmaster, how could you hope to maximize pleasure minus pain?

As for premise (P2), it is plausible in virtue of the principle that ought implies can. The ought implies can principle states that if you ought to perform an act, then you can perform that act. That is, an obligation to φ presupposes an ability to φ. For example, it cannot be true that I ought to pick you up at the

[77] Howard-Snyder (1997). Responses to Howard-Snyder include Carlson (1999) and Qizilbash (1999), to which Howard-Snyder (1999) has replied. For further discussion, see Mason (2003), Miller (2003), pp. 53-54, Moore (2006), and Andrić (2016). Bergström (1996) anticipates this discussion.

airport if I cannot pick you up. That is, if my car is being repaired or the airport is too far away, then I have no obligation to pick you up. The ought implies can principle is equivalent to the proposition that if you cannot perform an act, then it is not the case that you ought to perform the act, which straightforwardly gives us premise (P2).

Finally, premise (P3) is meant to be plausible in virtue of the definition of utilitarianism. If utilitarianism is true, the argument goes, you ought to maximize pleasure minus pain. Therefore, by *modus tollens*, if it is not the case that you ought to maximize pleasure minus pain, then utilitarianism is false.

Although Frances Howard-Snyder does not intend the ought implies can objection to constitute an *action guidance* objection to utilitarianism, it qualifies as such in this discussion.[78] While a notion of action guidance is not directly appealed to in the objection, whether utilitarianism is doxastically guiding is relevant to evaluating premise (P3). Had utilitarianism been doxastically guiding for you, you would in many circumstances have the ability to maximize pleasure minus pain: namely by using utilitarianism to find out which act is right according to the theory – an act which will have the property of maximizing pleasure minus pain.[79]

The main problem with the ought implies can objection becomes clear when we consider how to exactly understand premise (P1).[80] That you "cannot maximize pleasure minus pain" can be interpreted in either of the following two ways, and the ought implies can objection is unsuccessful on either interpretation:

> (A) You cannot perform an act that has the property of maximizing pleasure minus pain.
>
> (B) You cannot perform the act "maximize pleasure minus pain."

On the A-interpretation of "cannot maximize pleasure minus pain," premise (P1) is false, since I *can* perform an act that has the *property* of maximizing pleasure minus pain. In any choice situation, there is an easily performed act such as "drink a glass of water" or "open the door" that has the property of maximizing pleasure minus pain.[81] Clearly, I can perform some such act – I

[78] Cf. Howard-Snyder (1997), pp. 241-242.

[79] This is only true if we know what utilitarianism says. The farmer who does not know what utilitarianism says would still be unable to maximize pleasure minus pain.

[80] Several authors have, in various ways, noted this problem with the ought implies can objection. See Bergström (1996), Carlson (1999), pp. 93-95, Moore (2006), pp. 84-87, and Andrić (2016), pp. 71-74.

[81] I am assuming a fine-grained theory of act-individuation, according to which "drinking a glass of water" and "maximize pleasure minus pain" are different acts, even if drinking a glass of water is a way *by which* you maximize pleasure minus pain. On a coarse-grained view, these are one and the same acts, and the discussion would have to be presented in a slightly different

can drink a glass of water, open a door, or do some other simple act whose performance will maximize pleasure minus pain. Likewise, while I cannot perform the act "defeat Judith Polgár," I *can* perform an act that would make me defeat Judith Polgár: this is just a long conjunctive act such as "move E4, D3, etc." Few of us can defeat Polgár, but most of us can "move E4, D3, etc."

On the B-interpretation of "cannot maximize pleasure minus pain," premise (P1) states that you cannot perform a specific act – the act "maximize pleasure minus pain." However, if you cannot perform the act "maximize pleasure minus pain," then it is not among your available alternatives, and so utilitarianism does not tell you to perform it. In this case, we should instead reject premise (P3):

> (P3) If it is not the case that you ought to maximize pleasure minus pain, then utilitarianism is false.

This is because utilitarianism can be true even if it is not the case that you ought to perform the act "maximize pleasure minus pain." As a result, the ought implies can objection fails on either interpretation.

3.6 The Epistemic Objection

According to the epistemic objection, utilitarianism is incompatible with our *ability to gain moral knowledge*. Several authors have raised this objection in connection to the ought implies can principle. For example, H. J. McCloskey writes that "[w]e are not truly free to do the right or obligatory act if we cannot in advance know what is the right act. 'Ought implies can', and 'Can' implies 'Can know'."[82] Similarly, Lars Bergström notes that "'Ought' implies 'can', and 'can' implies 'knows how'."[83] McCloskey and Bergström appear to consider these conceptual truths; while others indicate that the problem is that we *intuitively* have the ability to obtain moral knowledge, but that utilitarianism rules this out.[84] With this in mind, consider the following argument against utilitarianism:

> THE EPISTEMIC OBJECTION
>
> (P4) I can come to know of at least some available acts that they are morally right.

way. The coarse-grained theory does not make life easier for the objector, however: I can clearly drink a glass of water, and if that act is *the same* act as the act of maximizing pleasure minus pain, then I can maximize pleasure minus pain as well.

[82] McCloskey (1973), p. 62.

[83] Bergström (1996).

[84] See Frazier (1994) as well as Lenman's comments cited in the beginning of the chapter.

(P5) If utilitarianism is true, then I cannot come to know of any available acts that they are morally right.
(C) Utilitarianism is false.

Unlike the argument by Lenman which I discussed briefly in Section 3.1, the epistemic objection is properly about action guidance, and so an action guidance objection. As is the case with the ought implies can objection, doxastic guidance plays an indirect role in evaluating the premises of this argument. If utilitarianism had been doxastically guiding, we could have used it to come to know that some available acts are morally right – and in such a case, premise (P5) would have been false. However, because utilitarianism is not doxastically guiding for us, we lack this tool for ascertaining what is morally right. So the question of whether utilitarianism is doxastically guiding for us is relevant to evaluating premise (P5).

The conclusion of the objection follows from the premises. What can we say about the premises? Premise (P4) can be supported in either of two ways. First, we can argue that it is a conceptual truth that, if an act is morally right, then we can know it is morally right. Granting that at least some of my available acts are morally right, it follows that I can come to know that some of my available acts are morally right, which is enough to support (P4). Second, we can point to cases where we intuitively have an ability to gain moral knowledge. For example, it is intuitively true that I can come to know that brushing my teeth this evening is morally right, that drinking coffee right now is morally right, or that riding my bicycle to work is morally right. These acts are clearly morally right, and it should be easy for me to figure this out. This too supports (P4).

As regards premise (P5), it is plausible in virtue of my previous discussion of how our acts have chaotic and wildly unpredictable effects on the future – that it is impossible to learn which of our acts maximize pleasure minus pain. That said, premise (P5) must be understood in a restricted way: for it to be plausible, we must focus on actions *in the real world* and not on actions *in imagined cases*. This is because in thought experiments we may come to know which acts are morally right according to utilitarianism by *stipulating* that they maximize pleasure minus pain. Nevertheless, this is not a serious restriction on the argument, because (P4) remains plausible even if we consider only actions in the real world.

Mark Lucas has suggested two ways in which to reject premise (P5), both of which are arguments for us having alternative means for knowing that an act is morally right, other than simply thinking about what the true moral theory says. Interestingly, these arguments may work even if utilitarianism is not doxastically guiding.

First, Lukas notes that we could argue for a foundationalist epistemology, according to which “some moral beliefs are so obviously true that they are in effect self-justifying."[85] Just like I can know that I have two hands, as in G. E. Moore's famous argument, I can know that tooth brushing is morally right, and this is so even if I cannot use the true moral theory to find out that tooth brushing is morally right. In this case, I would come to have this knowledge by direct intuition. Second, Lukas suggests that we can adopt a "hybrid coherentist/externalist epistemology about the justification of beliefs about certain acts."[86] On this epistemological view, we would first on a sub-conscious or semi-conscious level note various features of an act. These observations will give rise to a moral intuition that the act is right, wrong, obligatory, etc. and would also justify the content of this intuition. He writes that:

> This conclusion is based on limited information. It is based just on the features of [the] act that I have directly observed or that I have inferred to exist. And it is based on beliefs and moral intuitions all of which are subject to revision should more information arise. Nevertheless, it is in virtue of such a process, I think, that I form my belief and through which my belief gets its justification.[87]

I believe that neither of these two arguments against (P5) are plausible on closer examination – although to be fair, Lukas merely sketches these responses and is not strongly endorsing either of them. Both of the responses share a problem that is anticipated by Lukas in his discussion of the first view, that "such a view would turn us in to soothsayers of a sort."[88] To see the problem, suppose that I know that utilitarianism is true and that tooth brushing is morally right. On the basis of these two pieces of knowledge, I can come to know that tooth brushing produces at least as much pleasure minus pain as any alternative act available to me. In this case, knowledge of two purely philosophical facts would give me knowledge of a purely empirical fact – and an empirical fact that is not knowable by means of employing our best instruments of observation and theories of science. That we can obtain empirical knowledge in such a way, by means of mere armchair philosophy, seems deeply implausible.[89] Presumably, such soothsaying is impossible because when our moral intuitions work with empirical information, they are limited to considering the empirical information which we are already aware of.

Instead of challenging premise (P5), I propose that the utilitarian challenge premise (P4). The same argument that is used to argue that utilitarianism is not doxastically guiding (i.e., that our acts have wildly unpredictable chaotic

[85] Lukas (2008), p. 6.
[86] Lukas (2008), p. 7.
[87] Lukas (2008), p. 8.
[88] Lukas (2008), p. 6.
[89] Importantly, this argument does not presuppose that knowledge is closed under implication or known implication. Even if it is not always true that knowing P and knowing P→Q enables us to know that Q, it still seems plausible to think that knowing that utilitarianism is true and knowing that an act is right would enable us to know that this act maximizes pleasure minus pain.

effects) also makes it plausible to think that we do not know which acts are morally right even in ordinary choice situations. Surprisingly weak claims are needed for this argumentative strategy to work. We need only defend the claim that a few types of significant consequences for human health or well-being matter for whether an act is morally right. Since we do not know of any ordinary act that it will or will not have such significant consequences, we do not know of any ordinary act whether it is morally right. Moreover, there are good candidates for what may constitute such significant consequences. For example, if an act brings about a million deaths in a thousand years, that seems to matter for whether it is morally right. But for any act available to us, we cannot know whether it will bring about a million deaths in a thousand years. To give a more specific example, suppose that you will live for thousands of years because of future advances in medicine. As you witness massive death and destruction in the year 3019, you trace this calamity back to an evening in 2019, where you stood in your bathroom with toothbrush in hand. Apparently, you learn, had you not brushed your teeth that evening in 2019, then a million lives would have been saved a thousand years later. Everything else being equal, brushing your teeth was surely morally wrong in this situation. One reason to think that it was wrong is that it is fitting to *regret* your choice of brushing your teeth, and the reason to think that regret is fitting is surely that standing there a thousand years later you realize that brushing your teeth was, in fact, morally wrong. Presumably, you would not in the year 3019 say that "Yes, I caused a million deaths, but since I did not *know* that I caused a million deaths, what I did was nevertheless right."

Now, we might have different views on how demanding the true moral theory is. If a moral theory tells you to give all your money to charity and spend your whole life helping others, then perhaps the theory is too demanding. We might also think that in some cases a lack of knowledge excuses certain kinds of behavior, especially in cases where little is at stake. But the case under consideration is one where the theory demands very little of you, and where a lot is at stake: to save a million lives, you need only to put down your toothbrush. In such a situation, putting down your toothbrush is clearly what you should do, regardless of your epistemic situation at the time. You do not deserve any blame for brushing your teeth, of course, because of your ignorance about the future. But to put it down is nevertheless obligatory. To sum up, what the wildly unpredictable chaotic effects of our acts show is not *only* that utilitarianism is not doxastically guiding for us, but *also* that each of our acts has a large potential to sooner or later affect the lives of a million people in very significant ways, pertaining to intuitively salient moral dimensions such as those of life, death, happiness, suffering, and freedom. This fact should, regardless of our confidence in utilitarianism, make us reject (P4) and along with it the epistemic objection.

Let me take stock of the argumentation in this chapter. I have examined a number of action guidance objections, and can conclude that none of them looks promising. But I admit that my discussion may still fail to satisfy the action guidance objector. Many of us have an inner voice telling us that there is something *wrong* with utilitarianism because it is not action guiding, and that voice is not always silenced by being given a comprehensive list of failed objections, as I have tried to do. The situation seems similar to the difficulty moral realists have in convincing people that there are non-natural irreducible moral facts. Regardless of which arguments are offered in defense of the realist position, many feel that the theory *must* be false and that there simply *has* to exist a better objection to it. Perhaps the best action guidance objection to utilitarianism is just that *intuitively*, if a moral theory is not doxastically guiding, then it is false. Since utilitarianism is not doxastically guiding, it is therefore false. However, even if we have the intuition that the true moral theory is doxastically action guiding, this intuition seems to give us reason to believe in its content *only if* there is a plausible *explanation* for why its content would be true. It is not like the intuition that pain matters morally – an intuition that seems trustworthy regardless of whether its content can be explained or not. But as my discussion above shows, it is difficult to find an explanation for why the true moral theory would be action guiding. With this note, I end my discussion of action guidance. At the end of the day, it is not the utilitarian's job to make her opponents' objections work for them. If the objector wants an action guidance objection against utilitarianism that succeeds, she needs to find better objections.

4. Moral Intuitions

Utilitarianism conflicts with strongly held moral intuitions, such as intuitions about what is morally right and wrong, and these conflicts give us reason to believe that utilitarianism is false. In this chapter, I first explain what moral intuitions are and why they give us reason to believe in their content. Next, I introduce three intuitive objections to utilitarianism based on the following three thought experiments: Experience Machine, Transplant, and Utility Monster. Finally, I consider three responses to these objections, and show how each of them is unsuccessful. These discussions set the stage for Chapters 5 and 6, where I discuss two more promising replies on behalf of the utilitarian, both of which, in different ways, concern the role that imagination plays in thought experimentation.

4.1 Intuitions and Moral Theories

Moral intuitions are tacitly, if not explicitly, treated by us as giving reasons to believe in moral claims. For example, we say that it *seems* wrong to lie on a job application and that it *appears* that we should prioritize the welfare of our children. In some of these cases, we report having a moral intuition and take that intuition to support our position. Moreover, when we face novel moral questions we often stop and wait for a moral intuition to surface, and we seldom proceed without having elicited at least some intuitions to rely on.

Even if moral intuitions play this central role in our moral thinking and reasoning practices, that does not mean that they are *epistemically significant*, where an intuition is epistemically significant if and only if it gives the intuiter defeasible reason to believe in its content.[90] I will later suggest two reasons to think that intuitions are epistemically significant. But for now, let us think about how intuitions can be employed in the evaluation of moral theories *if* this is so.

In general, there are two ways in which moral intuitions can matter for the evaluation of moral theories. On the one hand, moral intuitions can *support* a theory and so give you reason to believe that the theory is true. For example, suppose that you morally intuit that giving to a specific charity is right and

[90] This way of phrasing the issue leaves open whether it is the *intuition* or the proposition that we *have* the intuition that constitutes evidence for claims.

that utilitarianism implies that doing so is right. In this case, the intuition gives you reason (although only weak such) to believe that utilitarianism is true. On the other hand, moral intuitions can *conflict* with a theory and so give you reason to believe that the theory is false. For example, suppose that utilitarianism implies that choosing a lamb steak at a restaurant is wrong, but that intuitively it is right for you to order the steak. In this case, the intuition gives you reason to believe that utilitarianism is false.

In this and the following two chapters, I explore the consequences for utilitarianism of the idea that moral intuitions are epistemically significant. I discuss three intuitive objections to utilitarianism, and I consider various ways in which utilitarians may defend their theory against these objections. First, however, I will clarify what intuitions are, and present some reasons to think that they are in fact epistemically significant.

4.2 Intuitions and their Epistemic Significance

Several philosophers have written about how to understand intuitions, and I will not explore this topic in any greater detail here.[91] However, it will be useful to have a basic understanding of what intuitions are for the discussion that follows.

To begin with, intuitions share a number of characteristics. First, we use a certain vocabulary to report having them. When you intuit that *p*, you can report having this intuition by saying that "intuitively *p*," "it seems like *p*," "it appears that *p*," or "it looks like *p*." Second, intuitions have *propositional content*. You intuit that something is the case, such that giving to a charity is morally right, or that letting your cat starve is wrong. Third, intuitions are *immediate*. When you intuit that killing animals for food is wrong, this intuition appears in your mind suddenly and without warning; it does not slowly take shape during deliberation. Fourth, intuitions are *non-voluntary*. You cannot decide which intuitions to have or when to have them. For example, I cannot decide to intuit that torturing innocents is morally obligatory – I intuit that torturing innocents is morally wrong whether or not I want to. Fifth, intuitions *vary in strength*, so while I intuit that it is morally wrong to let my cat starve, I more strongly intuit that it is morally wrong to kill my cat. Sixth, and finally, intuitions are *mental states* with, or at least accompanied by, a particular phenomenology.

Let me further comment on the vocabulary used to report having intuitions. While terms such as "seems" and "appears" can refer to moral intuitions, they can also be used for other purposes. It is important to not conflate these different uses of these terms. One such alternative purpose is what we may call "epistemic reporting." For example, suppose that I have acquired significant

[91] See, for example, Bealer (1998) and Pust (2017).

reason to believe in utilitarianism, even though the theory does not strike me as true. Whenever I think about utilitarianism it "leaves me cold," so I do not intuit that it is true. Nevertheless, when reflecting on my available evidence in this situation I may truthfully tell you that "in the light of my evidence it seems that utilitarianism is true." In this case, "seems" refers not to a mental state of "seeming," but to having on balance (a little) more reason to believe that utilitarianism is true rather than false. More obviously, terms like "seems" and "appears" can be used to hedge statements, without thereby referring to any underlying mental states.

As should be clear, intuitions share many characteristics with beliefs, so it is natural to ask whether they *are* beliefs. While nothing important in this book hangs on intuitions being distinct from beliefs, as opposed to constituting a special kind of beliefs, I think it is reasonable to think that intuitions are mental states distinct from other propositional attitudes, such as beliefs and desires.[92] For example, I can intuit that *p* is wrong, but still distrust this intuition (e.g., perhaps I think that the intuition is unduly influenced by my cultural background) and so decline to form a judgment about *p*. This is not possible if to intuit that *p is* to believe that *p*. Moreover, intuitions have a different phenomenology than do beliefs – they have a psychologically insisting quality to them that "pushes" us towards believing in their contents. This is difficult to make sense of if intuitions *are* beliefs.

Why believe that having a moral intuition that *p* gives me defeasible reason to believe that *p* – that moral intuitions are, in the terminology that I employ, epistemically significant? To begin with, moral intuitions are seemings, and seemings include mental states such as:

(1) It seems to me that a cat stands outside my window.
(2) It seems to me that the dish contains chili.
(3) It seems to me that two plus two equals four.

All of (1)-(3) constitute the same type of mental state – that is, one of a proposition *appearing* or *seeming* to be true. While (1) and (2) are not what we typically would call intuitions, (3) is a clear example of an intuition – more precisely, a *mathematical* intuition. In general, intuitions appear to be a kind of *intellectual* seemings which contrast with *perceptual* seemings like (1) and (2). On this account, intuitions constitute a proper subset of seemings: every intuition is a seeming, but some seemings are not intuitions.

A first argument for why moral intuitions are epistemically significant leans directly on a comparison between them and other seemings. The argument goes as follows. Seemings such as (1)-(3) are clearly epistemically significant. If it seems to me that the dish contains chili, in the context of tasting the dish, then that seeming gives me defeasible reason to believe that it *does*

[92] Bealer (1998), pp. 208-210, Huemer (2008), p. 99, and Pust (2017).

contain chili. Next, presumably the best explanation as to *why* seemings like (1)-(3) are epistemically significant is that they *are seemings*: that they are mental states with the same psychologically insisting quality to them, one which pushes us towards believing in their contents. For example, that a cat seems to stand outside my window justifies the corresponding belief in virtue of how *it just seems* to me that there stands a cat in that location. Similarly, that two plus two seems to equal four justifies the corresponding belief in virtue of how *it just seems* to me that two plus two equals four. While it is possible that only *some* such seemings are epistemically significant and that others are not, it is difficult to explain what criteria would rule in only some seemings as epistemically significant. For example, it cannot be that all epistemically significant seemings have a perceptual basis, as (3) lacks such a basis, yet is epistemically significant. So, the argument goes, we should conclude that seemings in general are epistemically significant. Because moral intuitions are seemings, they are also epistemically significant.

A second argument for why moral intuitions are epistemically significant draws on a comparison to other philosophical intuitions. Why do I have reason to believe that, whenever it seems to me that a cat stands outside the window, this seeming gives me reason to believe that a cat stands outside my window? Presumably, because it *seems* that the intuition gives me such a reason. But this is an appeal to an *epistemic philosophical* intuition. Or consider what reason we have to believe that there cannot exist quadratic triangles. In this case, we refer to an *ontological philosophical* intuition that it *seems* like there cannot exist any such objects. Both epistemic and ontological philosophical intuitions – philosophical because they have philosophical propositions as their subject matters – are clearly epistemically significant. Therefore, we should expect moral intuitions to be epistemically significant as well, because they are also philosophical intuitions.[93]

Recall that the claim that moral intuitions are epistemically significant is a fairly weak one. As I defined the notion of epistemic significance earlier in this chapter, that a moral intuition is epistemically significant means only that it *defeasibly* gives the intuiter a reason to believe in its content, such as to believe that giving to a charity is right. This claim is compatible with the epistemic status of moral intuitions being *undermined* in various ways, and their not giving us any actual reason for belief.

In what follows, I assume not only that moral intuitions are morally significant, but that moral intuitions in general are not undermined. In other words, I assume that global skepticism about moral intuitions is false – where global skepticism could arise either because no moral intuition is epistemically significant, or because all moral intuitions are undermined. I will not attempt to argue for this view in this book, but will assume it for dialectical reasons. If

[93] This type of argument could be challenged if intuitions are produced by sufficiently heterogeneous processes. See Nado (2014).

moral intuitions never give us any reasons for belief in their content, then no intuitive objection to utilitarianism works. But utilitarians cannot help themselves to this defense, because by appealing to global skepticism, the utilitarian will at the same time lose most or all *positive* support for her theory. Simply put, the positive case for utilitarianism relies heavily on theoretical moral intuitions, including the intuitive plausibility of the utilitarian theory itself, and on evaluative intuitions such that pleasure is good and that pain is bad. The case for utilitarianism also relies on moral intuitions about particular cases, such as how to deal with emergencies, the correct prioritization of resources, and the permissibility of medical triage under some circumstances. For the above reasons, I will assume that the following are the rules of the game. In defending their theory, utilitarians can give us reason to distrust *specific* moral intuitions that conflict with utilitarianism. However, they may not let such arguments lead us to distrust *all* moral intuitions, or to in other ways undermine the positive case for utilitarianism. In other words, utilitarianism must defend their theory while surviving this defensive ordeal intact, with substantial remaining positive support. A utilitarian's defense against intuitive objections must therefore be surgical in nature: she must find a way to handle the troublesome intuitions while retaining enough intuitive support. The arguments that I discuss in the rest of this book are all attempts at such a defense.

4.3 Three Intuitive Objections

To narrow the focus of my discussion, I will only consider moral intuitions about particular acts. These *particular* moral intuitions must be distinguished from *generic* moral intuitions, which have as their contents generic statements such as "torture is wrong" and "it is right to take care of your family." Because "torture is wrong" is a generic statement, it is compatible with various particular acts of torture being right or obligatory, like how the generic statement "lasagna is tasty" is compatible with specific dishes of lasagna not being tasty. That lasagna is tasty means only that *typically* or *in general* lasagna is tasty, or that *paradigmatic examples* of lasagna are tasty. Similarly, that torture is wrong means only that typically or in general torture is wrong, or that paradigmatic examples of torture is wrong. As a result, even if we have the generic moral intuition that torture is wrong, and even if we learn that a specific act of torture is right according to utilitarianism, we have not thereby identified a conflict between the intuition and utilitarianism. Instead, to get a conflict between the intuition and the theory, we need to establish that if utilitarianism is true, then torture is not *typically* or *in general* wrong. This is difficult, in part because the truth conditions of "torture is wrong" are less transparent to us than the truth conditions of "the act of torturing person P at the time T is wrong". That being said, while I will focus on particular moral intuitions, I do not thereby *deny* that generic moral intuitions are relevant for evaluating utilitarianism. For example, suppose that the generic statement “it is wrong to not pay your taxes” is intuitively true. Moreover, suppose that utilitarianism implies that it is obligatory to *not* pay your taxes under a wide range of circumstances, including many paradigmatic and typical cases. For example, perhaps you should in such cases refrain from paying your taxes and instead give the money to animal charities. Under these circumstances, the truth of utilitarianism may imply that the generic statement “it is wrong to not pay your taxes” is false, and imply that the generic statement “it is obligatory to not pay your taxes” is true. In this case, the intuition that it is wrong to not pay your taxes will give one reason to believe that utilitarianism is false. The reason that I focus on particular moral intuitions is therefore not that they provide the *only* way of arguing for or against utilitarianism, but that they provide an easy and straightforward way to construct strong intuitive objections to the theory.

Setting aside generic moral intuitions, there are two ways in which utilitarianism can conflict with particular moral intuitions. First, utilitarianism can conflict with the intuited *deontic status* of acts – let us call these intuitions *deontic* particular moral intuitions. Second, utilitarianism can conflict with the intuited explanation of *why* acts have that deontic status – let us call these intuitions *explanatory* particular moral intuitions. For an example of the latter conflict, suppose that giving my child food produces the most pleasure minus pain, and suppose that my moral intuitions agree with utilitarianism that giving my child food is obligatory. Even so, there can be a conflict between my intuitions and utilitarianism about *why* feeding my child is obligatory. Intuitively, it may seem that “giving my child food produces the most pleasure minus pain of any alternative acts available to me” does not explain *why* giving my child food is obligatory (this is an example of an explanatory particular moral intuition). However, in that case utilitarianism must be false, since it gives precisely this (incorrect) explanation. In other words, utilitarianism can be false not only by having *incorrect implications* for which acts are right, wrong, or obligatory; but also by giving *incorrect explanations* as to why acts are right, wrong, or obligatory. Now, one reason to focus here on the deontic particular moral intuitions, rather than on the explanatory particular moral intuitions, is that the former are typically stronger. Moreover, when we consider how utilitarianism conflicts with deontic particular moral intuitions, we get the conflicts with the explanatory particular moral intuitions for free. If utilitarianism does not have the correct implication in a case, then it cannot give the correct explanation either. That is, if utilitarianism does not even *imply* that giving my child food is obligatory, then neither can it *explain* it being obligatory.

Figure 2 summarizes the distinctions between different kinds of intuitions which have been introduced in this section:

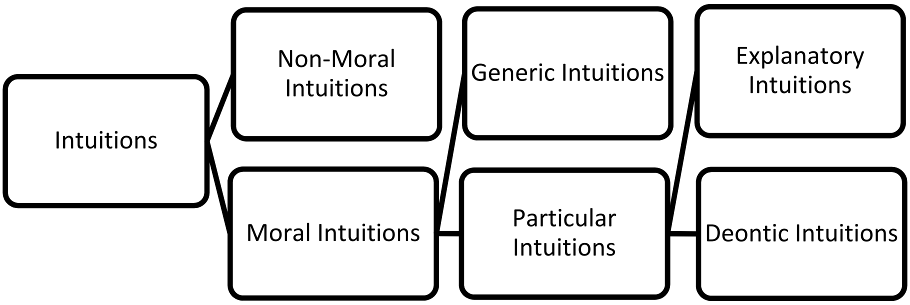

Figure 2. Kinds of intuitions

Now, when focusing on the deontic particular moral intuitions (which I will henceforth refer to simply as moral intuitions), note that intuitive objections to utilitarianism can be represented by the following general schema:

> (P1) If utilitarianism is true, then the act *A* has the deontic status *D*.
> (P2) *A* does not have *D*.
> (C) Therefore, utilitarianism is false.

An important question is how to support the premises of such an argument. While we can support premise (P2) by appealing to our moral intuitions, it is less obvious how to support premise (P1).[94] The problem is that if *A* is an act in a real world scenario, and *D* is either the deontic status of rightness or obligatoriness, then we have little reason to believe that (P1) is true in such an argument. As I note in Chapter 2, the future consequences of our acts are almost completely hidden from us. So we have barely any clue as to which acts are right or obligatory according to utilitarianism in real world scenarios. And we cannot support (P1) by an appeal to intuition – to how it *seems* that, if utilitarianism is true, then *A* has *D*. For such intuitions to be reliable, they would have to be sensitive to complex empirical facts about the far future, which they are not. Finally, note that what I just said is true *even if utilitarianism is false* because, as I pointed out in the previous chapter, *everyone* must accept that the long-term consequences of our acts matter for rightness and obligatoriness to at least some degree.

When intuitive objections concern the wrongness of acts rather than their rightness or obligatoriness, it is far easier to support premise (P1). Almost every act is wrong according to utilitarianism, since there are so many suboptimal acts available to us at any given moment. Therefore, we have significant reason to believe of any act that it is wrong according to utilitarianism. As I pointed out earlier, this is similar to a lottery where only one of a thousand

[94] Cf. Tersman (1991); (1993), pp. 122-123.

tickets will win; in such a case, I have significant reason to believe of *any* ticket that it will *not* win. This fact has peculiar consequences for the evaluation of utilitarianism. While utilitarianism will for this reason have *incorrect* implications in all cases where intuitively acts are right or obligatory, it will also have the *correct* implications in all cases where intuitively acts are wrong (since again, for any act about a real world case that we consider, we have significant reason to believe that it is wrong according to utilitarianism). Here it is difficult to see whether it is the objector to or the defender of utilitarianism who has the upper hand. In what follows, I therefore focus exclusively on cases where we can avoid these blanket judgments about rightness and wrongness. As a result, I discuss only *imagined* cases, considered by means of thought experimentation. In imagined cases, we can stipulate that acts have certain features, such as to maximize pleasure minus pain, and so avoid the above complications. When thought experimenting we can therefore, at least on the face of it, justifiably believe premise (P1) in intuitive objections to utilitarianism, which leaves us free to evaluate the intuitive support for premise (P2).

With these lengthy introductory remarks out of the way, let me now present three imagined cases together with the corresponding intuitive objections to utilitarianism. I will return to these cases and objections in the discussion that follows.

> EXPERIENCE MACHINE
>
> William has the chance to plug into an experience machine. If he plugs in, he will be extremely well off in terms of pleasure minus pain. He will have these experiences for the rest of his life. Therefore, plugging in produces the most pleasure minus pain of any act available to him.[95]

The experience machine objection goes as follows: If utilitarianism is true, then it is morally obligatory for William to plug into the machine. But intuitively, it is not morally obligatory for William to plug in. At the very least, William is morally permitted to *not* plug in, even if he is also morally permitted to plug in. Therefore, utilitarianism is false.

> TRANSPLANT
>
> Six patients lie sedated in their beds. Sarah can maximize pleasure minus pain by killing patient Six and use her organs to save the lives of

[95] Nozick originally presented such a case when objecting to hedonism as a theory of value. See Nozick ([1974] 2012), pp. 42-45. Here, I employ a similar case, but present it as an objection to utilitarianism.

> the other five patients. If she does not kill patient Six, the other five patients will die, and Sarah will produce a less than optimal amount of pleasure minus pain.[96]

The transplant objection goes as follows: If utilitarianism is true, then it is obligatory for Sarah to kill patient Six and transplant the organs. But intuitively, it is not morally obligatory for Sarah to kill patient Six. At the very least, she is morally permitted to *not* kill patient Six. Therefore, utilitarianism is false.

> UTILITY MONSTER
>
> Tim has resources available to him that can either help a thousand individual humans feel some amount of pleasure or help a non-human creature feel much more pleasure. Tim will maximize pleasure minus pain if and only if he gives the creature all of his resources.[97]

The utility monster objection goes as follows: If utilitarianism is true, then it is obligatory for Tim to give the non-human creature all of his resources. But intuitively, it is not morally obligatory for Tim to do so. At the very least, it is morally permissible for him to distribute the resources among the thousand individual humans instead. Therefore, utilitarianism is false.

Taken together, these three objections represent a broad group of intuitive objections that are employed against utilitarianism. Each objection works by exploiting a particular objectionable quality of utilitarianism. The experience machine objection exploits utilitarianism's exclusive focus on pleasures and pains – that it is only how we feel and not how we live that matters according to the theory. Second, the transplant objection exploits that there is no principled ban on violence and killing according to utilitarianism – that it finds no inherent problem with killing a person to save the lives of others; that the ends always justify the means. Third, the utility monster objection exploits how according to utilitarianism the *distribution* of pleasure and pain among individuals does not matter – that it is only the *total amount* of pleasure minus pain that matters.

How can utilitarians respond to these objections – and, by extension, to other intuitive objections as well? First, the utilitarian can argue that even if the moral intuitions *do* give us reason to believe in their content, they do *not* give us reason to believe that utilitarianism is false. I examine such strategies in Section 4.4 and Chapter 5. Second, the utilitarian can argue that the relevant moral intuitions do not even give us reason to believe in their content. I will

[96] The case was first discussed by Thomson (1976), p. 206. Transplant shares all or most important features with the so-called "footbridge" trolley cases.
[97] The case is due to Nozick ([1974] 2012), p. 41.

examine such strategies in Sections 4.5 and 4.6. Third, the utilitarian can argue that while the moral intuitions give us some reason to reject utilitarianism, utilitarianism can be successfully defended by an appeal to other more trustworthy intuitions. Such a defense is the topic for Chapter 6.

4.4 Scope Restriction

Moral theories can be thought of as moral laws, holding over a wide range of circumstances. For example, a moral theory has a wide *geographical* reach: if a moral theory is true, then it is not only true in Sweden, but also in Germany, the United Kingdom, and every other country on Earth. Conversely, if the moral theory is false in any country, it is also false in Sweden. Similarly, a moral theory is commonly thought to have a wide *modal* reach: if a moral theory is true, then it is true not only in the actual world, but also in many merely possible worlds, or alternative ways in which reality could be like. Conversely, if a moral theory is shown to be false in such a possible world, it must also be false in the actual world. This explains why there is no *principled* problem with evaluating moral theories merely by considering our moral intuitions about imaginary circumstances.

The above points suggest a way of defending utilitarianism against intuitive objections, which we may name *scope restriction*. Say that the "scope" of a moral theory is the set of possible worlds in which a moral theory is true if it is true in the actual world. For the intuitive objections that I mentioned earlier to be successful, the scope of moral theories must extend to the imagined cases of Experience Machine, Transplant, and Utility Monster. Therefore, to defend utilitarianism against these objections, we need only to show that the scope of moral theories fails to extend to these scenarios, in which case our moral intuitions about these cases become irrelevant to the evaluation of utilitarianism. That is, even *if* utilitarianism *does* in such a case conflict with our moral intuitions about these cases, we *still* could not draw upon these conflicts to argue that utilitarianism is false. It would be like arguing against the laws of nature in the actual world – like the second law of thermodynamics – by demonstrating that they fail to hold in the world of *The Lord of the Rings*.[98] The proper response to such an argument is straightforward. The laws of nature are restricted in scope, and their scope does not extend to imaginary worlds like *The Lord of the Rings*. Similarly, the utilitarian could argue that the moral laws are restricted in scope, and that their scope does not extend to cases like Experience Machine, Transplant, and Utility Monster. Therefore, whether or not our

[98] Presumably, most uses of magic, even the more subtle kind used by Gandalf and the other Maia, enables the magician to decrease entropy in isolated systems, rendering the second law of thermodynamics false.

intuitions about such imaginary circumstances are to be counted on, they provide no evidence against utilitarianism.

The problem is that, to restrict the scope of moral theories, we must deny a standard philosophical view. This is the view that the scope of moral theories is *all and only the metaphysically possible worlds*. Because the worlds of Experience Machine, Transplant, and Utility Monster are metaphysically possible, they will fall within the scope of moral theories according to the standard view. There are various rationales for why the scope of moral theories would extend this far in modal space. For example, perhaps the scope of *philosophical laws* in general, such as theories about personal identity, knowledge, and reference, extend to these worlds. Since moral theories are philosophical laws in this sense, they share this scope. But plausible objections have also been offered against the standard view.[99] For the utilitarian's argument to get going, I will simply assume that we can reject the standard view, and focus on how the utilitarian could take her argument from there. As we shall see, even when granting this controversial assumption, it is difficult to make the utilitarian defense against intuitive objections work.

Even when we refute the standard view of scope restriction, much work remains before we have an adequate defense of utilitarianism in hand. To begin with, we need to argue for a plausible *alternative* restriction of scope. Moreover, we need to show that this alternative restriction nets positive results for the plausibility of utilitarianism. For example, it is not helpful to the utilitarian theory to get rid of cases like Experience Machine, Transplant, and Utility Monster, if we also get rid of imagined cases that are used to support the theory.

One salient possibility is to limit the scope of moral theories to the *nomologically* possible worlds – i.e., the worlds that share the laws of nature with the actual world. However, there are numerous problems with such a proposal. For one thing, it would be surprising if the moral laws *exactly* match the laws of nature, because what could account for this overlap? Moreover, we can easily conceive of nomologically impossible worlds that should intuitively be included in the scope of moral theories. For example, consider a world which is exactly like ours, but where a small miracle occurs whenever a bird flaps its wings, giving it slightly more lift and so violating the laws of nature. Surely this world should not be excluded from the scope of moral theories for such a trivial reason. If utilitarianism is true in the actual world, then it is true in this "small miracle"-world as well. So the correct scope restriction seems unlikely to be purely nomological in this way.

Furthermore, even if a nomological scope restriction is plausible, it is not clear that utilitarianism will benefit from such a restriction. First, the cases of

[99] For arguments against strong supervenience for moral claims, see Tännsjö (2010), pp. 47-50, Hattiangadi (2018), and Rosen (Unpublished manuscript).

Experience Machine, Transplant, and Utility Monster are nomologically possible, at least with minor unimportant modifications to these cases. Second, there are counterpart cases to Experience Machine, Transplant and Utility Monster that are clearly nomologically possible, yet which elicit nearly as strong moral intuitions. For example, we can substitute the experience machine for drugs, entertainment, virtual reality, or being lied to by loved ones, all of which make similar "inauthentic" pleasures and pains possible. The transplantation can be changed for a "footbridge" where to maximize pleasure minus pain we must push a man from a bridge in front of a train to save the lives of five others (i.e., one of the classic trolley cases). Finally, Utility Monster can be changed for a version of the "repugnant conclusion" thought experiment, where to maximize pleasure minus pain we must cause a billion people with slightly above zero in well-being to exist, rather than causing a thousand people with a moderately comfortable lifestyle to exist. In these cases we have nearly as strong anti-utilitarian intuitions, and these cases are also clearly nomologically possible.

Instead of tying the scope of moral theories to *metaphysical* or *nomological* necessity, we could tie it to a *sui generis* type of necessity, such as *normative* or *moral* necessity.[100] But this move leads to another issue. The description of Experience Machine is compatible with – that is, the case can be "situated in" – any of several possible worlds. Therefore, it is not correct to think of a thought experiment as *being* a possible world; rather, a thought experiment is better thought of as being *compatible* with various possible worlds. While some possible worlds in which we can situate Experience Machine may be morally impossible, what we need to defend utilitarianism is a significantly more ambitious claim: that *all* of them are morally impossible. Why is this so? When we perform thought experiments, we look to the closest possible worlds in which a described case is true. This is the possible world that differs *the least* from the actual world. And even if there is only *one* possible world for which the description of Experience Machine is true and that is also morally possible, then we are likely to imagine *precisely* this world when conducting the thought experiment, simply because it shares its moral laws with the actual world (and so does not differ from the actual world in this respect). But it seems unlikely that there is not a single morally possible world in which Experience Machine can be situated. So restricting the scope of moral theories to the worlds that are normatively or morally possible seems unlikely to help the utilitarian.

A different kind of scope restriction restricts the scope of moral theories, not to a particular set of *possible worlds*, but to a particular set of *choice situations*. For example, we might think that the scope of moral theories extends only to *familiar* or *commonly encountered* choice situations, as opposed to

[100] For example, Rosen (Unpublished manuscript) has argued that moral principles hold with normative necessity rather than metaphysical necessity.

unfamiliar or unusual ones. Experience Machine, Transplant, and Utility Monster – as well as their realistic counterparts mentioned earlier – all have one feature in common: they are *unusual cases*. So this kind of scope restriction, it seems, would nicely exclude these cases from consideration. Of course, it is a fairly odd view about scope restriction to hold in the first place, as it is not clear what arguments could be given in its favor.

In any case, even if this type of non-modal scope restriction would be reasonable, this "familiar circumstances" defense of utilitarianism faces the following problem. In *what way* are the circumstances thought to be familiar? One possibility is that we require choice situations to be societally, culturally, biologically, or technologically familiar. But why would *these* features of choice situations decide the scope of *moral* theories? It seems that the familiarity involved must be of a distinctively *moral* kind. But Experience Machine, Transplant, and Utility Monster involve quite familiar moral issues: how to value pleasure and pain had from living a real or authentic life, how to determine whether violence or killing are permissible means to help others, and how to prioritize the well-being of different people. These cases are merely clad in a fantastic or unusual attire to strengthen our moral intuitions. So it seems that this proposal results either in an implausible view of scope restriction, or that it fails to properly exclude cases like Experience Machine, Transplant, and Utility Monster.

Moreover, in this section I have focused on difficulties for excluding cases from the scope of moral theories. But even if we get this far, and do successfully exclude cases like Experience Machine, Transplant, and Utility Monster, we must also take care not to jettison cases that elicit utilitarian friendly intuitions. For example, consider a case of medical triage, where a doctor must choose between using her resources to save one person or save five people (the death of the one is not a means for saving the five). In this case, our moral intuitions strongly match the utilitarian verdict: the doctor should save the five. But if Transplant is excluded from consideration because it is unusual, then surely the case of medical triage is excluded as well, since it too is unusual.

I admit that the above arguments are not conclusive, but at least they suggest that scope restriction is a difficult path to take for the utilitarian. There is no clear and easy approach that will let us defend utilitarianism by excluding the imagined cases that I have considered. The strategies that I discuss in what remains of the book are less ambitious. If the strategy of scope restriction is a *metaphysical* strategy, which defends utilitarianism by arguing that moral reality is in a certain way, then the rest of the strategies that I will consider are *epistemic* strategies, in that they let us defend utilitarianism by arguing that our epistemic position or abilities are in a certain way.

4.5 Debunking Arguments

I will now consider another type of strategy for defending utilitarianism: namely that of so-called *debunking arguments*. But before I begin, let me say something about the terminology that I employ. I have stipulated that a moral intuition is epistemically significant if and only if it gives the intuiter defeasible reason to believe in its content. I have also assumed that moral intuitions are epistemically significant in this way. Now, as I pointed out earlier in this chapter, even when an intuition is epistemically significant it may still lack justificatory force. That is, that a moral intuition is epistemically significant means only that it gives the intuiter a *defeasible* reason to believe in its content. To refer to moral intuitions whose justification is not in such ways "defeated," let us say that a moral intuition is epistemically *trustworthy* if and only if it *actually* gives the intuiter reason to believe in its content. While I have assumed that all moral intuitions are epistemically significant, I have left it open which moral intuitions are epistemically trustworthy. Although utilitarians cannot argue that *no* moral intuition is epistemically trustworthy – per the rules of the game mentioned earlier – utilitarians could still argue that *some* intuitions are not trustworthy and therefore pose no problem for utilitarianism. The strategy considered in this section constitutes one such argument.

Joshua Greene and Peter Singer have argued that we have evolved emotional responses to violence and that this renders some of our moral intuitions untrustworthy.[101] In their arguments, Greene and Singer try to *debunk* anti-utilitarian intuitions, by pointing to their suspect evolutionary or psychological origins. Singer writes that:

> For most of our evolutionary history, human beings have lived in small groups, and the same is almost certainly true of our pre-human primate and social mammal ancestors. In these groups, violence could only be inflicted in an up-close and personal way – by hitting, pushing, strangling, or using a stick or stone as a club. To deal with such situations, we have developed immediate, emotionally based responses to questions involving close, personal interactions with others.[102]

Greene gives a more detailed account of these emotionally based responses:

> First, this automatic setting responds more to harm caused as a means to an end (or as an end), rather than as a side effect […] Second, it responds more to harm caused *actively*, rather than passively. […] [T]hird, it responds more to harm caused directly by *personal force*, rather than more indirectly. [...] Putting these three features together, it seems that our alarm gizmo responds to

[101] See Singer (2005) and Greene (2013), and especially Greene's remarks on pages 264, 274 and 328. For criticism of Singer and Greene, see Tersman (2008), Sandberg and Juth (2010), and Meyers (2015).

[102] Singer (2005), pp. 347-348.

> actions that are *prototypically violent* – things like hitting, slapping, punching, beating with a club, and, of course, pushing.[103]

Here is how the above argument could apply to one of the cases that I considered earlier. In Transplant, Sarah has the option to kill a patient so that she can use the patient's organs to save the lives of five other patients. When contemplating whether Sarah should perform this act, its description triggers what Greene calls our "alarm gizmo" because the action is "prototypically violent." As a consequence, we intuit that it would be wrong for Sarah to kill the patient – let's call this the "do not kill" intuition. The argument is that because of its evolutionary and psychological origin, the "do not kill" intuition is untrustworthy.

For the sake of the argument, let us assume that Singer and Greene get the evolutionary background and the psychological details of the "alarm gizmo" right. Even so, giving an account of *why* we have the "do not kill" intuition does not by itself render this intuition untrustworthy. For example, an evolutionary or psychological explanation for why I see colors does not by itself make my color perceptions untrustworthy; similarly, an evolutionary or psychological explanation for why I have moral intuitions does not by itself make these intuitions untrustworthy.

Here it is important to see that the utilitarian debunker is a *selective* debunker.[104] The selective debunker tries to get rid of only *some* moral intuitions, while the *global* debunker tries to get rid of *all* moral intuitions. That the utilitarian is a selective debunker makes it difficult for her to employ a strategy sometimes favored by global debunkers in meta-ethics. A global debunker can argue that it is a *remarkable coincidence* if the evolutionary advantageous moral intuitions track truth. This fact supposedly supports that the intuitions do not in fact track truth, because it would be too *unlikely* for them to do so. Whatever its ultimate merits, such a global debunking argument is not readily available to the utilitarian debunker. Recall that the utilitarian debunker has already accepted that many moral intuitions are trustworthy, in particular those supporting utilitarianism, and many of these trustworthy intuitions clearly have evolutionary explanations as well. Therefore, the utilitarian must accept that there *are* many epistemically trustworthy intuitions that have evolutionary explanations. But in that case, it is neither unlikely nor surprising that a moral intuition such as the "do not kill" intuition is *both* explained by evolutionary facts *and* is epistemically trustworthy.

Another approach is to appeal to the moral *relevance* or *salience* of acts. For example, Singer asks:

[103] Greene (2013), pp. 246-247.

[104] For discussion of debunking arguments targeting moral realism, see Street (2006), Enoch (2010), and Kahane (2011).

> [W]hat is the moral salience of the fact that I have killed someone in a way that was possible a million years ago, rather than in a way that became possible only two hundred years ago? I would answer: none.[105]

Singer's idea is that there is nothing morally *special* about prototypical violence. Since moral intuitions produced by means of what Greene calls our "alarm gizmo" identify as wrong precisely acts of prototypical violence, the argument goes, we should not trust intuitions produced by this mechanism. It does not track something morally relevant. But to support this claim about moral relevance or salience, Singer must also make an appeal to intuitions: that there *seems* to be nothing special about prototypical violence. And this intuition is probably not widely shared, as many non-utilitarians surely feel that there *is* a clear difference between prototypical and non-prototypical violence. For example, there is intuitively a moral difference between hitting a man with a club in the head than it is to either shout at him – even if the outcome in both cases is a terrible headache. Even utilitarians like myself will agree that there *seems* to be a difference between these cases – it is just that we for theoretical reasons do not *believe* that there is such a difference.

For similar reasons, it does not help to show that the utilitarian friendly intuitions are the "products of reason" and that the utilitarian unfriendly ones are the "products of emotion," for it is far from clear whether we should trust reason-based intuitions any more or less than emotion-based ones. Moreover, many intuitions such that "pleasure is good" and that "pain is bad" do not seem to be products of reason, but have a similar emotional character as intuitions about prototypical violence. While I do not know whether these intuitions have similar psychological origins to those about prototypical violence, at least they arise intensely and suddenly in the same way.

In conclusion, what the utilitarian needs is a specific account of *how* moral intuitions are formed and why some of them are trustworthy and some are not. That is, we need to say something about the underlying mechanisms behind intuition formation. Having done that, the utilitarian can argue that our moral intuitions about cases such as Experience Machine, Transplant and Utility Monster fail to become epistemically trustworthy through this process, although moral intuitions which are more friendly to utilitarianism remain trustworthy. Let us now turn to one such attempt at defending utilitarianism, which appeals to our non-conscious application of moral theories.

[105] Singer (2005), pp. 348.

4.6 Non-conscious Application

The argument that I discuss later in this section focuses on the *unusual* character of cases like Experience Machine, Transplant, and Utility Monster, and how this might affect the epistemic trustworthiness of our moral intuitions. To many, the use of such cases in ethics is annoying and unserious. But it is difficult to explain *why* their use would be problematic. As I have noted earlier in Section 4.4, the standard view is that moral theories hold by metaphysical necessity, which means that their scope extends even to unusual cases like Experience Machine, Transplant, and Utility Monster. But if that is the case then, everything else being equal, we should be able to test moral theories like utilitarianism against our moral intuitions about these cases.

Consider then a natural idea: that our intuition forming abilities simply *function badly* when we consider unusual cases, and that the resulting moral intuitions are therefore epistemically untrustworthy. For example, Richard Hare says that:

> [People's] intuitions are the product of their moral upbringings […] and, however good these may have been, they were designed to prepare them to deal with moral situations which are likely to be encountered […] there is no guarantee at all that they will be appropriate to unusual cases. Even in the unusual cases, no doubt, the usual moral feelings will be in evidence; but they provide no argument.[106]

And Robert Goodin writes that:

> What those counterexamples do – all that they do – is to conjure up a situation in which doing the utility-maximizing thing would lead to intuitively unappealing results. The circumstances they depict, however, are very far from those to which our standard intuitions are standardly shaped. (They involve things like promises to dying friends on otherwise unpopulated desert islands and "super efficient pleasure machines" and such like.) Precisely because of that, we may well decide that it is our intuitions rather than the prescriptions of our utilitarian moral theory that ought to be readjusted in such unusual circumstances.[107]

Hare refers to our moral upbringings and Goodin refers to how our intuitions are "standardly shaped." In both cases, the idea is that our moral intuitions are epistemically trustworthy only if they are formed on the basis of common, as opposed to unusual, cases. What makes the cases that I have considered in this chapter unusual? Presumably, they are unusual because they include experience machines, perfect transplant procedures, methods for keeping killings in the healthcare system a secret, and alien creatures with a strange psychology. These are not things that we encounter in everyday situations.

[106] Hare (1981), pp. 131-132.
[107] Goodin (1995), p. 6.

Now, since we are assuming that moral intuitions are not in general untrustworthy, we need to explain what makes *these* intuitions – intuitions formed on the basis of unusual cases – untrustworthy, while many other moral intuitions remain trustworthy. Two types of arguments suggest themselves. First, we can argue that the problem is a lack of *training*. Just as a doctor requires extensive experience to have proper medical intuitions, a philosopher would require extensive experience with imagined cases to have trustworthy moral intuitions. But the problem is that philosophers *do* seem to have extensive experience with cases like those of Experience Machine, Transplant, and Utility Monster – they encounter these cases all the time: in the seminar room, in writing, and in discussion. And despite this massive experience with thinking and reasoning about these cases, for the most part philosophers have not changed their moral intuitions about the cases, and many still have moral intuitions that conflict with utilitarianism. So this "training" version of the defense is unlikely to help utilitarianism.

Second, and more promising, we could argue that the problem is not a lack of experience on behalf of the intuiter, but some feature built into our intuition forming processes. In what remains of this section, I will evaluate one such proposal, namely James Wood Bailey's *inferential view* of intuition formation, on which moral intuitions are formed by means of the non-conscious application of moral theories.[108] According to this view, when we form moral intuitions:

> [W]e are engaged in a sub- or semiconscious act of applying some set of principles – along with a large amount of empirical knowledge [...] The view that moral intuitions are a process of using tacit knowledge in conjunction with moral principles is not embarrassing at all. Not only does this make the phenomenon of moral intuition similar to that of intuition as studied by psychologists, but it also gives a perfectly mundane account of the phenomenon of conflicts between the intuitions of different persons. [...] It may in fact be the case that if utilitarians reason by using the same general tacit knowledge that the intuition holder uses sub- or semiconsciously, they will not conclude that we should do such-and-such. And thus like M. Jourdain, who had been speaking prose all his life without knowing it, our intuitions could well be fundamentally utilitarian without our really being aware of this fact.

For present purposes, some qualifications are needed to transform Bailey's idea into a working defense of utilitarianism. First, we need to assume that the theory which is being non-consciously (I employ this term rather than those of "sub-consciously" and "semi-consciously") applied is *true* or at least has correct implications in most cases; if not, we will land in a general skepticism about moral intuitions. That is, if the non-consciously applied theory is false, then no moral intuition will give us reason to believe in its content, because

[108] Bailey (1997), pp. 34-37. See also Eggleston's (2010) discussion about "practical equilibrium."

our moral intuitions will just reflect (i.e. be the result of the application of) that false theory. Second, it seems more reasonable to think that we non-consciously apply the empirical statements that we tacitly *believe* to be true, rather than those that we *know* are true. For example, if I am wrong about some aspect of the world, but still believe that the world is in such a way, then that will surely not stop me from applying the empirical belief non-consciously. Third, we should assume that the empirical knowledge that is drawn upon is about the *real* world and not about imagined cases. To coin a term, let us call this modified version of Bailey's view the *qualified inferential view*.

The qualified inferential view suggests the following way of defending utilitarianism.[109] Consider Experience Machine. If the qualified inferential view about intuition formation is correct, and if utilitarianism is true, then our moral intuitions about Experience Machine are generated by non-consciously applying utilitarianism, together with our tacit empirical beliefs about the real world, to the case. Plausibly, the tacit empirical beliefs that we will draw upon include (a) that experience machines, like other machines, are prone to failure and (b) that the pleasure we gain from entertainment in the machine is not outweighed by the good we can do in the world outside of it, for example, by working to alleviate poverty. On the basis of utilitarianism and these tacitly held empirical beliefs, we will form the moral intuition that it is permissible not to connect to the machine – and we will do this *even* if we have stipulated that connecting to the machine maximizes pleasure minus pain. This means that the moral intuition does not give us any reason to believe that utilitarianism is false, because we would intuit that it is permissible to not connect to the machine even if utilitarianism is true.

However, the above defense of utilitarianism suffers from several problems. First, we typically do not believe of *any* acts that they maximize pleasure minus pain in the real world. And it is such empirical beliefs, rather than beliefs like (a) and (b, which are needed for the above argument to work. Therefore, if the qualified inferential view is correct, then the truth of utilitarianism should lead us to generate *no* moral intuitions about the permissibility of not connecting to the machine, since there are *no* tacit beliefs with which the utilitarian theory can combine to generate an intuition.[110] Clearly, however, we *do* have moral intuitions about what is morally permissible in the case under consideration: we have the intuition that it is permissible to not connect to the experience machine. It follows that utilitarianism is false, since had it been true, we would not have had this intuition. Therefore, contrary to expectation, what we get is an argument *against* and not in *support of* utilitarianism.

[109] My discussion mirrors Bailey's to some extent, although his defense of utilitarianism is more detailed, drawing on (among other things) assumptions from game theory.

[110] Alternatively, if we tacitly believe for every act that it fails to maximize pleasure minus pain, then we should expect to generate the moral intuition that it is wrong to not connect to the machine.

Second, the qualified inferential view is as such problematic. To begin with, it does not correctly account for how some moral intuitions are generated. For example, we have no real-world beliefs about creatures such as the utility monster, so if the qualified inferential view about intuition formation is correct, then we should expect our intuition generating apparatus to treat the creature either as a normal human or to ignore it altogether. However, this is not what happens in Utility Monster. We seem open to give the creature *some* additional moral weight due to its weird psychology; it is just that we are not prepared to give it the *full* moral weight accorded to it by utilitarianism. This is especially true if we imagine an alien creature that suffers immensely, more so than any existing creature is capable of suffering. In such a case, our moral intuitions seem to endorse giving significantly more resources to alleviate this creature's suffering than we would to any ordinary suffering human. It is unclear, however, how the qualified inferential view can explain the formation of this intuition, since we do not have any real-world empirical beliefs about such an intensely suffering creature.

In addition, the qualified inferential view seems unable to explain some instances of *changing* moral intuitions, or of changes to the *strength* of moral intuitions. After thinking for a long time about a particular thought experiment, it is common to find that one's intuitions change or weaken. But in such circumstances, we have typically not changed any of the relevant empirical beliefs about the real world. The qualified inferential view cannot account for such a change in intuitions or their strength, because if our empirical beliefs about the real world remain the same, then the generated intuitions must also remain the same.

I am for the above reasons skeptical of the qualified inferential view, both as a way of defending utilitarianism, and as a view about how our moral intuitions are generated. That said, the distinction between unusual and common cases merits further attention. The defense of utilitarianism that I propose in Chapter 5 draws on this distinction, although in a novel way distinct from the strategies of scope restriction and non-conscious application.

In general, I think that Hare, Goodin, and Bailey have the right idea about how to best defend utilitarianism against intuitive objections. Rather than, as Singer and Greene do, debunk the origins of our moral intuitions, we should investigate the underlying conditions behind intuition formation. In the next two chapters, I take a closer look at the conditions under which our intuitions are formed. In particular, I consider the role of *imagination* in eliciting moral intuitions, and in making these intuitions epistemically trustworthy.

5. Misimagining

In this chapter, I argue that when we conduct thought experiments to test utilitarianism against our moral intuitions, we sometimes unknowingly carry out the wrong thought experiment. In many such cases, we elicit moral intuitions that we believe give us reason to reject utilitarianism but that in fact do not.

5.1 Imagining the Wrong Case

The strategy for defending utilitarianism that I advance in this chapter differs from those that I discussed in the previous chapter. Earlier I considered *restricting the scope* of utilitarianism to not extend over various thought experiments. I also considered the strategies of debunking moral intuitions and proposing a specific mechanism behind intuition formation, which I called the qualified inferential view. All these strategies share an assumption that is challenged in this chapter: that we manage to successfully *carry out* the thought experiments in the first place. I will argue that in some cases when we are testing utilitarianism using thought experiments, we unknowingly carry out the wrong thought experiment. To coin a term for this phenomenon, we *misimagine* the case – we try to imagine it, we fail to, and we imagine another case instead. Importantly, we sometimes misimagine a crucial aspect of a case: whether an act does or does not maximize pleasure minus pain.

Because it might appear implausible that we ever misimagine cases when we conduct thought experiments, the rest of this section is dedicated to arguing that it is not strange or uncommon to think that we sometimes misimagine cases. The rest of the chapter is organized as follows. In Section 5.2, I propose an explanation for *why* we misimagine some cases. I suggest that the cases we are especially prone to misimagine are those with stipulated features "unnatural" to them. In Section 5.3, I conclude that each of the three cases being considered – Experience Machine, Transplant, and Utility Monster – are precisely cases of this kind. Each of these cases contains a stipulated feature that is "unnatural to" or "fits badly with" the case as it is described – namely stipulations about the total amount of pleasure minus pain produced by acts. In Sections 5.4 and 5.5, I discuss two ways in which to prevent misimagining from occurring. Finally, in Section 5.6, I consider some potential objections to my view.

Similar ideas to mine can be found in the literature. For example, the argument that I discuss is related to the complaint that thought experiments are *underdescribed* – that they leave open different ways of filling out their details.[111] However, the phenomenon I point to is different: We might misimagine cases even if they are fully described, and we might successfully imagine cases even if they are wildly underdescribed. Moreover, it strikes me as odd to call cases like the ones I consider in this book *underdescribed*. For example, Experience Machine seems to me fully described; while its full description contains few details, this is because the *case* contains few details. As I see it, the focus on cases being underdescribed distracts from the real problem facing thought experimenters – that we may unknowingly imagine the wrong case.

Another idea related to mine was considered in the previous chapter, namely that unusual cases are suspect. As I have already noted, my argument bears some similarity to this complaint. But my approach is subtly different: it finds no fault with cases *as such* being unusual. As I demonstrate in Sections 5.4 and 5.5, cases may sometimes *avoid* the problem of being misimagined by being made *more* unusual. Sometimes, by making cases more unusual, important aspects of cases become more natural to them. Surprisingly, therefore, sometimes the problem is not that a case is *too unrealistic*, but that it is *too realistic*.

Finally, my discussion draws on Sharon Hewitt's discussion of hedonism and the experience machine, where she points out that our intuitions might fail to be sensitive to stipulation.[112] While Hewitt does not put her argument in terms of us *misimagining* cases, our approaches have much in common.

What would be well-established examples of misimagining cases? First, consider the phenomenon of *imaginative resistance*.[113] As is well known in the philosophical literature, some propositions are especially difficult to imagine being true – they resist being imagined. For example, try to imagine that a sunset is horrifying, that torturing innocent people is fun, or that an episode of pain in your life is intrinsically good. Alternatively, we may consider the following case proposed by Brian Weatherson:

> JACK AND JILL
>
> Jack and Jill were arguing again. This was not in itself unusual, but this time they were standing in the fast lane of I-95 having their argument. This was causing traffic to bank up a bit. It wasn't significantly worse than normally happened around Providence, not that you could have told that from the reactions of passing motorists. They were convinced that Jack and Jill, and not the

[111] For discussion of the problem of thought experiments being underdescribed, see Friedman (1987), pp. 200-201, Wilkes (1988), pp. 1-48, Sorensen (1992), pp. 246-24, Häggqvist (1996), pp. 136-159, and Wilson (2016), pp. 136-140. For a response to Wilkes, see Brooks (1994).

[112] Hewitt (2010), especially pp. 337-343.

[113] For a well-known treatment of imaginative resistance, see Gendler (2000). For a more recent overview, see Gendler and Liao (Forthcoming).

> volume of traffic, were the primary causes of the slowdown. They all forgot how bad traffic normally is along there. When Craig saw that the cause of the bankup had been Jack and Jill, he took his gun out of the glovebox and shot them. People then started driving over their bodies, and while the new speed hump caused some people to slow down a bit, mostly traffic returned to its normal speed. So Craig did the right thing, because Jack and Jill should have taken their argument somewhere else.[114]

In this case, we are stipulating that Craig did the right thing – that he did what was morally right to do in the situation. But we also find this difficult to imagine. In cases like these, either of two things might happen. First, we might fail to imagine *any* case whatsoever – our imagination simply shuts down. Second, we might imagine *a* case, but not the case that we were asked to imagine. For example, when trying to imagine a horrifying sunset, I might instead imagine an ordinary beautiful non-horrifying sunset. This is an example of misimagining: I am asked to imagine a case *C*, but when trying to imagine *C*, I imagine a different case *C**. Similarly, when I try to imagine the case of Jack and Jill, I will imagine a situation where it is *not* morally right for Craig to shoot them, rather than no case at all, or a case where it is morally right for Craig to shoot. This, again, is a case of misimagining.

Cases of imaginative resistance exist at the extreme end of a continuum, and other propositions are less difficult, although still hard, to imagine. For example, Simon Stevin famously asked his readers to imagine a chain draped over a frictionless plane. The chain is easy to imagine, but it is difficult to imagine the plane. Any flat space that we are acquainted with exhibits *some* friction, and so performing this thought experiment taxes our imagination. Here it is easy to unknowingly imagine an ordinary plane with some friction, especially if we do not attend to the details of the case.

Misimagining also occurs frequently when reading or listening to fiction, where we depend on elaborate descriptions to get the details right. If I read a text too fast, I often fail to imagine the case properly. I will imagine people looking different, wearing different clothes, possessing different motivations, and so on. Moreover, if I lack a complete understanding of the words used or concepts employed in a text, I might misimagine various parts of the case being described. Finally, in some cases, I might simply not remember the correct details, leading me to misimagine the case. Even if not all such instances of misimagination are conducted *unknowingly*, some clearly are.

[114] Weatherson (2004), p. 1.

5.2 Explaining Misimagining

What could explain that we tend to misimagine some cases, such as Jack and Jill, but not others? One idea is that when a feature of a case is *unnatural* to it, imagining *other* features of the case stops us from – or "cancels out" – imagining this unnatural feature.[115] That a feature is unnatural to a case means simply that it fits or coheres badly with the rest of the case as it is described- As it stands, this is a fairly shallow explanation – we will still want to know *why* our imagination strives after such coherence between the features of imagined cases. As such, this explanation should be compatible with a number of "deeper" explanations of the phenomenon of misimagination. Nevertheless, despite its shallowness, the explanation will be sufficient for my purposes in this chapter.

The unnaturalness explanation accounts for several cases of misimagination. For example, imagining the sunset stops us from imagining it as horrible, because its being horrible is unnatural to the case. Similarly, imagining Jack and Jill arguing in the street stops us from imagining Craig's act of killing them as morally permissible, because it being morally permissible is unnatural to the case. To say somewhat more about this proposal, let me now briefly introduce an account of what cases and thought experiments are, and of what we do when we perform thought experiments.

To begin with, let us say that a *case* is a certain way the world could be like. Regardless of whether a case is a set of propositions, it can at least be represented by such a set. What we do when we describe a case using a short vignette – as I previously did for Experience Machine, Transplant, and Utility Monster – is to list these propositions. Accordingly, a case is not a whole possible world – a full representation of an alternative reality – because the propositions that represent cases like Transplant are true in (i.e., "can occur in" or "are compatible with") any of a number of possible worlds.

To continue, let us say that a *thought experiment* is simply a case that has been formulated for a specific purpose: to test a hypothesis by means of thinking about the case. This means that while all thought experiments are cases, not all cases are thought experiments. For example, some cases are formulated

[115] How does my suggestion relate to the vast literature on imaginative resistance, which I have not engaged with to any substantial degree? I am not entirely sure, but I suspect that the phenomenon of misimagination is different from many of the concerns that are addressed in those discussions. For example, the problem that I am concerned with is not that some cases *resists* being imagined, that authorial authority *breaks down* in some cases, that we do not *want* to imagine some cases, or that the imagining of some cases has a *peculiar phenomenology*. Instead, the problem is that we sometimes unknowingly imagine a different case than the one that we intended to imagine, which has consequences for the practice of testing moral theories against our intuitions about cases. But perhaps I am wrong, in which case the above discussion would need to be better informed by the literature on imaginative resistance. See also Weatherson's taxonomy of different problems which are raised by cases like Jack and Jill, in Weatherson (2004), pp. 1-2.

only for *illustrative* purposes, and are therefore not thought *experiments*, and some cases are formulated only for testing theories in a lab or computer simulation, and are therefore not *thought* experiments. On the above account, Experience Machine, Transplant, and Utility Monster are thought experiments, since they are formulated for the purpose of testing utilitarianism, and for doing so by means of thinking about what the theory implies in these cases.

What does it mean to *perform* a thought experiment? I will assume that to perform a thought experiment is to create a "model" of it by means of our imagination.[116] Therefore, just as there can be computer models and wooden miniature models, there can be "thought models." Along these lines, I assume that when conducting thought experiments, we *imagine* that the propositions representing it are true, not in the actual world, but in the model – the thought experimental "fiction."

When modeling a case by means of our imagination, we can use different *types* of imagination. To begin with, there are *sensory* and *propositional* imagination.[117] Sensory imagination can be used to "paint a picture in your mind" or "seeing with your mind's eye" what a mountain looks like. It is related to its perception analogues – that is, visualizing resembles seeing. Sensory imagination can occur using various sensory moods, and not just the visual one. Therefore, we can sensorily imagine smells, sounds, pain, heat, and so on. For example, I may imagine how painful it would be to have a headache right now, or how it would feel to freeze in my office. In contrast to sensory imagination, propositional imagination does not employ any sensory moods, but is simply about representing a proposition as being true. It is what we use when we imagine ten billion people living lives barely worth living – we assume or suppose that they live such lives. Sensorily imagining a feature of a case seems to entail propositionally imagining it, because to sensorily imagine that a beautiful mountain stands in front of me, I must suppose or assume that there is such a mountain. But propositionally imagining a feature does not entail sensorily imagining it. For example, I can propositionally imagine ten billion people without sensorily imagining even a single person.

Another distinction, which is especially relevant to my arguments in this chapter, is that between *active* and *passive* imagination. For example, when reading a book, I might imagine that a character who has not been fully described has black hair, a slender physical build, and wears a t-shirt with jeans. However, I might never have intended to imagine any of these details – they sprung into my mind uninvited, without being prompted by the author. Similarly, when reading a story about a far-away country, I might imagine that this

[116] A mental modelling account of thought experiments is defended by Nancy Nersessian (1992), (2007); Nenad Miščević (1992), (2007), and Tamar Szabó Gendler (2004). Rachel Cooper (2005) defends an account of thought experiments as models, although not as exclusively mental models.

[117] McGinn (2004), pp. 128-139; (2009), p. 595.

country has a functioning healthcare system. Again, however, I may never have intended to imagine such a system – my brain filled in these details automatically. These are examples of passive imagination, which contrast with cases of active imagination, where we intentionally imagine various features of cases.[118] Some reflection shows that the distinction between active and passive imagination is orthogonal to that between sensory and propositional imagination. As a result, there can be active and passive sensory imagination, and active and passive propositional imagination. For example, when I imagine a character with black hair, I sensorily passively imagine her hair, and when I imagine the functioning healthcare system, I propositionally passively imagine this fact.

Consider now cases of misimagining. That we misimagine a case is not necessarily due to lacking the *ability* to imagine the features of the case. In the Jack and Jill case, we *can* imagine that killing Jack and Jill is morally permissible. For example, we can imagine that Jack and Jill's arguing in this particular spot will set off a massive bomb, killing a million people and devastating the city – in which case they should indeed have taken their argument elsewhere.[119] Under those circumstances, it is not implausible that Craig did the right thing. Moreover, nothing in the description of the case as given by Weatherson prevents us from imagining this much; nevertheless, our passive imagination is uncooperative, and does not help us fill out the case in the required way. What happens in these cases is that our passive imagination runs with what is, broadly understood, *natural* to assume in a case, and this might or might not be helpful for imagining the propositions that are stipulated to be true of the case.[120] In the Jack and Jill case, our passive imagination seems to draw on assumptions of what consequences arguments in the street *usually* lead to, and these consequences do not include millions dying from bombs.[121] Similarly, when reading works of fiction in which we have either forgotten about or have not been told about various characters' moods, motivations, clothing, etc., our passive imagination seems to draw upon what is natural to

[118] As an alternative to spelling out the active-passive distinction in terms of *intentionally* imagining features of a case, we could define it in terms of *attentively* or *consciously* imagining such features. I am not sure what is the best way to go here – there seems to be advantages and disadvantages associated with each of these proposals. In any case, I hope the phenomenon that I try to describe is clear enough.

[119] Cf. Weatherson (2004), p. 20.

[120] One's views here might or might not mirror one's views about truth in fiction. Note, however, that the question of *what is true* in a work of fiction is different from the question of *what we imagine* to be true when considering the work. See also Daniel Kahneman (2013), pp. 97-99, on the phenomenon of "substitution." According to Kahneman, sometimes our fast, intuitive system (System 1) answers, instead of a hard question that we are asked, a different question which is easier to answer. As Kahneman notes, we are often unaware that we have answered the wrong but easy question, and not the right but hard question.

[121] The author's intentions matter as well, although perhaps just indirectly. There is a reason why Weatherson chooses a case where it is natural to assume that Jack and Jill's arguing has no particularly adverse consequences.

assume is true in the case – such as assumptions from the real world or from the fiction-specific genre. For example, if we have not been told of any details about a particular dragon in a fantasy epic, we will passively imagine that it has scales and is difficult to slay, because that is the natural assumption to make in fantasy fiction.

One would think that our active imagination of stipulated features of cases always overrules our passive imagination. But in some cases the opposite seems to happen. For example, in the Jack and Jill case our passive imagination *stops* us from imagining that Craig is acting permissibly. It seems that our passive imagination quickly takes onboard various natural assumptions about cases, and once these details are in place, we are stopped by our passive imagination from actively imagining certain propositions being true – even when we are explicitly asked to imagine them.

From the above discussion, it follows that certain cases are especially susceptible to being misimagined. These are cases that contain propositions that are "unnatural" to them, or that fit badly with the rest of the case as described, and where it is easy for our imagination to draw upon alternative, more natural, assumptions in order to make the case more coherent. That Craig is acting morally permissible is such a proposition. As we will see next, other examples of "unnatural" propositions include those describing the amount of pleasure minus pain produced by acts in Transplant, Experience Machine, and Utility Monster.

5.3 Revisiting the Objections

Just as we can fail to imagine that Craig is morally permitted to kill Jack and Jill, we can fail to imagine that an act maximizes pleasure minus pain. To begin with, consider the following variant of Transplant:

> FAILED TRANSPLANT
>
> Six patients lie sedated in their beds. Sarah can kill and use the organs of one patient to attempt to save five others. If she does, all six patients will die. In the ensuing outrage, the public will stop trusting in the healthcare system, leading to many fatalities and missed treatments. Killing the patient maximizes pleasure minus pain, while not killing her does not.

At least initially, I tend to misimagine this case. I successfully imagine every proposition except for the last one – i.e., that killing the patient maximizes pleasure minus pain. Instead, I imagine a different case where killing the patient does not maximize pleasure minus pain. Only with some effort, I can force myself to imagine the case correctly; for example, I can imagine a bomb that will kill millions of people if Sarah does not kill the patient – that would explain why killing her maximizes pleasure minus pain. But this is not what I first imagine when I encounter Failed Transplant. Because it is liable to make us misimagine it, Failed Transplant is obviously not a useful case for objecting to utilitarianism. When I try to imagine Failed Transplant, I successfully imagine a case in which it is intuitively morally wrong to kill the patient. But the content of that moral intuition does not give me any reason to reject utilitarianism, because in the case that I do in fact imagine, as opposed to the case that I try to imagine, killing the patient does not maximize pleasure minus pain. Utilitarianism does not imply that Sarah should kill the patient in the case that I imagine. In essence, I have the *right moral intuition*, but it is *about the wrong case*.

Consider now the original variant of the case:

> TRANSPLANT
>
> Six patients lie sedated in their beds. Sarah can maximize pleasure minus pain by killing patient Six and use her organs to save the lives of the other five patients. If she does not kill patient Six, the other five patients will die, and Sarah will produce a less than optimal amount of pleasure minus pain.

For this case too, I propose, there is a risk that we misimagine it. In Failed Transplant the problem is that we actively imagine that propositions like the following are true: "In the ensuing outrage, the public will stop trusting in the healthcare system, leading to many fatalities and missed treatments." In contrast, in Transplant the problem is that we passively imagine such propositions. We passively imagine what is natural to assume to be true in a case, and it is natural to assume that someone will find out about the transplant, and that this will result in fatalities due to a lack of trust in the healthcare system. It is also natural to assume that transplants are not always successful due to, for example, tissue rejection, incompetence, infection, and so on. Therefore, it is likely that our passive imagination will interfer with our attempt to actively imagine that killing the patient does in fact maximize pleasure minus pain in Transplant.

Corresponding arguments can be made for the other two cases that I have discussed:

> EXPERIENCE MACHINE
>
> William has the chance to plug into an experience machine. If he plugs in, he will be extremely well off in terms of pleasure minus pain. He will have these experiences for the rest of his life. Therefore,

> plugging in produces the most pleasure minus pain of any act available to him.

> UTILITY MONSTER
>
> Tim has resources available to him that can either help a thousand individual humans feel some amount of pleasure or help a non-human creature feel much more pleasure. Tim will maximize pleasure minus pain if and only if he gives the creature all of his resources.

In Experience Machine, it is natural to imagine that machines often break down, that they do not always receive the proper level of maintenance, that they can be hacked and controlled by malicious actors, that others are not good at determining what will actually give us more pleasure and less pain, and that mere entertainment is nearly always less conducive to pleasure minus pain than is helping other humans and animals. In Utility Monster, it is natural to imagine that giving the resources to the monster will face steep declining marginal returns on the investment, and that the monster, like any other creature known to us, has a maximum level of pleasure that can be obtained by spending resources. In both cases, these natural assumptions interfere with us imagining that the relevant acts (i.e., connecting to the machine and giving the monster the resources) do in fact maximize pleasure minus pain.

Again, recall that the problem is not that we *cannot* imagine that these acts maximize pleasure minus pain. To do this, we need merely imagine a feature that makes it natural to assume that pleasure minus pain is maximized. For example, we can imagine that a bomb will kill millions of people if the agent does not kill the patient, plug in to the experience machine, or give the creature the resources. But such a "brute force" approach is not useful for constructing objections to utilitarianism, for obvious reasons: in such revised cases it is also intuitively obligatory to kill the patient, to plug into the experience machine, and to give the resources to the creature.

My proposal is that the mere risk that these cases are being misimagined should lead us to consider revised cases instead. These revised cases should be ones where it is not unnatural to assume that the acts under consideration do, in fact, maximize pleasure minus pain. In doing so, we will of course need to stay true to the spirit of each objection – we cannot simply add that a bomb will explode if Sarah operates on the patient. I will now consider two general ways in which to carry out such revisions. First, I will consider *adding details* to the cases that make it more natural to assume that the relevant acts maximize pleasure minus pain. Second, I will consider *removing details* from the cases – not in the sense of making the list of stipulated propositions shorter, but by including propositions that ensure there is less room for our passive imagination to fill out the cases in the wrong way.

5.4 Adding Details

Consider first how, by adding propositions to the description of a case, we can make it more natural to assume that an act does in fact maximize pleasure minus pain. For example, consider this revised version of Transplant:

> REVISED TRANSPLANT
>
> Civilization have broken down and Sarah – a highly qualified doctor – is alone responsible for an island of high technology located in a barren wasteland: an autonomous hospital in the middle of nowhere. What remains of humankind is disorganized and spread out, and no communications exist between the hospital and the outside world. As it is, Sarah has admitted six patients to the hospital. All of them are lone wanderers who recently arrived from the surrounding wastelands, and who now lie sedated in her wards. No other patients will ever arrive to her hospital again – detailed scans show that these are the last brave souls who travelled through the desert before it became impassable. Moreover, Sarah knows from extensive experience that she, together with her advanced robotic workforce, can perform any transplant whatsoever with no risks to the patients. She also knows that she can safely escort any survivors to safer lands. One of the patients is patient Six. Sarah can maximize pleasure minus pain by killing patient Six and use her organs to save the lives of the five other patients. If she does not kill patient Six, she will survive, but the other five patients will die. In such a case, Sarah will produce a less than optimal amount of pleasure minus pain.

Revised Transplant is a different case than Transplant, because it is represented by a different set of propositions. That being said, Transplant still "implies" Revised Transplant in a sense, because by imagining every proposition of Revised Transplant, you will also imagine every proposition of Transplant (note that the description of Transplant does not mention any context or specifics of the transplantation procedure).

The details that are added in Revised Transplant are strategically selected. The goal with each revision to the original case is to make it more natural to assume that killing patient Six maximizes pleasure minus pain. First, I have removed any effects the transplantation procedure can have on society in general. In Revised Transplant we imagine a hospital in the middle of nowhere, that humankind is disorganized and spread out, and that no communications exist between the hospital and the outside world. Second, I have removed the possibility that the transplantation could fail: we imagine a highly qualified doctor aided by an island of high technology and an advanced robotic workforce. Third, I have removed potential effects on future patient care, such that people will not dare to be treated at the hospital in fear of being sacrificed for the greater good, by stipulating that the six patients are the last ones to be treated at the facility. The case also introduces some unnatural features of its own – for example, that the survivors would be able to escape the desert afterwards. Therefore, I added that the survivors will be safely transported to safer lands.

While the revisions made to Transplant are substantial in many ways, I take it that they still respect the spirit of the original objection, as they change nothing that is relevant or important to the original case. The original objection, I assume, is proposed as a way to see how utilitarianism unacceptably permits killing innocent people by means of direct physical force, or prototypical violence. This feature remains fully represented in Revised Transplant.

Do our moral intuitions differ between Transplant and Revised Transplant? If they do not, then this proposed defense of utilitarianism will fail. Even if we often do misimagine Transplant, we can instead appeal to our intuitions about Revised Transplant – a case which, I have argued, we are less prone to misimagine. I can only report on my own intuitions here. When I am confronted with Transplant, I have the intuition (at least initially) that it would be wrong for Sarah to kill patient Six. In imagining this case, I imagine a doctor standing in a gloomy cellar with a scalpel in hand, the peaceful hospital above blissfully unaware of the crime being secretly committed. But when I consider Revised Transplant, I weakly intuit that Sarah should kill patient Six. To kill patient Six still strikes me as a horrible thing to do, but it also seems like a necessary evil and so the *right* thing to do. Now, your intuitions might differ from mine with respect to these cases – I admit that my intuitions are perhaps corrupted by my utilitarian beliefs. But even if your intuitions differ from mine, for utilitarianism to be made more plausible we do not require that the moral intuitions *change*. It is enough that their *strength* change. And presumably even stalwart anti-utilitarians will have weaker intuitions about it being wrong for Sarah to kill patient Six in Revised Transplant, in contrast to intuitions they have on the basis of considering Transplant.

Clearly, there are various ways in which to make revisions like those above, and I have only suggested one of several possible modifications of Transplant. For example, instead of stipulating various facts about the downfall of civilization or the technological feasibility of organ transplantation, we could instead stipulate facts about how people *react* to such operations. John Harris, for example, asks us to consider a world where organs are forcibly donated by means of a "survival lottery":

> No one was considered to have an absolute right to life or freedom from interference, but everything was always done to ensure that as many people as possible would enjoy long and happy lives. In such a world a man who attempted to escape when his number was up [and was therefore chosen to become a

> donor] or who resisted on the grounds that no one had a right to take his life, might well be regarded as a murderer.[122]

As the case is described by Harris, no one would be afraid to use the healthcare system merely because, within that system, people are regularly killed to save others. Indeed, such killings are made routine, and lead to less deaths overall. Perhaps this gives us another way to avoid passively imagining that forced donations would cause fear and fatalities due to people avoiding life-saving treatment. On the other hand, Harris' case also introduces an unnecessary dimension of justice or fairness: the person benefits from the existence of the organ lottery, but refuses to contribute when his number is up. In general, constructing cases like these will therefore require some creativity to get right, although that should not deter us, as this much is true for all cases which are used to test moral theories.

For a suggestion of how to construct a revised version of Experience Machine, consider the following case proposed by Sharon Hewitt:

> We know how to keep everyone happy and healthy, and the machine can carry this out as well as – if not better than – we can. There aren't any experiences for us to have that haven't already been had, and that haven't already been fully analyzed as to their phenomenal and physiological characteristics and catalogued in the experience machine. What things remain to be learned about the universe are going to be things we know, a priori, can have no effect on improving anyone's health or the subjective quality of anyone's life – things like the exact number of atoms there are, or historical facts that are so specific that they don't reveal to us any general truths about the world or human nature that we weren't already aware of.[123]

In the light of this description, consider the following attempt to revise Experience Machine. First, we stipulate that the machine has a 1000-year track record of perfect run-time. Second, we assume that the protagonist – William – has sampled what the machine has to offer on several earlier occasions. These two assumptions make it more natural to assume that the machine does in fact provide superior experiences. Third, we add to the case that there are no morally worthy goals to strive for in the "real" world. This stipulation helps us to refrain from passively imagining that we can produce more pleasure minus pain by not connecting and instead engage in worthy charitable causes.

I believe that another major revision to Experience Machine is called for. Instead of considering whether William should *plug himself in*, we should consider whether he *should plug someone else in*, such as an old friend. This change in perspective helps us focus on what William *morally* should do, instead of what he *prudentially* should do. More importantly, it helps us ignore

[122] Harris (1975), p. 83.
[123] Hewitt (2010), p. 342.

another problem for utilitarianism. The problem I have in mind is that we intuitively have a very significant moral leeway in deciding what to do with our *own* lives, as opposed to the lives of others. Thus, even if I prefer to not perform an act *A*, and even if *A* causes me no pleasure and much pain, and even if performing *A* restricts my freedom, equality, and so on, I *still* seem morally permitted to perform *A*, given that it achieves whatever moral goals are worthy to achieve outside of my own life. While this is a problem for utilitarianism – and a worrying one – it is also a distraction in the current context. Experience Machine is meant to illustrate that utilitarianism does not properly distinguish between having authentic experiences, or living a life connected to reality, compared to living a life connected to the machine. Moreover, many other moral theories will also struggle to account for extensive moral freedom to decide how to live our own lives. Therefore, we will do best to remove this feature from the case, leaving this particular problem for another day. Doing so will also conveniently allow us to stipulate that William can interfer and unplug his friend if the machine malfunctions. This makes the assumption that William's friend will have hedonically superior experiences in the machine more natural to the case. Finally, we should imagine that William's friend cannot be consulted about whether to be plugged in or not – again, to avoid the problem of how a person is intuitively morally permitted to decide for herself how to live her own life.

These modifications to Experience Machine results in the following case:

> REVISED EXPERIENCE MACHINE
>
> For 1000 years, a sophisticated machine civilization has run a number of highly reliable experience machines. William lives in this civilization, and he has sampled what experiences a specific machine offers at several occasions. William has the opportunity to plug in an old friend, John, to this machine. While William cannot ask John what he prefers, he knows from careful and extensive testimony of millions of others over the past millennia, as well as from his own experiences, that John will have hedonically superior experiences by being plugged in. Moreover, there is nothing morally worthwhile to do in the world around William and John: poverty, disease, and mental illness have all been eradicated. Finally, William will continuously monitor John's well-being in the machine and unplug him if the experiences are no longer hedonically superior. If John is plugged into the machine, William maximizes pleasure minus pain; however, if John stays unplugged, William does not maximize pleasure minus pain.

In this revised version of the case, I intuit that it is obligatory for John to plug Williams into the machine.

Finally, let us consider Utility Monster. It is difficult to remove the most problematic feature of this case, which are the intense pleasures that the nonhuman creature will receive from being given Tim's resources. The problem is that even if we stipulate that the creature is not human, we will still use our basic picture of a human being to imagine what the creature's life is like, and humans seem limited to a certain maximal intensity in pleasure. Moreover, humans also suffer from declining marginal returns on pleasure-raising investments.

What we can do is to consider a case where investing additional resources does not raise the *intensity* of the felt pleasure, but rather adds to its *duration*. In doing so, we can consider an explicitly human creature, rather than a "monster." With that in mind, consider the following revised variant of the case:

> REVISED UTILITY MONSTER
>
> Tim has resources available to him that can make 1000 individual humans each feel a ten second episode of strong pleasure. Alternatively, Tim can give a single human the same resources, which would make this human feel 11,000 seconds of equally strong pleasure. Tim will maximize pleasure minus pain if and only if he gives this human the resources.

When considering this case, my intuitions weakly support giving the single human the resources. Clearly, some will have more egalitarian intuitions and think that Tim should distribute his resources among the 1000 individual humans instead. But presumably this intuition is at least weaker than the intuition that Tim should give the 1000 humans the resources in Utility Monster. Of course, nothing said so far will convince those who think that, contrary to the utilitarian assumption, pleasures are less valuable the more intense they become. So the case is still not ideal. Perhaps we should conclude that utility monster cases are simply not well suited as objections to utilitarianism, because the intensity of the pleasure had by the utility monster is both essential to the case and a highly unnatural assumption to make.

In my discussion so far, I have focused on how to use the active imagination of stipulation propositions to add details to existing thought experiments. However, there exists another way to add details to cases. Science fiction and fantasy writers have long made use of our passive imagination for this purpose. They first teach us how a fictional world works, and then they rely on readers to passively imagine the right details without being explicitly instructed to do so. We are often not even told the details at any point in the story, but will simply infer them from other descriptions of the world. For example, I passively imagine that Gandalf cannot throw fireballs due to the fictional context established by Tolkien when he wrote *Lord of the Rings*. Tolkien never explicitly states that Gandalf cannot throw fireballs, but it nevertheless follows from how magic is presented as subtle and inconspicuous in Tolkien's works. By learning how a fictional world works in this way, we can thus "train" our passive imagination to find certain things natural in it.

Perhaps we can use the same approach when we test moral theories like utilitarianism. For example, we may read an elaborate sci-fi fiction novel where plugging into experience machines, performing transplants from one to five patients, and giving certain people high levels of resources really *do* reliably maximize pleasure minus pain. Of course, we need to become acquainted with this world to such a degree that we actually believe and accept this. But when we do become acquainted with it, we should be able to read only short passages like those in Experience Machine and Transplant, and carry out the thought experiments without our passive imagination getting in the way. This would remove the need for more elaborate descriptions of cases.

As it is, many existing fictional works are unsuitable for testing utilitarianism against our moral intuitions. In most fictional works, killing, torturing, or saving 1000 innocent strangers but letting your loved one die, always have worse consequences than some alternative act. In most fictional works, it is natural to assume that some less repugnant solution is available and optimal. The hero who decides to save her loved one, rather than 1000 innocent people, always manages to save *both* her loved one *and* any innocents. But it should be possible to write stories where these expectations are upended. Reading such stories might help philosophers better test moral theories such as utilitarianism. For example, we could consider various hyper-realistic stories, like those included in the computer game *This War of Mine*. In this game, the player controls a group of refugees trying to survive in a warzone. For example, a player may have to choose between taking in a small girl who would otherwise die or to push the group's resources over the brink, where the latter act would result in the death of them all. Such stories, which would include realistic and forced hard choices, may help us imagine that certain acts do in fact produce the most pleasure minus pain, however horrible they appear to us. Fictional works can similarly be produced that would help us passively fill out the details of Experience Machine, Transplant, and Utility Monster. My suspicion is that when these cases are situated in such fictional settings, our intuitions will become more friendly to utilitarianism – either because our moral intuitions change, or because their strength change.

5.5 Removing Details

The approach that I considered in the previous section tries to flesh out cases to make a central claim of a case – that an act maximizes pleasure minus pain – more natural to it, and so easier to passively imagine. But there is another approach available to us, which is to construct cases that have less room for our passive imagination to fill them out incorrectly. For example, consider the following case:

> MINIMAL TRANSPLANT
>
> Sarah stands on a ten by ten-meter platform located in empty space. There are no other objects in the universe, and all of existence will end in ten minutes. Six persons lie sedated in their beds on the platform in front of her. Sarah can use a hyper-advanced surgery tool to kill one patient and successfully save the lives of the five others. Upon doing so, the five will immediately wake-up and live a pain-free and pleasant existence for nine minutes. If Sarah does not operate, the five will die, but the sixth person will wake up and live a pain-free and pleasant existence for nine minutes. If Sarah operates, she maximizes pleasure minus pain. If she does not operate, she does not maximize pleasure minus pain.

When considering Minimal Transplant, there is less room for us to passively imagine adverse effects on society, such that people will fear using the healthcare system. Our moral intuitions about this case are also, I presume, more friendly to utilitarianism. For my own part, I intuit that Sarah should operate in these circumstances. But even if you intuit that operating is morally impermissible, your intuition is likely weaker than when you considered Transplant.

Examples like Minimal Transplant, just like does Revised Transplant, show that we can avoid the problem of aspects of a case being unnatural to it by making the case not more realistic, but less realistic. Many look down on unrealistic thought experiments as being unserious or unnecessary. But the above discussion suggests one reason to employ unrealistic thought experiments in our thought experimental practices: to make cases easier to imagine correctly, and so avoid misimagining them.

5.6 Objections and Responses

Let me end this chapter by considering a few potential objections to the arguments that I have put forward, as well as responses to these objections.

Objector: “You are too pessimistic about my ability to avoid misimagining cases. If I try hard enough, I can successfully imagine the cases you describe. So I do not need to add to or remove any details from the cases that you discuss.”

Response: I am not claiming that we always misimagine cases like Experience Machine, Transplant, and Utility Monster. I am arguing that we sometimes misimagine them and that it is not always transparent to us *when* we do so. Moreover, I worry that we do not always try hard to imagine cases correctly, even if we have the ability to successfully imagine them with enough effort. We often seem to imagine cases carelessly, quickly, and automatically. Finally, if you do in fact successfully imagine Transplant, then you should find that your intuitions do not differ between this case and that of Revised Transplant. But is that really so? I suspect that for many of us, our intuitions *do* differ between these cases. And one explanation for *why* they differ is that we do *not*, in fact, successfully imagine Transplant.

Objector: "I disagree – that is not the best explanation for why we have different intuitions on the basis of considering these cases. While I have different moral intuitions in response to the revised cases, that is primarily because you introduce additional morally relevant factors. So it is not surprising that my moral intuitions differ – they differ because the morally relevant factors differ."

Response: I agree that this is another possible explanation for why our intuitions differ between these cases. Even so, we cannot escape introducing potentially morally relevant factors to cases. Just because we do not *stipulate* them, we may still introduce them when we fill out cases via our passive imagination. For example, when I imagine Transplant, I imagine Sarah standing in a cellar with scalpel in hand. Surely, any number of such unintentionally added features are potentially morally relevant. That features are not explicitly stipulated does not mean that they are not being imagined. So, I would argue, the more detailed cases are not worse off in this respect than the less detailed ones. Furthermore, the detailed cases are better off in one respect: they help us avoid misimagining the cases.

Objector: "But why not simply get rid of the details in Transplant and the other original cases?"

Response: Because the details are needed for us to form intuitions about these cases. As Hewitt notes, if "our feelings were always responsive to […] abstract stipulation […] we would not need to appeal to concrete thought experiments in the first place."[124] For example, consider the thought experiment: "Imagine that a person *P* can maximize pleasure minus pain by performing an act *A*, but not by performing its (only) alternative act *A**, and that *A* is an act of prototypical violence. Should *P* perform *A*?" In response to this case, you will presumably not elicit any moral intuitions that conflict with utilitarianism. The case as described is too abstract. This shows that objections to utilitarianism are not just incidentally adding details to cases – such as transplants, doctors, patients, and so on – they essentially depend on them. Since we have to add at least some details to elicit moral intuitions, we should take care to only add details that are natural to the case as described.

Objector: "I worry that your proposal is too radical. If you are right, then all thought experiments in philosophy become problematic."

[124] Hewitt (2010), p. 339.

Response: I think the proposal is far less radical than one might believe. The problem of misimagination concerns only cases that include stipulated propositions that are *unnatural* to them. Many cases that are used in philosophy, such as the famous Gettier thought experiments, are clearly not of this kind. Moreover, consider a strange thought experiment proposed by Lippert-Rasmussen. Lippert-Rasmussen notes that "it is possible to imagine worlds in which forced donations of one's organs need in no way undermine one's ability to control one's life" and "in which the forced donation of most of a person's body will actually enhance his autonomy."[125] He then asks us to consider the following case:

> [Imagine] that people are born with huge bodies they can barely move, bodies with two hundred legs and arms. At any given moment, they can at best sense and control 1 percent of their bodies, although they can readily determine which percent that is. Since their bodies heal very easily, their ability to control their lives is promoted best if 99 percent of each body is removed in such a way that these abnormal individuals end up with what are, for us, normal human bodies.[126]

This example is highly unrealistic and unusual, but nothing that is stipulated in it is *unnatural* to the case as described: it is not unnatural to assume that for the individuals we are told about, removing 99% of their bodies *will* enhance their autonomy; on the contrary, it would be unnatural to assume that doing so would *not* enhance their autonomy. But if even such extremely unrealistic cases can be unproblematic, then surely there is no general problem with our thought experimental practices here. There is only a specific problem with cases such as Experience Machine, Transplant, and Utility Monster, which upend our expectations about what would naturally be considered true in these cases.

To conclude, I propose that the utilitarian can give a partial response to intuitive objections by pointing out that, in cases such as Experience Machine, Transplant, and Utility Monster, certain assumptions are unnatural to the cases as they are described. Next, the utilitarian can argue that we tend to misimagine such cases – and to instead imagine cases which, unlike the ones we intended to imagine, do not give rise to moral intuitions that are problematic for utilitarianism. Finally, the utilitarian can propose revised variants of the original cases – either by adding or removing details – which contain less or no such unnatural assumptions, and which are also more favorable to utilitarianism. In this chapter, I have considered some such revised cases, including Revised Experience Machine, Revised Transplant, Revised Utility Monster, and Minimal Transplant. To be sure, the end result is not a conclusive vindication of utilitarianism, but it does make the theory seem more plausible.

[125] Lippert-Rasmussen (2008), p. 109.
[126] Lippert-Rasmussen (2008), p. 109.

6. Sensory Imagination

In this chapter, I present and defend a theory about how sensory imagination affects the epistemic trustworthiness of moral intuitions. I argue that if this theory is true, then utilitarianism can be partially defended against some intuitive objections.

6.1 Sensory Imagination and Intuitive Objections

This chapter offers a partial defense of utilitarianism against some intuitive objections which are formed on the basis of thought experiments. In the discussion that follows, I appeal to the importance of *sensory imagination* in thought experimentation. As I noted in the previous chapter, while propositional imagination is about *supposing* or *assuming* something to be the case, sensory imagination is about bringing up a picture of the thing or about presenting the thing to yourself using a "sensory mood." The general idea behind this chapter can be summarized as follows. Depending on *how* we imagine cases, we might form different moral intuitions on their basis. So it is important to know whether certain *ways* of imagining cases make our moral intuitions any more or less epistemically trustworthy. As I will argue, sensory imagination of certain features of cases enhance the epistemic trustworthiness of our moral intuitions. Moreover, in several such cases, the right kind of sensory imagination makes our moral intuitions more supportive of utilitarianism.

Let me first explain how my argument is a *defense* of utilitarianism. An analogy is helpful here. Suppose that I propose to you an empirical theory – the *no boat theory* – according to which there are no boats on the Swedish lake *Storsjön*. To see whether the no boat theory is false, you can watch for boats at Storsjön: if you find a boat, then you have falsified my theory. So, let us assume that every afternoon you watch the lake, although you do this from afar and in foggy weather. Suddenly, you have the visual impression of a boat traversing Storsjön. Excited, you move closer to take a better look at the boat. But when doing so, you no longer have the visual impression of a boat; you see only a large grey rock that has the approximate size of a small boat. Clearly, you have been fooled. Now, this second visual impression constitutes a defense of the no boat theory, *even if* the first visual impression still gives you *some* reason to believe that a boat is on Storsjön. Even if the evidential value of the first visual impression is not *cancelled out* by the evidential value of the second visual impression, it is at least *outweighed* by it. My defense of utilitarianism is analogous to such a defense of the no boat theory. I suggest that in some cases, sensorily imagining features of cases increases the epistemic trustworthiness of our moral intuitions. On my view, to merely propositionally imagine a case is like watching Storsjön from afar in foggy weather. That is, just as we can watch a lake under *bad perceptual conditions,* we can imagine a case under *bad imaginary conditions*. Sensorily imagining a case is like moving closer to the lake – we improve the imaginary conditions by the use of sensory imagination, just like we improve the perceptual conditions by moving closer to take a better look. A visual impression formed on the basis of bad perceptual conditions might give me *some* reason to reject the no boat theory, even if I have *another* visual impression that is formed on the basis of better perceptual conditions, and so gives me *greater* reason to accept this theory. Similarly, a moral intuition formed on the basis of bad imaginary conditions might give me *some* reason to reject utilitarianism, even if I also have *another* moral intuition that is formed on the basis of better imaginary conditions, and so gives me *greater* reason to accept this theory.

The view that sensory imagination can improve the epistemic trustworthiness of our moral intuitions strikes me as a common-sense view – although when it is expressed precisely it will have to be qualified in various ways, as we will see in the next section. For example, human rights advocates typically try to make us imagine what it is like to be tortured by describing the atrocities in ways that prompt us to sensorily imagine them taking place. Intuitively, when we carry out such episodes of sensory imagination, we are not merely becoming more convinced about the atrocities being wrong, but also attain more reason to believe that they are wrong.

Moreover, when morally deliberating, we often try to put ourselves "in the shoes of others." That is, we sensorily imagine being them or seeing the world "from their eyes." Many philosophers have expressed ideas that imply that sensory imagination plays such an important role for learning about what is right or wrong. For example, Derek Parfit writes with respect to the utility monster thought experiment that:

> Nozick tells us to suppose that this imagined person [the utility monster] would be so happy, or have a life of such high quality, that this is the distribution that produces the greatest sum of happiness, or the greatest amount of whatever makes life worth living. [...] For this to be true, this Monster's quality of life must be millions of times as high as anyone we know. Can we imagine this? Think of the life of the luckiest person that you know, and ask what a life would have to be like in order to be a million times as much worth living. [...] Act Utilitarians might say that, if we really could imagine what such a life would be like, we might not find Nozick's objection persuasive.[127]

[127] Parfit (1984), p. 389.

In this quote, Parfit must mean that we cannot *sensorily* imagine the monster's quality of life, because we can clearly *propositionally* imagine it. The utilitarian could argue, Parfit seems to suggest, that if we successfully sensorily imagined the monster's quality of life, we would no longer find it counterintuitive that it should be given all our resources. Presumably, Parfit would also argue that once we sensorily imagine the monster's quality of life successfully, our moral or evaluative intuitions will also become more epistemically trustworthy.

More generally, the idea of "putting ourselves in the shoes of another" – which suggests sensorily imagining being in her shoes – when considering moral issues has been adopted in various ways by many philosophers, including by Richard Hare in formulating his theory of preference utilitarianism, and by John Rawls in proposing his "veil of ignorance."[128]

6.2 The Sensory Imagination Theory

The theory that I have in mind can be stated as follows:

> THE SENSORY IMAGINATION THEORY
>
> Everything else being equal, the higher the degree to which a person has sensorily imagined the morally relevant features of an act, the higher is the epistemic trustworthiness of her moral intuitions about this act.

To see how this theory applies to a concrete case, consider a prison guard who must decide whether to put a prisoner in solitary confinement. Before deciding what to do with the prisoner, the guard sensorily imagines the fear, loneliness, and anxiety which would be caused by confinement. After this episode of imagination, the guard has a moral intuition that it is wrong to put the prisoner in confinement. If the sensory imagination theory is true, the guard's moral intuition gives her more reason to believe that this act is wrong than it would had she not sensorily imagined the prisoner's suffering.

The sensory imagination theory must be clarified and qualified in several ways. First, I will assume that the added epistemic trustworthiness from sensory imagination can go beyond learning non-moral facts. Accordingly, if we compare the first prison guard with a second guard – who knows the same non-moral facts, but has only propositionally imagined what it is like to be in solitary confinement – the first guard will have more trustworthy moral intuitions. That is, I assume that sensory imagination can improve the trustworthiness of moral intuitions *over and above* helping us learn non-moral facts.

[128] Hare (1981), Rawls ([1971] 2005), pp. 136-142.

Thus, sensory imagination helps us not only see *more* of a case, but to see a case *more clearly*. (That being said, my arguments are likely relevant to defending utilitarianism even *if* sensory imagination only helps us learn new non-moral facts.)

Second, the sensory imagination theory is compatible with an episode of sensory imagination *on the whole* subtracting from the epistemic trustworthiness of a moral intuition. The theory only claims that the epistemic trustworthiness is increased *everything else being equal*. For example, suppose that you visualize a terrible murder. You may be so shaken by this visualization that you cannot "think straight" and so reflexively condemn the murder. In this case, your moral intuition that the murder is wrong might on the whole be less epistemically trustworthy because you sensorily imagined it. But the use of sensory imagination *as such* still adds to the epistemic trustworthiness of the intuition. What epistemic weight is added by the sensory imagination is simply cancelled out by an unintended side effect: your inability to think straight. This is also true for other distorting influences that can be prompted by the use of sensory imagination, such as overwhelming emotional bursts of empathy, anger, sadness, and love. Similarly, to improve one's perceptual conditions (e.g., by taking a closer look at an object) might bring about strong emotional reactions. These reactions might lower the epistemic trustworthiness of our perceptual sightings *on the whole,* even if the improved perceptual conditions *as such* still improve their trustworthiness.

Third, the sensory imagination theory is not only relevant to *thought experimentation*. For example, suppose that an animal activist knows all the facts about how animals suffer in factory farms. Compare her to a second activist with the same knowledge, but who has additionally read a book that vividly describes the conditions in factory farms. Reading the book prompted her to sensorily imagine what it is like being an animal in a factory farm, including the pain, depression, and boredom from isolation. Intuitively, the second animal activist's moral intuitions are more trustworthy in this case, even if she did not carry out a thought experiment. The sensory imagination theory accounts for this improved trustworthiness.

Fourth, the sensory imagination theory states that sensory imagination of *morally relevant features* of acts increases the trustworthiness of moral intuitions about those acts – not just any sensory imagination will do. For example, sensorily imagining the details of a murderer's shoelaces makes no difference. Moreover, it is sensory imagination of morally relevant features *of acts* that is needed, as a feature may be morally relevant *in a thought experiment* even if it is not a morally relevant feature *of an act* being imagined in that thought experiment.

Fifth, in formulating the sensory imagination theory, "morally relevant features" refer solely to *fundamental* and *non-derivative* morally relevant features. For example, whether the prisoner in the above case is adequately clothed is a non-fundamentally morally relevant feature of the case. It is non-fundamentally morally relevant because it is morally relevant only in so far as it brings about something that is morally relevant, such as pleasure or the absence of pain. Therefore, to sensorily imagine the prisoner's clothes will typically not improve the epistemic trustworthiness of the guard's intuition. But there is one exception: the sensory imagination theory should be understood as allowing for *two* ways of sensorily imagining fundamental and non-derivative morally relevant features. On the one hand, one can *directly* sensorily imagine such features, as when the guard sensorily imagines the pleasure and pain felt by the prisoner. On the other hand, one can *indirectly* sensorily imagine these features by directly sensorily imagining features that are, by the imaginer, *closely associated* with the morally relevant features. For example, the guard can sensorily imagine facial expressions signaling pain or screams of pain from the prisoner, and this will let her indirectly sensorily imagine the prisoner being in pain. Similarly, to sensorily imagine the prisoner being adequately clothed may constitute a way for the guard to indirectly sensorily imagining the prisoner feeling well and comfortable, and may so increase the epistemic trustworthiness of the guard's moral intuition.

Sixth, we may ask *when* the relevant sensory imagination has to take place, relative to the formation of the moral intuition, in order to increase the epistemic trustworthiness of that intuition. The sensory imagination clearly needs to be in the present or past – not in the future. But how far in the past does sensory imagination count? For example, suppose that the prison guard sensorily imagined what solitary confinement is like ten years ago, and only now forms the intuition that confinement of a prisoner is wrong. Does that episode of sensory imagination increase the trustworthiness of her intuition? Some kind of causal criterion seems reasonable in this case. For example, we might say that the sensory imagination counts only if it is part of the (immediate or close) cause of the moral intuition.[129]

6.3 Supporting the Theory

A first piece of evidence that favors the sensory imagination theory consists of various *epistemic* intuitions, namely intuitions about how, in certain circumstances, the trustworthiness of our moral intuitions changes. For example, in cases like that of the prison guard, it seems very plausible that sensory imagination raises the trustworthiness of moral intuitions. Similarly, in the case of the animal activist, it seems plausible that her sensory imagination of the suffering of factory farm animals raises the trustworthiness of her moral intuitions. In general, we seem to think that imagining the inner lives of other people can help us learn about our moral obligations to them, beyond learning new non-moral facts – the sensory imagination theory explains why this is so.

[129] Thanks to David Alm for this suggestion.

Finally, we may also have epistemic intuitions about what role sensory imagination plays outside of thought experimenting. For example, it seems plausible that fictional works of literature can teach us about what is morally right or wrong. If sensory imagination can make moral intuitions about poverty, torture, charity, helping others, and so on more epistemically trustworthy, that could explain why reading fiction can teach us about what is right or wrong.

The sensory imagination theory is also supported by a comparison to the importance of *sensory perception* in making our moral intuitions more trustworthy. To begin with, it seems plausible to think that:

(1) Sensory perception makes our moral intuitions more epistemically trustworthy.

For example, aid workers who have spent some time in poor countries will also in general have more epistemically trustworthy moral intuitions about whether, for example, it is morally obligatory to give more money to stop poverty. People who themselves have felt the effects of oppression are at least in some respects better at judging whether it is wrong to oppress others. Moreover, this is not merely because these people have greater knowledge of relevant non-moral facts. Had they had the same non-moral knowledge, but lacked the relevant sensory perception, their moral intuitions would not have been as trustworthy. It is because they have *seen* the relevant acts performed up front that their intuitions are more trustworthy. Similarly, it seems plausible that seeing documentaries or playing certain computer games, both of which involve sensory perception, can increase the trustworthiness of our moral intuitions beyond letting us learn non-moral facts. This too is explained by (1).

Now, there are clear similarities between sensory perception and sensory imagination. To begin with, both are intentional states. Just as I see *that* something is the case, I sensorily imagine *that* something is the case. Moreover, both involve a phenomenal component, a sensory experience. But most importantly, what makes sensory perception epistemically important is that things are *presented to us in a sensory way*. Since things are presented to us in a sensory way also in the case of sensory imagination, we should expect sensory imagination to be similarly epistemically important. That is, it seems plausible that:

(2) If sensory perception makes our moral intuitions more epistemically trustworthy, then so does sensory imagination.

Taken together, (1) and (2) support the sensory imagination theory.

We can also argue for the sensory imagination theory by means of a second analogy – this time not to sensory perceptions, but to *physical intuitions*. For example, in arguing that thought experiments are mental models, Tamar Gendler claims that sensory imagination in thought experiments allows various non-moral intuitions about physical spaces to become more trustworthy (although she does not use this piece of terminology to state her point, it seems to describe her position correctly).[130] In her article, Gendler asks us to:

> Think about your next-door neighbor's living room, and ask yourself the following questions: If you painted its walls bright green, would that clash with the current carpet, or complement it? If you removed all its furniture, could four elephants fit comfortably inside? If you removed all but one of the elephants, would there be enough space to ride a bicycle without tipping as you turned?[131]

Gendler points out that we typically use sensory imagination to answer these questions:

> Perhaps you believed (perhaps tacitly) that *some* indoor spaces are too small to ride a bike in (closets, for instance), and that others (banquet halls, for instance) are certainly large enough – but did you have, even tacitly, beliefs about where the border between these lay, and, in particular, beliefs about where your neighbor's living room stood with respect to that border? Didn't you, instead, *discover* something about bikes and living rooms by *imagining having a certain experience*? Likewise with the color case. While you may have known beforehand that your neighbor's rug looks like *this*, and that green looks like *that*, was it really a matter of deductive or inductive inference that led you to the conclusion that – were they adjacent – you would judge them to clash? Wasn't it instead as if you performed an *experiment-in-thought*, on the basis of which you got some new information about your own judgments, which (perhaps because of tacit beliefs that you hold) you took to be relevant data in answering the question at hand?[132]

Similar ideas can be found in the writings of Nancy Nersessian, who is another defender of thought experiments as mental models.[133] From this picture of mental modelling and the importance of sensory imagination, it seems plausible to think that:

(3) Sensory imagination makes our physical intuitions more epistemically trustworthy.

Moreover, the *way* in which sensory imagination makes our physical intuitions more trustworthy in these cases seems analogous to how sensory imagination (according to the sensory imagination theory) makes our moral intuitions more trustworthy. For example, you can learn whether a bed fits through a door by sensorily imagining trying to move it past the doorway, forming a picture of it moving through, adding room for persons to lift it, and so on.

[130] Gendler (2004). See especially pp. 1156-1161.
[131] Gendler (2004), p. 1156.
[132] Gendler (2004), p. 1159.
[133] Nersessian (2007).

Your physical intuition that the bed does or does not fit through the door will clearly be more trustworthy if you have in a detailed way visualized the widths and heights of the door and the bed. That is, your physical intuitions are more epistemically trustworthy if you have sensorily imagined the *physically relevant features* of the door and bed. With this in mind, it seems reasonable to think that:

> (4) If sensory imagination makes our physical intuitions more epistemically trustworthy, then it makes our moral intuitions more trustworthy as well.

And taken together, (3) and (4) support the sensory imagination theory.

6.4 Morally Relevant Features

We can sensorily imagine morally relevant features not only directly and indirectly, but also by means of different sensory moods. These moods include not only the visual, but also the olfactory, gastronomical, tactual, auditory, and proprioceptional moods. They include imagining sensations, such as unpleasant sensations of sadness and anxiety, physical pains such as sharp and dull pain, and pleasant sensations such as elation and satisfaction. Just as we can visualize a mountainside, we can sensorily imagine hearing sounds, feeling pressure, tasting sweetness, fearing a wolf, feeling pain, and so on. In many of these non-visual cases, our language lacks imagination counterpart terms, as "visualizing" is the imagination counterpart term to seeing. Therefore, when applying the sensory imagination theory, we may want to invent new counterpart terms and speak of auditializing, tactializing, tastializing, and so on to refer to the relevant phenomena.

Now, what are the (fundamentally and non-derivatively) morally relevant features of acts? On one view, they are the features that (fundamentally and non-derivatively) matter for the act's moral evaluation. Accordingly, different moral theorists may have different opinions on the matter. But even so, there is a strong case to be made for pleasure and pain making up at least *part* of what is fundamentally and non-derivatively morally relevant. Even many non-utilitarians agree that pleasure and pain is fundamentally morally relevant, including Rossians and consequentialist pluralists, as well as object preferentialist utilitarians. These theories disagree with utilitarianism about whether *only* pleasure and pain are fundamentally morally relevant, and *how* these morally relevant features affect the deontic status of acts. Moreover, if a moral theory cannot account for how pleasure and pain is fundamentally morally relevant, that seems to count strongly against such a view. So it seems reasonably to think that, at the very least, sensorily imagining the pleasure and pain resulting from an act will improve the epistemic trustworthiness of our moral intuitions. This is so even if there are many additional morally relevant features of acts that we may also sensorily imagine in order to improve the epistemic trustworthiness of our moral intuitions.

6.5 Defending Utilitarianism

Having granted that sensorily imagining the pleasure and pain resulting from acts will improve the trustworthiness of our moral intuitions, let us now return to Revised Transplant. In this case – the one where Sarah runs an autonomous hospital in a barren wasteland – we sensorily imagine that the five patients wake up and feel relief after a successful operation, knowing that their lives are saved. We then indirectly sensorily imagine them living pleasant lives many years afterwards, such as by visualizing them being transported out of the desert, settling down in a village, forming families, enjoying their vocations, and so on. When contrasting this outcome with that of not killing patient Six and letting the five die, we sensorily imagine only patient Six walking out of the desert, settling down in a village, and so on. When I have imagined the case in these different ways, sensorily imagining the pleasure and pain resulting from performing the act or its alternative, I more strongly intuit that it would be obligatory for Sarah to kill patient Six in order to save the five other patients.

Let us now approach Revised Experience Machine in the same way. We sensorily imagine that John is plugged into the experience machine, having a large number of pleasant sensations. We sensorily imagine that he sees wonderful sunsets, experiences deep happiness, feels love and relaxation, and enjoys focusing on the meaningful activities that he most loves doing. Finally, we sensorily imagine John instead living in the outside world. We imagine the same sensations, but as slightly less pleasurable. The sunset is less bright, the happiness less deep, the love is less intense, and so on. In this case, I more strongly intuit that it is obligatory for William to plug John into the machine, as compared to when I did not sensorily imagined these features of the case.

In contrast to Revised Transplant and Revised Experience Machine, sensory imagination makes less of a difference to the two utility monster cases that I have considered – both the revised and the original one. These cases involve the pleasures of many people, which is difficult for us to sensorily imagine.[134] Moreover, the original case involves pleasures which are so intense that we *cannot* sensorily imagine them. Therefore, it seems unlikely that the sensory imagination theory can help us defend utilitarianism against the objections raised by these cases. That said, if the sensory imagination theory is

[134] Several philosophers have pointed out the problem of imagining large number cases. See Tännsjö (2002), p. 344, Broome (2004), pp. 56-59, 210-214, and Huemer (2008), p. 908. For discussion, see Pummer (2012).

true, this means that we will also *lack* one way of increasing the trustworthiness of our moral intuitions about such cases, because we cannot sensorily imagine some of their features. This suggests that cases which are more easily sensorily imagined should play a proportionally greater role in the evaluation of utilitarianism and other moral theories, as these cases can provide us with more trustworthy moral intuitions.

Finally, note that in the above cases I considered how the *strength* of our moral intuitions may change when sensorily imagining the morally relevant features of cases. But that is not the only way in which the sensory imagination theory could aid utilitarianism. First, that moral intuitions become more trustworthy can itself make utilitarianism more plausible. Second, the relevant episodes of sensory imagination may cause us to *change* our moral intuitions in favor of utilitarianism. For example, after sensorily imagining Revised Transplant, we might intuit that Sarah should kill patient Six; while before we imagined the case in this way, we intuited that she should let the five die. In that case, even if the new intuition is weaker than the original one, the result is still that utilitarianism becomes more plausible.

6.6 Objections and Responses

Finally, let me consider some potential objections to my arguments.

Objection: "I do not share your intuitions. So I remain unconvinced by this defense of utilitarianism."

Response: Clearly, that I rely on only my own intuitions is a weakness in the above arguments. I would be happier if I had had access to systematic, empirical data on whether and how people's moral intuitions change in response to the relevant use of sensory imagination. Perhaps this is a matter for future research. But that said, I would be surprised if a significant number of people did not *somewhat* change their intuitive responses when sensorily imagining pleasures and pains in these ways, and for their responses to become more friendly to utilitarianism.

Objection: "The view that sensory imagination yields intuitions with greater epistemic trustworthiness is question-begging. It is stacked in favor of those that put relatively much stock in facts about experiences, such as utilitarianism."

Response: The sensory imagination theory does indeed support utilitarianism, but appealing to this theory in defense of utilitarianism is not question-begging. To begin with, I have argued that the sensory imagination theory is independently plausible – I suggested several arguments for the sensory imagination theory in Section 6.3, and none of the arguments depended on utilitarianism being true. Moreover, for the sensory imagination theory to aid in the defense of utilitarianism, we need only assume that pleasures and pains are fundamentally morally relevant. As I noted earlier in this chapter, many non-utilitarians will agree with this claim.

Objection: "Is it not more reasonable to require that we imagine the *potentially* morally relevant features of acts, rather than their *actually* morally relevant features?"

Response: As I have formulated the sensory imagination theory, it is compatible with us *also* increasing the epistemic trustworthiness of moral intuitions by sensorily imagining the potentially morally relevant features of acts. That is, the theory gives us *one* way in which sensory imagination can increase the epistemic trustworthiness of our moral intuitions. This might not be the only way, however. In any case, the potentially morally relevant features will even more obviously include pleasure and pain. So this only strengthens the case for utilitarianism.

In philosophical practice, where we are uncertain about which features are morally relevant, we will want to construct thought experiments so that a large number of theorists can benefit from thinking about them. Conveniently, we can probably use one and the same episode of sensory imagination to let us indirectly sensorily imagine several different potentially morally relevant features of an act. For example, we can indirectly sensorily imagine a person being in pain and having her rights violated by directly sensorily imagining an act of physical violence.

Objection: "Using the above methodology, utilitarianism becomes less plausible. For example, consider the case of Transmission suggested by Tim Scanlon.[135] Jones has suffered an accident in a TV transmitter room and is in incredible pain. To help him, we must interrupt the transmission. But doing so will cause millions of television viewers to receive slightly less pleasure, and so this alternative act produces less than maximal pleasure minus pain. Therefore, utilitarianism tells us to continue the transmission and to leave Jones in pain. Intuitively, of course, we should interrupt the transmission. Moreover, sensorily imagining Jones' pain will actually make us *less* inclined to have intuitions which accord with utilitarianism. Sensorily imagining Jones' pain strengthens our moral intuition that it is obligatory to interrupt the transmission, and this moral intuition will also (per the sensory imagination theory) become more epistemically trustworthy."

Response: I admit that cases like these will continue to trouble utilitarianism for the foreseeable future, and that in some such cases our anti-utilitarian intuitions become both stronger and more epistemically trustworthy. But that said, I would argue that cases like these are ultimately not optimal for testing moral theories, simply because we cannot sensorily imagine all of the morally relevant features in them, such as large numbers of people receiving a certain amount of pleasure and pain. Moreover, when I *do* try to sensorily imagine many sensations of, for example, slight disappointment (resulting from having

[135] Scanlon (2000), p. 235.

the transmission turned off) and comparing that to the greatly lessened pain for Jones, the result is not that I become ever more certain that it is morally right to turn off the transmission the more disappointments that I imagine. Rather, I find it more and more difficult to elicit any intuitions about the case whatsoever. My intuitive faculties seem to have a hard time comparing the large decline in pain to the many small disappointments being imagined. It is like my brain just gives up trying to compare these sensations. In so far as others share my reactions to this case, the implications for utilitarianism are admittedly not fantastic – it would be better had we intuited that we *should* turn of the transmission. But neither are they any worse than before we used our sensory imagination to engage with the case.

7. Conclusions

Almost 100 years ago, Ralph Blake wrote a defense of hedonistic act utilitarianism, which he introduced with the following remarks:

> In current discussions of ethical theory it has become the tradition to treat [utilitarianism] on the more or less definite assumption that its rejection is a foregone conclusion. Its falsity is so thoroughly taken for granted that it is thought possible to dispose of its claims in very short order indeed. It is usually treated with ill-disguised contempt as an antiquated heresy that has so long since been definitely refuted that its truth can scarcely be contemplated as a genuine possibility at all. Its "fallacies" are summarily pointed out, in a few brief pages, and we are hurried on to a consideration of theories more worthy the attention of a mature mind.[136]

I imagine that had Blake been alive today, he would have been glad to learn that, for the most part, we have left behind the objections that he complained of as "simply puerile." Today's objections to utilitarianism are more sophisticated than they were a hundred years ago. That said, I suspect he would still be disappointed to learn that a century later the theory's "falsity" is "so thoroughly taken for granted that it is thought possible to dispose of its claims in very short order." I hope to have made some contribution to dispelling such pessimistic views. While utilitarianism may be false, it is not *obviously* false, and many objections that are directed against it become far less convincing upon closer examination.

Let me sum up what has been achieved in this book. In the previous six chapters, I considered two types of objections to utilitarianism, and partially defended the theory against both. I first discussed objections to the effect that utilitarianism is not action guiding. In Chapter 2, I distinguished between two types of action guidance, which I called doxastic and evidential guidance. Roughly, for a theory to be doxastically guiding is for it *to be able to teach us what to do*, whereas for a theory to be evidentially guiding is for it *to be practically relevant to us in decision making*. I gave more precise definitions of these kinds of action guidance, and I also argued that while utilitarianism is not doxastically guiding, it is at least evidentially guiding. The distinction between doxastic and evidential guidance became useful in Chapter 3, where I considered several action guidance objections to utilitarianism. I identified

[136] Blake (1926), p. 1. Note that in his article, Blake refers to hedonistic act utilitarianism as "hedonism."

five such objections, and sometimes considered multiple interpretations of a given objection. An important difference between these objections lie in their *conclusions*, which include that utilitarianism is a *bad* moral theory, that it *is not even a moral theory*, that it is an *uninteresting or unimportant* moral theory, and, most importantly, that it is a *false* moral theory. I concluded that none of these objections are promising upon closer examination. That said, my discussion did not give a complete defense against action guidance objections to utilitarianism, because I did not consider purely *meta-ethical* action guidance objections. Moreover, I admit, there might exist action guidance objections that we have not yet found, or that we have not formulated clearly enough.

Next, I discussed objections to the effect that utilitarianism conflicts with our moral intuitions. In Chapter 4, I presented three thought experiments – Experience Machine, Transplant, and Utility Monster – together with corresponding intuitive objections. I also evaluated and criticized three strategies for responding to these objections: scope restriction, debunking arguments, and non-conscious application. In Chapters 5 and 6, I moved on to consider two alternative defenses of utilitarianism against intuitive objections, which I considered more promising. In Chapter 5, I discussed the idea that we might fail to carry out the right thought experiments when we test utilitarianism against our moral intuitions – that we might *misimagine* these cases. I considered two ways in which we can revise thought experiments to avoid misimagining cases, and I showed that on both approaches the intuitive objections to utilitarianism either fail or become weaker. In Chapter 6, I argued that when we sensorily imagine the morally relevant features of the cases being considered, our moral intuitions become friendlier to utilitarianism.

Just as with the action guidance objections, my defense of utilitarianism against the intuitive objections was merely partial. That said, I think that once we consider the possibility of misimagining cases and the importance of sensory imagination, and once we adjust our philosophical and methodological practices accordingly, utilitarianism looks like a more plausible view. It certainly does not deserve the peripheral place it has been given in recent discussions of consequentialist theories.

Let me end this book with a general observation. The idea that moral theories must be action guiding seems to require a meta-ethical picture on which morality is less "objective" and more "subjective." If moral facts are like scientific facts about atoms and molecules, then a moral theory that is perfectly action guiding would appear, for that very reason, suspect – because what is the chance that, of everything that could be morally relevant in the universe, it is *precisely* that which we can easily learn about and use for applying the true moral theory? Indeed, if morality is "objective," one would suspect that an action guiding moral theory just reflects our ingrained prejudices. On the other hand, the intuitive objections seem most plausible on a meta-ethical picture where morality is more "subjective" and less "objective." The intuitive objections depend on our moral intuitions being epistemically trustworthy, and the best argument for their being so draws upon a comparison to our perceptual faculties – i.e., how we by sight, smell, and hearing perceive an objective reality. Action guidance objections therefore seem to require a radically different view of moral reality than do intuitive objections – one on which moral facts are "up to us." Intuitive objections, in contrast, seem to require a view on which moral facts are rather "up to us to find." For these reasons, I suspect that irrespective of the ultimate fate of these two groups of objections against utilitarianism, it will be difficult to make *both* work.

References

Andrić, Vuko. 2016. “Is Objective Consequentialism Compatible with the Principle That ‘Ought’ Implies ‘Can’?” *Philosophia*, 63–77. https://doi.org/10.1007/s11406-015-9668-5.

———. 2017. “Objective Consequentialism and the Rationales of ‘“Ought” Implies “Can.”’” *Ratio* 30 (1): 72–87. https://doi.org/10.1111/rati.12108.

Aydede, Murat. 2014. “How to Unify Theories of Sensory Pleasure: An Adverbialist Proposal.” *Review of Philosophy and Psychology* 5 (1): 119–33. https://doi.org/10.1007/s13164-014-0175-6.

Bailey, James Wood. 1997. *Utilitarianism, Institutions, and Justice*. New York: Oxford University Press.

Bales, R. Eugene. 1971. “Act-Utilitarianism: Account of Right-Making Characteristics or Decision-Making Procedure?” *American Philosophical Quarterly* 8 (3): 257–65.

Barrow, Robin. 1991. *Utilitarianism: A Contemporary Statement*. Aldershot: Edward Elgar Publishing Limited.

Bealer, George. 1998. “Intuition and the Autonomy of Philosophy.” In *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*, 201–39. Lanman; Boulder: Rowman & Littlefield Publishers, Inc.

Bentham, Jeremy. (1780) 2008. *An Introduction to the Principles of Morals and Legislation*. London: Dodo Press.

Bergström, Lars. 1996. “Reflections on Consequentialism.” *Theoria* 62 (1–2): 74–94. https://doi.org/10.1111/j.1755-2567.1996.tb00531.x.

Blake, Ralph Mason. 1926. “Why Not Hedonism? A Protest.” *The International Journal of Ethics* 37 (1): 1–18.

Bradley, Ben. 2006. “Against Satisficing Consequentialism.” *Utilitas* 18 (2): 97–108. https://doi.org/10.1017/S0953820806001877.

Brandt, Richard. 1963. “Toward a Credible Form of Utilitarianism.” In *Morality and the Language of Conduct*, edited by Héctor-Neri Castañeda and George Nakhnikian, 107–43. Detroit: Wayne State University Press.

Brooks, D. H. M. 1994. “The Method of Thought Experiment.” *Metaphilosophy* 25 (1): 71–83. https://doi.org/10.1111/j.1467-9973.1994.tb00469.x.

Broome, John. 2004. *Weighing Lives*. Oxford; New York: Oxford University Press.

Burch-Brown, Joanna M. 2014. “Clues for Consequentialists.” *Utilitas* 26 (1): 105–119. https://doi.org/10.1017/S0953820813000289.

Bykvist, Krister. 2010. *Utilitarianism: A Guide for the Perplexed*. London; New York: Continuum International Publishing Group.

Carlson, Erik. 1995. *Consequentialism Reconsidered*. Dordrecht: Kluwer Academic Publishers.

———. 1999. “The Oughts and Cans of Objective Consequentialism.” *Utilitas* 11 (1): 91–96. https://doi.org/10.1017/S0953820800002284.

———. 2002. “Deliberation, Foreknowledge, and Morality as a Guide to Action.” *Erkenntnis* 57 (1): 71–89. https://doi.org/10.1023/A:1020146102680.

Cooper, Rachel. 2005. “Thought Experiments.” *Metaphilosophy* 36 (3): 328–47. https://doi.org/10.1111/j.1467-9973.2005.00372.x.

Cowen, Tyler. 2006. “The Epistemic Problem Does Not Refute Consequentialism.” *Utilitas* 18 (4): 383–399. https://doi.org/10.1017/S0953820806002172.

Dorsey, Dale. 2012. “Consequentialism, Metaphysical Realism and the Argument from Cluelessness.” *The Philosophical Quarterly* 62 (246): 48–70. https://doi.org/10.1111/j.1467-9213.2011.713.x.

Driver, Julia. 2012. *Consequentialism*. Abingdon; New York: Routledge.

———. 2014. “The History of Utilitarianism.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2014. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2014/entries/utilitarianism-history.

Eggleston, B. 2010. “Practical Equilibrium: A Way of Deciding What to Think about Morality.” *Mind* 119 (475): 549–84. https://doi.org/10.1093/mind/fzq036.

Elgin, Samuel. 2015. “The Unreliability of Foreseeable Consequences: A Return to the Epistemic Objection.” *Ethical Theory and Moral Practice* 18 (4): 759–66. https://doi.org/10.1007/s10677-015-9602-8.

Enoch, David. 2010. “The Epistemological Challenge to Metanormative Realism: How Best to Understand It, and How to Cope with It.” *Philosophical Studies* 148 (3): 413–38. https://doi.org/10.1007/s11098-009-9333-6.

Feldman, Fred. 1997. *Utilitarianism, Hedonism, and Desert: Essays in Moral Philosophy*. Cambridge; New York: Cambridge University Press.

———. 2004. *Pleasure and the Good Life: Concerning the Nature, Varieties and Plausibility of Hedonism*. Oxford; New York: Oxford University Press.

———. 2006. “Actual Utility, The Objection from Impracticality, and the Move to Expected Utility.” *Philosophical Studies* 129 (1): 49–79. https://doi.org/10.1007/s11098-005-3021-y.

———. 2019. “Two Visions of Welfare.” *The Journal of Ethics* 23 (2): 99–118. https://doi.org/10.1007/s10892-019-09287-1.

Fogal, Daniel, and Olle Risberg. Forthcoming. “The Metaphysics of Moral Explanation.” In *Oxford Study in Metaethics, Vol. 15*.

Frazier, Robert L. 1994. “Act Utilitarianism and Decision Procedures.” *Utilitas* 6 (1): 43–53. https://doi.org/10.1017/S095382080000131X.

Friedman, Marilyn. 1987. “Care and Context in Moral Reasoning.” In *Women and Moral Theory*. Totowa: Rowman & Littlefield.

Gendler, Tamar Szabo. 2000. “The Puzzle of Imaginative Resistance.” *The Journal of Philosophy* 97 (2): 55. https://doi.org/10.2307/2678446.

Gendler, Tamar Szabó, and Shen-yi Liao. Forthcoming. “The Problem of Imaginative Resistance *.” In *The Routledge Companion to Philosophy of Literature*, edited by Noël Carroll and John Gibson, 1st ed., 405–18. Routledge. https://doi.org/10.4324/9781315708935-35.

Gendler, Tamar Szabó. 2004. “Thought Experiments Rethought—and Reperceived.” *Philosophy of Science* 71 (5): 1152–63. https://doi.org/10.1086/425239.

Goldman, Alvin I. 1976. *A Theory of Human Action*. Princeton: Princeton University Press.

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

———. 2009. "Demandingness as a Virtue." *The Journal of Ethics* 13 (1): 1–13. https://doi.org/10.1007/s10892-007-9025-4.

Greaves, Hilary. 2016. "Cluelessness." *Proceedings of the Aristotelian Society* 116 (3): 311–39. https://doi.org/10.1093/arisoc/aow018.

Greene, Joshua David. 2013. *Moral Tribes: Emotion, Reason, and the Gap between Us and Them*. New York: The Penguin Press.

Gren, Jonas. 2004. *Applying Utilitarianism: The Problem of Practical Action-Guidance*. Göteborg: Acta Universitatis Gothoburgensis.

Häggqvist, Sören. 1996. *Thought Experiments in Philosophy*. Stockholm: Almqvist & Wiksell International.

Hare, Caspar. 2011. "Obligation and Regret When There Is No Fact of the Matter About What Would Have Happened If You Had Not Done What You Did." *Noûs* 45 (1): 190–206. https://doi.org/10.1111/j.1468-0068.2010.00806.x.

Hare, R. M. 1981. *Moral Thinking: Its Levels, Method, and Point*. Oxford; New York: Clarendon Press; Oxford University Press.

Harris, John. 1975. "The Survival Lottery." *Philosophy* 50 (191): 81–87. https://doi.org/10.1017/S0031819100059118.

Hattiangadi, Anandi. 2018. "Moral Supervenience." *Canadian Journal of Philosophy* 48 (3–4): 592–615. https://doi.org/10.1080/00455091.2018.1436034.

Haybron, Daniel M. 2010. *The Pursuit of Unhappiness: The Elusive Psychology of Well-Being*. Oxford; New York: Oxford University Press.

Hewitt, Sharon. 2010. "What Do Our Intuitions about the Experience Machine Really Tell Us about Hedonism?" *Philosophical Studies* 151 (3): 331–49. https://doi.org/10.1007/s11098-009-9440-4.

Hooker, Brad. 2000. *Ideal Code, Real World*. Oxford; New York: Oxford University Press.

Howard-Snyder, Frances. 1997. "The Rejection of Objective Consequentialism." *Utilitas* 9 (2): 241–248. https://doi.org/10.1017/S0953820800005306.

———. 1999. "Response to Carlson and Qizilbash." *Utilitas* 11 (1): 106–111. https://doi.org/10.1017/S0953820800002302.

Huemer, Michael. 2008. *Ethical Intuitionism*. Houndmills: Palgrave Macmillan.

Hurley, Paul E. 2011. *Beyond Consequentialism*. Oxford: Oxford University Press.

Jackson, Frank. 1991. "Decision-Theoretic Consequentialism and the Nearest and Dearest Objection." *Ethics* 101 (3): 461–82.

Kagan, Shelly. 1992. "The Limits of Well-Being." *Social Philosophy and Policy* 9 (2): 169–189. https://doi.org/10.1017/S0265052500001461.

———. 1998. *Normative Ethics*. Boulder: Westview Press.

Kahane, Guy. 2011. "Evolutionary Debunking Arguments." *Noûs* 45 (1): 103–125.

Kahneman, Daniel. 2013. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Labukt, Ivar. 2012. "Hedonic Tone and the Heterogeneity of Pleasure." *Utilitas* 24 (2): 172–99. https://doi.org/10.1017/S0953820812000052.

Lang, Gerald. 2008. "Consequentialism, Cluelessness, and Indifference." *The Journal of Value Inquiry* 42 (4): 477–85. https://doi.org/10.1007/s10790-008-9136-6.

Lazari-Radek, Katarzyna De, and Peter Singer. 2014. *The Point of View of the Universe: Sidgwick and Contemporary Ethics*. Oxford: Oxford University Press.

———. 2017. *Utilitarianism: A Very Short Introduction*. New York: Oxford University Press.

Lenman, James. 2000. "Consequentialism and Cluelessness." *Philosophy & Public Affairs* 29 (4): 342–70. https://doi.org/10.1111/j.1088-4963.2000.00342.x.

Lippert-Rasmussen, Kasper. 2008. "Against Self-Ownership: There Are No Fact-Insensitive Ownership Rights over One's Body." *Philosophy & Public Affairs* 36 (1): 86–118. https://doi.org/10.1111/j.1088-4963.2008.00125.x.

Lukas, Mark. 2008. "Reconciling Consequentialism with Ordinary Moral Knowledge." *UC Berkeley: Kadish Center for Morality, Law and Public Affairs*. https://escholarship.org/uc/item/58r6d1bs.

Mason, Elinor. 2003. "Consequentialism and the 'Ought Implies Can' Principle." *American Philosophical Quarterly* 40 (4): 319–31.

———. 2004. "Consequentialism and the Principle of Indifference." *Utilitas* 16 (3): 316–321. https://doi.org/10.1017/S0953820804001190.

———. 2013. "Objectivism and Prospectivism about Rightness." *Journal of Ethics and Social Philosophy* 7 (2): 1–21. https://doi.org/10.26556/jesp.v7i2.72.

McCloskey, H. J. 1973. "Utilitarianism: Two Difficulties." *Philosophical Studies* 24 (1): 62–63. https://doi.org/10.1007/BF00376360.

McConnell, Terrance. 1989. """Ought" Implies "Can"" and the Scope of Moral Requirements." *Philosophia* 19 (4): 437–454.

———. 2018. "Moral Dilemmas." In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2018. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2018/entries/moral-dilemmas.

McGinn, Colin. 2004. *Mindsight: Image, Dream, Meaning*. Cambridge: Harvard University Press.

———. 2009. "Imagination." In *The Oxford Handbook of Philosophy of Mind*, 595–606. Oxford; New York: Clarendon Press; Oxford University Press.

Meyers, C. D. 2015. "Brains, Trolleys, and Intuitions: Defending Deontology from the Greene/Singer Argument." *Philosophical Psychology* 28 (4): 466–86. https://doi.org/10.1080/09515089.2013.849381.

Mill, John Stuart. (1871) 2007. *Utilitarianism*. New York: Dover Publications, Inc.

Miller, Dale E. 2003. "Actual–Consequence Act Utilitarianism and the Best Possible Humans." *Ratio* 16 (1): 49–62. https://doi.org/10.1111/1467-9329.00205.

Miščević, Nenad. 1992. "Mental Models and Thought Experiments." *International Studies in the Philosophy of Science* 6 (3): 215–26. https://doi.org/10.1080/02698599208573432.

———. 2007. "Modelling Intuitions and Thought Experiments." *Croatian Journal of Philosophy* 7 (2): 181–214.

Moore, Eric. 2006. "Objective Consequentialism, Right Actions, and Good People." *Philosophical Studies* 133 (1): 83–94. https://doi.org/10.1007/s11098-006-9007-6.

Moore, G. E. (1903) 2004. *Principia Ethica*. Mineola: Dover Publications.

Mukerji, Nikil. 2016. *The Case Against Consequentialism Reconsidered*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-39249-3.

Mulgan, Tim. 2001. *The Demands of Consequentialism*. Oxford; New York: Clarendon Press; Oxford University Press.

———. 2007. *Understanding Utilitarianism*. Stocksfield: Acumen.

Nado, Jennifer Ellen. 2014. “Why Intuition?” *Philosophy and Phenomenological Research* 89 (1): 15–41. https://doi.org/10.1111/phpr.644.

Nersessian, Nancy J. 1992. “In the Theoretician’s Laboratory: Thought Experimenting as Mental Modeling.” *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1992: 291–301.

———. 2007. “Thought Experimenting as Mental Modeling.” *Croatian Journal of Philosophy* 7 (2): 125–61.

Norcross, Alastair. 1990. “Consequentialism and the Unforeseeable Future.” *Analysis* 50 (4): 253–56. https://doi.org/10.2307/3328263.

Nozick, Robert. (1974) 2012. *Anarchy, State, and Utopia*. Malden; Oxford: Blackwell Publishing.

Österberg, Jan. 1988. “A Problem for Consequentialism.” In *Not Without Cause: Philosophical Essays Dedicated to Paul Needham on the Occasion of His Fiftieth Birthday*, edited by Lars Lindahl, Jan Odelstad, and Rysiek Sliwinski, 278–83. Uppsala: Uppsala University, Department of Philosophy.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford; New York: Oxford University Press.

Peterson, Martin. 2013. *The Dimensions of Consequentialism: Ethics, Equality and Risk*. Cambridge: Cambridge University Press.

Portmore, Douglas W. 2014. *Commonsense Consequentialism: Wherein Morality Meets Rationality*. Oxford: Oxford University Press.

Pummer, Theron. 2012. “Intuitions about Large Number Cases.” *Analysis*, 1–9. https://doi.org/10.1093/analys/ans134.

Pust, Joel. 2017. “Intuition.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2019. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2019/entries/intuition.

Qizilbash, Mozaffar. 1999. “The Rejection of Objective Consequentialism: A Comment.” *Utilitas* 11 (1): 97–105. https://doi.org/10.1017/S0953820800002296.

Quinton, Anthony. (1973) 1989. *Utilitarian Ethics*. Peru: Open Court Publishing Company.

Rawls, John. (1971) 2005. *A Theory of Justice*. Cambridge: Belknap Press.

Rosen, Gideon. Unpublished manuscript. “Normative Necessity.”

Rosenqvist, Simon. 2019. “Review of Holly M. Smith’s Making Morality Work.” *Utilitas* 31 (3): 353–55. https://doi.org/10.1017/S095382081800033X.

Rydéhn, Henrik. 2019. *In Virtue of. Determination, Dependence, and Metaphysically Opaque Grounding*. Uppsala: Department of Philosophy.

Sandberg, Joakim, and Niklas Juth. 2010. “Ethics and Intuitions: A Reply to Singer.” *The Journal of Ethics* 15 (3): 209–26. https://doi.org/10.1007/s10892-010-9088-5.

Scanlon, Thomas M. 2000. *What We Owe to Each Other*. Cambridge: Belknap Press of Harvard University Press.

Scarre, Geoffrey. 1996. *Utilitarianism*. London; New York: Routledge.

Shaw, William H. 1999. *Contemporary Ethics: Taking Account of Utilitarianism*. Malden; Oxford: Blackwell Publishers.

Sidgwick, Henry. (1907) 1984. *The Methods of Ethics*. Indianapolis: Hackett Publishing Company.

Singer, Peter. 1993. *Practical Ethics*. Cambridge; New York: Cambridge University Press.

———. 2005. "Ethics and Intuitions." *The Journal of Ethics* 9 (3–4): 331–52. https://doi.org/10.1007/s10892-005-3508-y.

Slote, Michael, and Philip Pettit. 1984. "Satisficing Consequentialism." *Proceedings of the Aristotelian Society, Supplementary Volumes* 58: 139–76.

Smart, J. J. C. 1973. "An Outline of a System of Utilitarian Ethics." In *Utilitarianism For and Against*, 3–74. New York: Cambridge University Press.

Smith, Holly M. 1988. "Making Moral Decisions." *Noûs* 22 (1): 89–108. https://doi.org/10.2307/2215557.

———. 2012. "Using Moral Principles to Guide Decisions." *Philosophical Issues* 22 (1): 369–86. https://doi.org/10.1111/j.1533-6077.2012.00235.x.

———. 2018. *Making Morality Work*. Oxford: Oxford University Press.

Smuts, Aaron. 2011. "The Feels Good Theory of Pleasure." *Philosophical Studies* 155 (2): 241–65. https://doi.org/10.1007/s11098-010-9566-4.

Someren Greve, Rob van. 2014. "The Value of Practical Usefulness." *Philosophical Studies* 168 (1): 167–77. https://doi.org/10.1007/s11098-013-0124-8.

Sorensen, Roy A. 1992. *Thought Experiments*. New York: Oxford University Press.

Street, Sharon. 2006. "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies* 127 (1): 109–66. https://doi.org/10.1007/s11098-005-1726-6.

Tännsjö, Torbjörn. 1998. *Hedonistic Utilitarianism*. Edinburgh: Edinburgh University Press.

———. 2002. "Why We Ought to Accept the Repugnant Conclusion." *Utilitas* 14 (3): 339–359. https://doi.org/10.1017/S0953820800003642.

———. 2007. "Narrow Hedonism." *Journal of Happiness Studies* 8 (1): 79–98. https://doi.org/10.1007/s10902-006-9005-6.

———. 2010. *From Reasons to Norms*. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-90-481-3285-0.

Tersman, Folke. 1991. "Utilitarianism and the Idea of Reflective Equilibrium." *The Southern Journal of Philosophy* 29 (3): 395–406. https://doi.org/10.1111/j.2041-6962.1991.tb00599.x.

———. 1993. *Reflective Equilibrium: An Essay in Moral Epistemology*. Stockholm: Almqvist & Wiksell International.

———. 2008. "The Reliability of Moral Intuitions: A Challenge from Neuroscience." *Australasian Journal of Philosophy* 86 (3): 389–405. https://doi.org/10.1080/00048400802002010.

Thomson, Judith Jarvis. 1976. "Killing, Letting Die, and the Trolley Problem." *The Monist* 59 (2): 204–217. https://doi.org/10.5840/monist197659224.

Timmons, Mark. 2013. *Moral Theory: An Introduction*. 2nd ed. Lanham: Rowman & Littlefield Publishers.

Väyrynen, Pekka. 2006. "Ethical Theories and Moral Guidance." *Utilitas* 18 (3): 291–309. https://doi.org/10.1017/S0953820806002056.

Watson, Burton. 1963. *Mo Tzu: Basic Writings*. New York: Columbia University Press.

Weatherson, Brian. 2004. "Morality, Fiction, and Possibility." *Philosopher's Imprint* 4 (3): 1–27.

Wilkes, Kathleen V. 1988. *Real People: Personal Identity without Thought Experiments*. Oxford; New York: Oxford University Press.

Wilson, James. 2016. “Internal and External Validity in Thought Experiments.” *Proceedings of the Aristotelian Society* 116 (2): 127–52. https://doi.org/10.1093/arisoc/aow008.

Wright, Robert. 2000. *NonZero: The Logic of Human Destiny*. 1st ed. New York: Pantheon Books.

Index