Theory and Decision Library A: Rational Choice in Practical Philosophy and Philosophy of Science

Nikil Mukerji

The Case Against Consequentialism Reconsidered

Springer

Theory and Decision Library A:

Rational Choice in Practical Philosophy and Philosophy of Science

Volume 51

Series Editor
Julian Nida-Rümelin
Universität München, Munich, Berlin, Germany

This series deals with practical and social philosophy and also foundational issues in philosophy of science in general that rely on methods broadly based on rational choice. The emphasis in the Series A is on well-argued, thoroughly analytical and philosophical rather than advanced mathematical treatments that use methods from decision theory, game theory and social choice theory. Particular attention is paid to work in practical philosophy broadly conceived, the theory of rationality, issues in collective intentionality, and philosophy of science, especially interdisciplinary approaches to social sciences and economics. Assistant Editor: Martin Rechenauer (München) Editorial Board: Raymond Boudon (Paris), Mario Bunge (Montréal), Franz Dietrich (Paris & East Anglia), Stephan Hartmann (LMU Munich), Martin van Hees (Amsterdam), Isaac Levi (New York), Richard V. Mattessich (Vancouver), Bertrand Munier (Cachan), Olivier Roy (Bayreuth), Amartya K. Sen (Cambridge), Brian Skyrms (Irvine), Wolfgang Spohn (Konstanz), and Katie Steele (London School of Economics).

More information about this series at http://www.springer.com/series/6616

Nikil Mukerji

The Case Against Consequentialism Reconsidered

Springer

Nikil Mukerji
Faculty of Philosophy, Philosophy of Science,
and the Study of Religion
Ludwig-Maximilians-Universität München
Munich, Germany

ISSN 0921-3384 ISSN 2352-2119 (electronic)
Theory and Decision Library A:
ISBN 978-3-319-39248-6 ISBN 978-3-319-39249-3 (eBook)
DOI 10.1007/978-3-319-39249-3

Library of Congress Control Number: 2016940304

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

To my parents, Maria and Kiran Mukerji

Acknowledgements

A philosophical study is, above all, a *joint* effort. This one is no exception. While working on it, I benefited enormously from the support and encouragement of numerous fellow scholars. First and foremost, I owe thanks to Julian Nida-Rümelin and Martin Rechenauer, who supervised me in writing a thesis which ultimately gave birth to this book. Martin Rechenauer first suggested that I should write a piece about consequentialism. I profited a great deal from numerous discussions with him at every step of the way. Julian Nida-Rümelin's seminal book *Kritik des Konsequentialismus* (1993) and his other writings on consequentialism were a source of inspiration to me. Much of what I have to say on the following pages builds on his contributions. And I was very fortunate to be able to discuss my ideas with him regularly.

I would furthermore like to thank those with whom I was lucky enough to meet, interact, and cooperate, in particular, Kerim Peren Arin, Thomas Bonk, Matthew Braham, Christine Bratu, John Broome, Campbell Brown, Dale Dorsey, Julia Driver Driver, Grant Duncan, Andreas Edmüller, Gerhard Ernst, Wulf Gaertner, Jan Gertken, Johanna Grieshammer, Michael von Grundherr, Manfred Harth, Ludwig Heider, Dominik Heiss, Tim Henning, Lisa Herzog, Karl Homann, Robert Hümmer, Johanna Jauernig, Christoph Luetge, Erasmus Mayr, Paul McNamara, Ulrich Metschl, Julian Müller, Martin Peterson, Douglas Portmore, Hannes Rusch, Geoffrey Sayre-McCord, Walter Sinnott-Armstrong, Jörg Schroth, Christoph Schumacher, Andreas Suchanek, Steve Sverdlik, Mattias Uhl, and Martin VanHees. Also, I would like to thank Friedrich-Ebert-Stiftung (Bonn) for their financial support. Without them, I would not have been able to embark upon this project in the first place. Special thanks belong to Marianne Braun and Ursula Bitzegeio.

Last but not least, I owe thanks to my friends and family. In particular, I would like to thank Thomas Kaczmarek and Niko Kleinhammer. They have been the dearest of friends to me for so many years, and it is hard to overstate the debt I owe to them. Above all, however, I am indebted to my parents, Maria and Kiran Mukerji, who have always loved and supported me. For this reason, I dedicate this book to them.

Contents

List of Tables

Illustrations

Abbreviations

Abbr.	Meaning
BU	Bottom-Up Approach
CCND	Connection Between Constraints and Negative Duties
CGU	Coarse-Grained Utilitarianism
CU	Classic Utilitarianism
CUG	Classic Utilitarian Theory of the Good
DAC	The Distinction Between Act and Consequence
DAU	Desert-Adjusted Utilitarianism
DM	Definitional Method
FPA	Fixed Point Approach
FRA	Family Resemblance Approach
FRA_1	First Version of the Family Resemblance Approach
FRA_2	Second Version of the Family Resemblance Approach
HDD	Humpty Dumpty Defence
HSC	Hybrid Satisficing Consequentialism
HSC_1	Hybrid Satisficing Consequentialism – Version 1
HSC_2	Hybrid Satisficing Consequentialism – Version 2
NTW	Narrow Technical Welfarism
OSC	Objective-Subjective Consequentialism
OSC_1	Objective-Subjective Consequentialism – Version 1
OSC_2	Objective-Subjective Consequentialism – Version 2
PFPA	Provisional Fixed Point Approach
PU	Preference Utilitarianism
RE	Reflective Equilibrium Approach
SC	Subjective Consequentialism
SC_1	Subjective Consequentialism – Version 1
SC_2	Subjective Consequentialism – Version 2
SU	Satisficing Utilitarianism
SWF	Social Welfare Function
TD	Top-Down Approach
WTW	Wide Technical Welfarism

Introduction

When I first became interested in moral philosophy, one of the themes I immediately picked up on was the dispute between consequentialists and non-consequentialists. Their quarrel struck me as particularly conspicuous. Many authors would emphasize, right at the outset, that their approach to a given moral issue was a *consequentialist* one. Others would do the opposite, insisting that their view was decidedly *non-consequentialist*. Those who placed themselves in the consequentialist camp would usually begin with a few preliminary remarks in which they would motivate their position and answer to common criticisms of consequentialism, while those who thought of themselves as non-consequentialists would start out by explaining why taking a consequentialist stance on morality was a bad idea. To me, this suggested that the distinction between consequentialism and non-consequentialism was philosophically significant. It occurred to me that I had better make up my mind about it and pick the team I wanted to be on.

The paradigmatic consequentialist moral theory, viz., utilitarianism, seemed to me rather plausible and rational. Looking back, my first reaction to this doctrine was entirely predictable. As John Rawls writes, "[i]t is natural to think that rationality is maximising something and that in morals it must be maximising the good" (Rawls 1971/1999, 22). I thought that at least this aspect of utilitarianism, viz., its consequentialism, had to be true. Perhaps there were certain aspects of the doctrine that needed repair. It seemed clear to me, however, that the general approach was right and that the best moral theory was likely a close relative of utilitarianism.

But then I looked at critics of the consequentialist paradigm. They would argue that, in principle, consequentialism morally permits theft, torture, and murder as well as many other nasty things. Their arguments looked sound and crystal clear. And they persuaded me that consequentialism was not for me after all. I was quite convinced that this conviction would not waver. However, when I considered the replies of consequentialist authors, they would draw a rather different picture. They maintained that, on a more charitable reading, consequentialism is, in fact, capable of addressing the worries that critics have put forward. Much of what I read convinced me and I began to reconsider my earlier dismissal of consequentialism.

When I then looked at the replies of non-consequentialists, however, they, too, sounded rather persuasive. This baffled me. Some commentators have pointed out that one of the problems about consequentialism is that "[s]ome people find it very plausible", while "others find it very implausible" (Bergström 1996, 3). In my experience, however, the problem was different. It was very hard for me to make up my mind about it. I went back and forth between finding consequentialism plausible and finding it implausible. Eventually, this motivated me to dig deeper.

Before long, I came up with a theory that explained my confusion. I observed that critics of the consequentialist view of ethics commonly proceed as follows. First, they *define* – explicitly or implicitly – what a consequentialist theory is. That is, they assume that all and only consequentialist moral theories share a given property. Then, they go on to argue that this property gives rise to some fatal problems. And they conclude from this that we should, therefore, reject consequentialism. Further below, we will refer to this method of critique as the Definitional Method (DM). It turned out that consequentialists can easily rebut criticisms that their critics formulate based on DM. To this end, all they have to do is to claim that the notion of consequentialism on which opponents premise their objections does not coincide with their favoured view of consequentialism and that they, hence, do not apply to their moral theory. Further below, we will call this strategy the Humpty Dumpty Defence (HDD) – in mock *homage* to an argument that Humpty Dumpty uses in a discussion with Alice in Lewis Carroll's book *Through the Looking-Glass* (1871/1990).

I realized that the possibility of HDD made the debate about consequentialism so complicated because it impedes substantive discussion. Much of what I read reminded me of a fictitious dialogue between a sceptic and an antisceptic that I had once come across in Jerry Fodor's book *Concepts (1998)*. The dialogue goes like this:

Sceptic:	You can't ever infer with certainty from how things look to how they are.
Antisceptic:	Can too, because there is an intrinsic conceptual connection between how-things-look concepts and how-things-are concepts (. . .). Bop. I win.
Sceptic:	I don't acknowledge such intrinsic connections.
Antisceptic:	Then you don't have the concepts! Bop. I still win. (Fodor 1998, 70)

A stylized dialogue between a consequentialist and a non-consequentialist would run as follows:

Non-consequentialist:	All consequentialist theories are subject to objection O. We should, hence, reject them. This follows from the definition of consequentialism. Bob. I win.
Consequentialist:	I don't acknowledge your definition.
Non-consequentialist:	Non-consequentialist: Then you don't understand what consequentialism means! Bop. I still win.
Consequentialist:	Do too!
Non-consequentialist:	No, you don't!
Consequentialist:	Yeah, I do . . .

This depiction is certainly an exaggeration. The debate between consequentialists and non-consequentialists has surely amounted to more than *that*. Nevertheless, I believed, as I still do, that the presence of the verbal dispute poses a severe problem. It makes it tough to tell who is right. After all, the question whether consequentialism is subject to any particular objection ultimately comes down to one issue: how should we understand the notion of consequentialism? And this, it seems, cannot be settled by argument when HDD is possible. I thought, therefore, that I should look for ways to address this problem.

My first hunch was this. Maybe, I reasoned, it is possible to define the concept more broadly so as to make it impossible for consequentialists to apply HDD. But this approach turned out to be entirely on the wrong track.[1] To be sure, it is possible to state an idea which is certainly necessary for consequentialism (though it may not be sufficient). Below, we will call it the Core Idea of consequentialism. For critics who want to show that consequentialism is a defective moral outlook, this idea seems to be as good as a definition. If, after all, they succeed in showing that all doctrines which accord with it are morally objectionable, they have demonstrated (*a fortiori*) that all consequentialist theories are morally unacceptable. I realized, however, that there was a problem with this approach. As we will discuss below, the Core Idea is very abstract. That, in turn, makes it difficult to level any substantive criticisms against it. Hence, it became apparent to me that critics of consequentialism seemed to face a dilemma if they followed DM. If they start with a rather narrow account of consequentialism, this will allow them to formulate substantive criticisms. However, the narrowness of the definition will also give consequentialists the chance to employ HDD. If, on the other hand, critics start with a sufficiently broad definition of consequentialism that captures all consequentialist moral theories, substantive criticisms do not apply. This dilemma led me to an obvious question: Given that it seems impossible to use DM to show that consequentialism is untenable, is there another method we could use to this end?

At first, it was hard for me to see how there could be such a method. It was hard, that is, to see how one could conceivably support a general claim about an object of inquiry – in my case, consequentialism – if one did not start the discussion with a rigorous *definition* of it. I have come to believe, however, that there is such a method. The reason it took me a while to figure it out was that I naïvely bought into a certain view of language. On this view, there are only two kinds of general

[1] In retrospect, this should not have come as a surprise. Modern analytic philosophy teaches us that oftentimes rigorous definitions are not within reach. A case in point is the debate about the concept of knowledge in epistemology. In his *Theaetetus* Plato's protagonist Socrates defines knowledge as justified, true belief. This idea was famously demolished by Gettier (1963) who showed that there are cases of justified, true belief that are not, intuitively, cases of knowledge. Subsequently, philosophers embarked upon a quest to find an augmented definition that would stand up in cases of the sort that Gettier had proposed (for an instructive survey of this research, see Ichikawa and Steup 2012). This undertaking, however, might have been hopeless from the start because the idea of knowledge may be unanalysable, as an increasing number of epistemologists are claiming (e.g. Williamson 2000, 2–5). I am grateful to Martin Rechenauer for suggesting this example to me.

terms, viz., "basic terms" and "composite terms". The former are used to pick out observable characteristics, while the latter refer to more complex objects and are defined in terms of the former. Since I accepted this bifurcation, I had to put the term "consequentialism" into one of these two categories. I believed that it was certainly not a basic term. Hence, I reasoned, it must be a composite term which is defined by basic terms. It was hard to see, therefore, how one could conceivably avoid the need to define the term "consequentialism".

Then, however, I came across some of Wittgenstein's ideas and the scales fell from my eyes. Wittgenstein taught me an alternative view of general terms, according to which there is a third category, viz., that of "family resemblance terms". Such terms refer to a class of objects which is not united by a single shared feature. To be sure, every object in the class shares *some* feature with another. However, these common features may vary throughout pairs. This idea helped me to make sense of what some authors had noted in passing, viz., that we should understand the term "consequentialism" as a family resemblance term. Initially, I believed that this contradicted what I had established earlier. As I said above, there appears to be a Core Idea that lies behind all forms of consequentialism. For reasons that we will discuss in due course, however, this impression turned out to be incorrect. The fact that there is a Core Idea behind all forms of consequentialism does not imply that we cannot interpret the term "consequentialism" as a family resemblance term. When I had realized that this was so, I started to investigate whether the family resemblance view of consequentialism had any interesting methodological implications. It turned out that, indeed, it had. It became apparent that a methodological approach based on the idea of family resemblance – the Family Resemblance Approach (FRA), as I would later christen it – was capable of transcending the dilemma that plagued DM. This, of course, is not to say that FRA was free from problems. In fact, my further investigation made it clear that I would have to elaborate it to address problems that afflicted it. Nevertheless, I believed that the obvious next step was to apply the method because, as I like to say, "here, as anywhere in philosophy, the proof of the pudding is in the eating!" (Mukerji 2013b, 664). So I did. I elaborated FRA and used it to reconsider the case against consequentialism. That is, I used the method to put together a comprehensive argument against consequentialism which shows that, in fact, we do seem to have good reason to reject every consequentialist moral view. At first, I did so merely to play *advocatus diaboli*. By now, however, I have come to believe that my argument is sound and that consequentialism is, in fact, an untenable position. The reasoning that convinced me is novel in one respect, viz., in that consequentialists cannot dodge it the way they did previous objections. They cannot counter it, that is, simply by saying that they do not share the definition of consequentialism on which the argument is based. The argument is not, after all, based on any definition. So, HDD can do no lifting!

These preliminary remarks should suffice, I believe, to clarify the theme of this study. Moreover, they should make it obvious why anyone with an interest in the theoretical foundations of normative ethics would care about it. Before we start, however, let me sketch a rough picture of the reasoning. In Chap. 1 that follows this introduction, we will begin with a few preliminary remarks on normative ethics which is the subject area of our inquiry. In Sect. 1.1, we will distinguish this area of

moral philosophy from the other fields of inquiry in that area, viz., meta-ethics and applied ethics. As I understand it, normative ethics is the study of moral theories, and the claim that we are aiming to support in this study is a claim about a particular family of moral theories, viz., the family of consequentialist moral theories. In Sect. 1.2, we will anatomize the idea of a moral theory. We will elaborate on some important ideas that this notion comprises as well as crucial concepts to which it is connected. We will, in particular, discuss the moral concepts – right, wrong, obligatory, optional, and supererogatory – as well as the notions of an act, a choice situation, and a moral agent. This is important because we will frequently use these concepts in the subsequent chapters. Most importantly, though, we will factorize the idea of a moral theory into a theoretical component and a practical component. We will draw on this distinction when we dissect consequentialist moral theories in Chap. 4.

Having clarified the idea of a moral theory, we will turn to the meta-ethical foundations of our study in Chap. 2. Specifically, we will move on to the issue of moral theory evaluation. This is obviously a necessary step. After all, our aim is to put together a convincing case against consequentialism. To do that, we need to work out, more generally, how we should evaluate moral theories (or families thereof). In Sect. 2.1, we shall start by introducing an influential approach to theory evaluation in ethics. John Rawls proposed it. Hence, we will call it the Rawlsian Approach. It gives us a theoretical desideratum that can be used to judge the adequacy of moral theories. The idea is, very roughly, that a moral theory is acceptable to the extent that it systematizes and corrects our pre-theoretical moral intuitions. We can factorize this notion into (at least) three subsidiary criteria, viz., intuitive fit, systematicity (or connectedness), and consistency. In our study, we will focus on the first one because intuitive objections have occupied centre stage in the debate about consequentialism and seem, therefore, most promising. In Sect. 2.2, we will, then, discuss various interpretations of the Rawlsian Approach that relate to the sub-criterion of intuitive fit, viz., the Top-Down Approach (TD), the Reflective-Equilibrium Approach (RE), and the Bottom-Up Approach (BU). We will delineate these approaches in terms of a distinction between high-level and low-level intuitions. The former are abstract intuitions that cover all possible cases or, at least, a very wide range of cases. The latter are concrete intuitions that involve a single case or a small variety of cases. TD is the view that a moral theory is adequate, *ceteris paribus*, to the extent that it fits our high-level intuitions. RE is the view that a moral theory is adequate, *ceteris paribus*, to the extent that it has an overall fit with our intuitions at both ends of the high/low spectrum. And BU is the view that a moral theory is adequate, again *ceteris paribus*, to the extent that it fits our low-level intuitions about cases. We will try to establish that there are good reasons to reject TD and that either RE or BU is adequate. This is a crucial step in our argument. As will become apparent shortly, our reasoning depends critically on the assumption that a theory's fit with our low-level intuitions plays a role in its justification. However, if TD were the best interpretation of intuitive fit, this assumption would be false. Only RE and BU allow low-level intuitions to play a role. Hence, we have to show that TD is mistaken and that either RE or BU is the correct view. By itself, however, this result is not terribly helpful. This is because neither RE nor BU can give us any practical guidance in evaluating a moral theory. These approaches do not provide

a "pass-or-fail test" for moral theories. They merely define a philosophical ideal, viz., that a theory should fit the relevant intuitions. Therefore, we have to work out, in Sect. 2.3, how we can translate these approaches into a workable procedure that we can apply in our evaluation of consequentialist moral theories. We will call this method the Provisional Fixed Point Approach (PFPA). The idea behind PFPA is as follows. We look for "provisional fixed points" in our moral thinking. That is, we look for intuitive moral convictions which are unyielding, such that it is reasonable to expect that any acceptable moral theory should match them. Then, we test the moral theories in question against them. We check whether they do, in fact, match them and we reject them if they do not. In conjunction with the RE/BU approach to intuitive fit, PFPA gives us a method for our study that allows cases to play a crucial role. This method does not specify, however, which kinds of cases we should use. There are, in principle, two options. We can use realistic scenarios that have happened in real life or are, at least, likely actually to happen. Alternatively, we can use hypothetical examples which are very unlikely or outright impossible actually to occur. In Sect. 2.4, we will motivate and justify the use of trolley cases. These are a subclass of the latter stripe. We will talk about their distinct characteristics and their uses in moral inquiry. Then, we will explain why they are especially useful in the context of our study of consequentialism before we finally address possible objections against them.

Chapters 1 and 2 mostly reproduce conventional wisdom. This is not to say, of course, that what we say here is uncontroversial. But our aim in these chapters is not novelty. It is merely explicitness. Many ethical studies, I think, suffer from a lack thereof. Their arguments rest on premises that are, at least partly, unexplained. This holds, in particular, for the meta-ethical premises. It is often unclear, that is, which background assumptions are used to derive conclusions. Our aim is to avoid this, such that, if we are wrong, we are "at least wrong clearly" (Moore 2008, 38). Chapter 3 is different.[2] Here, we explore new territory. Equipped with a general method for testing moral theories, we turn to the pivotal element of our study. It is the question how we should make a case against consequentialism *methodologically*. In Sect. 3.1, we will consider the conventional approach, viz., DM. We will examine its problems and, in particular, HDD. After that, we will develop FRA in Sect. 3.2. Our first formulation of it, FRA_1, will turn out to be defective because it builds on a questionable assumption. Therefore, we will have to introduce a second formulation of the approach, FRA_2, which remedies this fault.

FRA_2 contains individual steps along which we will organize the remainder of our study in Chaps. 4 and 5. At this point, it is not possible to explain the content of these steps in much detail. The general idea, however, can be stated as follows. In Sect. 4.1, we will start by examining a paradigmatic consequentialist moral theory, viz., classic utilitarianism (CU). That is, we will factorize it into a set of logically distinct components $C_{11}, \ldots, C_{n1}$. Then, we will use an implication of our premise that the term "consequentialism" is a family resemblance term.

[2]Parts of Chap. 3 reproduce material from Mukerji (2013b).

If this assumption is correct, then there has to be at least one alternative, C_{i2}, to each paradigmatic component, C_{i1}, of CU that consequentialists can endorse. In Sect. 4.2, we will examine which relevant options we have to consider. In this connection, we will discuss, e.g., the distinction between subjective and objective forms of consequentialism, direct and indirect consequentialism, maximizing and non-maximizing theories, aggregative and non-aggregative, agent-neutral and agent-relative theories, and so on. These variants of consequentialism result from particular combinations of paradigmatic and/or non-paradigmatic components. One problem that we face is that there are uncountably many alternatives to each of the CU-components. To keep the study within manageable bounds, we will have to make certain assumptions as to which components are worthy of our attention.

In Chap. 5, we will take on the last step of our methodic procedure, FRA_2. Based on a series of trolley cases, we will attempt to show that every member of the family of consequentialist moral theories is subject to some serious objection. This chapter merely connects the dots and is accordingly short. Anybody who reads it in isolation may, in fact, get the impression that it is clownishly short. Moreover, they may say that it unjustifiably ignores many variants of consequentialism. It is important, therefore, to emphasize that the chapter has to be read against the background of our previous methodic steps. It may be an exaggeration to claim that everything we discuss is "equally close to the centre", as Adorno (2006, 71) thought it should be.[3] However, there are crucial bits and pieces of the overall argument in all chapters and sections of the text.

In Chap. 6, then, we will sum up and review our reasoning. Intellectual honesty commands that we reexamine the various controversial assumptions on which our thesis rests. Hence, we will discuss, by way of conclusion, how we might be wrong, which premises seem bold, and how we might strengthen our case against consequentialism further.

As a final preliminary matter, allow me a short comment on my writing style. I believe that in writing a philosophical text (or any text, for that matter), there are two important goals: *concision* and *readability*. However, these two goals are often in conflict since readability often requires some level of redundancy. No reader can remember all that she has read. Thus, it is often in the interest of readability to reprize certain aspects of earlier chapters and sections. This, of course, makes the text redundant. Since I tried to achieve a healthy balance between the two goals, the subsequent text is certainly not the most concise. At times, we will recapitulate ideas that we have discussed earlier. Furthermore, Chaps. 1, 2, 3, 4, and 5 contain summaries in which we will resume their central insights. I tried, however, to avoid lengthier rundowns of matters that are settled in previous parts of the text. When drawing on more complex ideas that we discussed earlier, I frequently include cross-references to the page or section where they first appear. This, I hope, will be helpful to those who want to turn back and reread the respective passages.

Without further ado, then, let us reconsider the case against consequentialism.

[3] I would like to thank Martin Rechenauer for making me aware of Adorno's quote.

Chapter 1
Normative-Ethical Foundations

Moral philosophers usually partition their subject into three fields: *normative ethics*, *applied ethics*, and *metaethics*. Though they may draw the lines between these areas in various ways, it is not uncommon to define normative ethics as the systematic study and evaluation of moral theories and applied ethics as the investigation of specific moral issues and concrete cases based on specific moral theories.

On this view, the contrast between normative ethics and applied ethics is twofold. We can distinguish them in the light of their *level of abstraction*, on the one hand, and their *focus on applicability*, on the other. Nevertheless, they have something in common which sets them off from the realm of metaethics. Both normative ethics and applied ethics are, as it were, on the same plain. They are both in the business of moral evaluation. Metaethics is concerned with different issues. It goes up a level, as it were, and considers moral discourse from a meta-perspective. In particular, it is concerned with the *ontological*, *semantic*, and *epistemological* questions that arise in the context of substantive moral evaluation.

We can characterize our study as belonging primarily to the field of normative ethics. We will home in on a particular class of moral theories, viz. "consequentialism". And we will try to support a general claim about it, viz. that we should reject every moral theory in it. As it will turn out below, however, parts of our discussion will also link up with certain issues in metaethics and moral epistemology, in particular.

In what follows, we will discuss some important ideas and concepts that we will frequently draw on in our inquiry. To start, we will briefly discuss normative ethics in Sect. 1.1. In Sect. 1.2, we will move on to the objects it studies, viz. moral theories and its components. In Sect. 1.3, then, we will close the chapter with a summary of our most important points.

N. Mukerji, *The Case Against Consequentialism Reconsidered*,
Theory and Decision Library A: 51, DOI 10.1007/978-3-319-39249-3_1

1.1 Normative Ethics

To characterize the object of our inquiry further, we should ask what the normative-ethical debate is all about. In fact, it is hard to give an all-encompassing answer to this question. Moral theories, which are the objects of normative-ethical inquiry, deal with a host of things. We can, e.g., morally assess a person's *feelings*, her *dispositions to act*, her *attitudes*, the *assertions*, *and judgements* she makes as well as the *convictions* and *motives* she has. We can assess *practical rules* that people follow as well as *social institutions* within which they act and interact. And we can, of course, morally evaluate the way people *act*.

One way to classify moral theories is in view of their *primary evaluative focal point* (Kagan 1992, 239), that is, regarding the object that they take to be of central concern.[1] Throughout the history of moral philosophy, many authors have, e.g., been *virtue theorists*.[2] These philosophers focus on character traits and are particularly interested in the qualities a person should aspire to possess. Others have been *rule theorists*. They seek to find out which sorts of rules people should follow.[3] Yet others have been *motive theorists*. Of course, there have also been those who primarily focused on the evaluation of acts. They are *theorists of right action*.

Consequentialist moral systems, in particular, can have many different objects. As Derek Parfit remarks: "Consequentialism covers, not just acts and outcomes, but also desires, dispositions, beliefs, emotions, the colour of our eyes, the climate, and everything else."[4] (Parfit 1986, 25) However, in this inquiry, we shall not look at such a "global consequentialism" (Pettit and Smith 2000) or "multiple object consequentialism," (Mulgan 2001b, 40) as some have called them. Rather, we shall look solely at what moral philosophers commonly call *act*-consequentialism. However, we will henceforth drop the qualifier 'act-' and refer to it simply as consequentialism.[5] Accordingly, we shall use the term 'moral theory' in a narrower sense than usual. We mean by it only those theories which take acts as their primary evaluative focal point. Having said this, let us define the notion of a moral theory a bit more formally.

[1] For this point, see also Crisp (2006, 101).

[2] Plato and Aristotle count as the founding fathers of this approach to ethics. For a survey article on virtue ethics, see Hursthouse (2012).

[3] Immanuel Kant is perhaps the paradigmatic exponent of a rule-based approach to ethics (cf., e.g., Kant 1785). For a helpful and comprehensive commentary on his moral system in his most widely read book *Groundwork of the Metaphysics of Morals* (*Grundlegung zur Metaphysik der Sitten*) see, e.g., Timmermann (2007).

[4] See, also, Parfit (2011, 374).

[5] Our reason for this limitation is simply to keep the discussion within reasonable bounds. Note, however, that doubts may be raised as to whether we should, in fact, regard a moral theory as consequentialist if it does not adopt acts as their primary evaluative focal points. For this point, see, e.g., the exchange between Howard-Snyder (1993) and Hooker (1994). See, also, Hooker (2003, 108–111).

1.2 Moral Theories

As Sumner (1987, 180) points out, we can distinguish between two purposes of moral theories. "A complete moral framework," he says, "must include both a theory of justification and a theory of decision-making."[6] In other words, its first purpose is to state general principles for the moral evaluation of acts. Its second objective is to give us practical instructions for making choices. We may, therefore, factorize moral theories into two parts, viz. a *theoretical component* which corresponds to the first purpose and a *practical component* which corresponds to the second. In what follows, we shall consider each of them in turn.

1.2.1 The Theoretical Component

The theoretical component of a moral theory deals with moral principles. To understand what it is all about, we need to comprehend, therefore, what a moral principle is. Generally speaking, such principles connect the non-moral properties of an act with its moral properties. To illustrate, consider the principle that lying is wrong. It states that an act which is a lie (non-moral property) is wrong (moral property).

We can distinguish between two types of principles: *pro tanto* principles[7] and *all-things-considered* principles. The verdict that lying is wrong, e.g., can be understood in two ways. The first is to interpret it as saying that an act which is a lie is wrong, *period*. This is the all-things-considered interpretation of the principle. Another way to interpret it is to say that an act is wrong *insofar as* it is a lie. This is the *pro tanto* interpretation. It does not say that an act which is a lie is wrong, period. It merely says that this fact is a "wrong-making feature" (Timmons 2002, 205) or, more generally, a "normative factor" (Kagan 1998, 17) which counts against the act.

Based on this distinction between *pro tanto* and *all-things-considered* moral principles, we can make another distinction between two kinds of moral theories, viz. *monistic* theories and *pluralistic* theories. Adherents of monistic theories claim that a moral theory ought to possess a unique method for weighing up right-making and wrong-making features. They claim, in other words, that we can condense *pro tanto* principles into a supreme *all-things-considered* principle. This supreme standard,

[6]For the distinction between principles for right action and principles for decision-making, see also Bales (1971), Brink (1989, 256), and Pettit (1984).

[7]It is more common to use the expression "prima facie" principle, as proposed by Ross (1930/2002). Kagan (1989, 17), however, has pointed out, rather convincingly, that this expression seems to connote that the principle merely appears to be genuine. A *pro tanto* principle, however, is a genuine principle which may – and that is the important point – be outweighed by other principles.

which takes account of all normatively relevant factors, would tell us in an all-encompassing way which acts possess a particular moral property. Insofar we might call it, as it commonly is called, a *criterion*. Depending on the moral property that it adjudicates on, we would call it a *criterion of rightness*, a *criterion of wrongness,* and so on for the other moral properties.

Pluralists deny that there is a unique method for weighing up the various *pro tanto* principles. Moreover, they deny the possibility that there can be one single criterion for determining whether an act possesses a particular moral property, e.g. rightness.

Note, then, that the disagreement between monistic and pluralistic moral theorists concerns not only the question which moral theory is the most appropriate one. The two camps disagree more fundamentally about the edifice of moral systems. On the pluralist picture, a moral theory is simply a set of *pro tanto* moral principles that are not necessarily connected with one another by a weighing method. In contrast, monists believe that a moral theory must give us a *unified* or *systematic* theoretical account of morality. There may be reasons to endorse either view.[8] But we do not need to adjudicate this debate here. Bearing in mind that we are interested in putting together a case against consequentialism, we can just use the picture of a moral theory that is more congenial to consequentialism. This picture is monistic.

I believe that we have said enough now to formulate a more formal account of a moral theory which should serve our present purpose. On the monistic view, the theoretical component of a moral theory can be characterized as follows.

The Theoretical Component
The theoretical component of a moral theory is a device that attaches a particular *moral status* to the *acts* available in a given *choice situation* to a *moral agent* as a function of certain *normative factors*.[9]

This characterization is, I believe, quite useful. For it concisely wraps up and highlights the various notions that are involved in the idea of a moral theory. We shall address them one by one in what follows.

[8]To support the monistic standpoint, it might be said, e.g., that a pluralistic account of morality seems somewhat arbitrary. It merely points to an "unconnected heap of duties" (McNaughton 1996) or a "patchwork quilt" (Kagan 1998, 295) of morally relevant factors, as we might metaphorically put it. That is, it does not explain why these factors *matter*. In contrast, a monistic account has some ultimate standard – an ultimate criterion of rightness and wrongness – which can explain why acts possess the moral status that they do. Pluralists can defend themselves against this charge by pointing out that monistic theories usually have a lower degree of fit with the common-sense moral judgements that we find intuitively plausible (cf., e.g., Ross 1930/2002, 24; Kappel 2006, 132). This fact, they might maintain, speaks against monism and in favour of their pluralist view.

[9]Our characterization of the theoretical component of a moral theory follows Peterson (2010, 156). It is also similar to what Brown (2011, 759) calls a "rightness function".

1.2.1.1 The Moral Status of Acts

The moral status of an act can be described using the moral concepts. The most frequently employed ones are *right* (or *permissible*), *wrong*, *obligatory*, *optional*, and *supererogatory*.[10]

In the modern moral-philosophical debate, it is usually assumed that moral status is an all-or-nothing affair. We take it for granted that an act is either right or not right, that it is either wrong or not wrong, and so on. This, it seems, has become a standard assumption. We might as well follow convention, then, and adopt this view.

Various scholars have pointed out that we may interpret the moral concepts in different ways (cf., e.g., Parfit 2011, 150). We can, e.g., distinguish a fact-relative sense of 'right', 'wrong', and so on from a *belief-relative* sense of these expressions. We will come back to this distinction on page 115. But, for now, we shall ignore it and turn our attention to a logical property that these moral concepts possess (independently of whether we interpret them, e.g., in the fact-relative or belief-relative sense). It is possible to define each one of them in terms of the others, using a set of rules contained in standard deontic logic (cf. McNamara 2010). We can, e.g., take 'right' as basic and define the other expressions based on it. Given that we understand what 'right' means, we can read 'wrong' as follows. We call an act 'wrong' if and only if it is not right. We call an act 'obligatory' if and only if it is wrong not to do it. And we call an act 'optional' if and only if it is right to do it and right not to do it. As Urmson (1958) has pointed out, the only exception is the concept of supererogation. Roughly, a supererogatory act is a permissible act which is good, but not mandatory, as it involves a level of self-sacrifice which seems unreasonable to require. This notion, of course, cannot be defined in terms of the other concepts. I would like to bracket it for now until we get back to it further below on page 150. The reason is this. If we ignore supererogation and assume standard deontic logic, we can represent the theoretical component of a moral theory (as we have just defined it) in a compact way. We can understand it as a proposition

[10]The use of these moral properties has been subject to various criticisms. Some philosophers have argued that "thin moral concepts" such as "morally right" and "morally good" should not be used by moral theorists. What G.E.M. Anscombe says in her famous article "Modern Moral Philosophy" (1958) can be interpreted in that way, as Driver (2009) points out. More recently, Roger Crisp has expressed a similar view (cf. Crisp 2006, 20–27). A recent rebuttal comes, e.g., from McElwee (2010). Other philosophers have suggested that moral properties such as rightness and wrongness are gradual rather than discrete. Bentham famously said that his principle of utility "approves or disapproves of every action whatsoever, according to the *tendency* which it appears to have to augment or diminish (. . .) happiness (. . .)." (Bentham 1838, 1; emphasis added, NM) John Stuart Mill has asserted that "(. . .) actions are right in *proportion* as they tend to promote happiness (. . .)." (Mill 1863, 9; emphasis added, NM) Both statements make it clear that these authors did not take rightness and wrongness as all-or-nothing affairs, as it is most commonly done nowadays. Alastair Norcross has supported this sort of "scalar morality" in recent times (cf. Norcross 2006). Martin Peterson has proposed a view called "Multi-Dimensional Consequentialism" which holds that "[m]oral rightness and wrongness are non-binary concepts, meaning that moral rightness and wrongness vary in degrees." (Peterson 2012, 186) Rebuttals to such views can be found in McElwee (2010) as well as in Lawlor (2009a, 74–80, b) and Lang (2013).

which gives us necessary and sufficient conditions for a given moral concept (or its respective negation) to apply to an act. This is what we called a *criterion*. Depending on the particular moral property at issue we can distinguish between a *criterion of rightness*, a *criterion of wrongness*, a *criterion of obligatoriness*, and a *criterion of optionality*. Throughout, we shall assume that we can represent the theoretical component of a moral theory as either one of these criteria *plus* the semantic rules of standard deontic logic. We will, therefore, confine our attention, for the most part, to the criterion of rightness.[11]

1.2.1.2 Acts

To complete our picture of the theoretical component of moral theories, we need to address the other notions that we have used to characterize it. These are the concepts of an *act*,[12] an *agent*, a *choice situation*, and a *normative factor*. We shall start with the idea of an act.

It may appear as though the clarification of the concept of an act is hugely important for any ethical inquiry. After all, "[i]f the central problem of ethics is that of right and wrong action", it seems we first need to ask the question: "what is

[11] It should be noted that standard deontic logic has been criticized for its lack of neutrality between moral theories. E.g., as Sayre-McCord (1986) points out, it has the unfortunate consequence that the possibility of moral dilemmas is *ab ovo* excluded. Most moral philosophers believe, however, that the thesis of the impossibility of moral dilemmas, whether they believe in it or not, ought to be seen as a substantive claim and should, hence, be the result of substantive moral inquiry rather than an implicit assumption from which it starts. One may worry, then, whether standard deontic logic is, in fact, a good starting point for our inquiry. In reply, I guess, the following should be said. Firstly, since this is a study of the case against consequentialism, we should make assumptions which consequentialists are likely to accept. They commonly reject moral dilemmas. Secondly, it is, as far as I can see, not even the case that our definitional stipulations exclude moral dilemmas. We assume only a part of the standard deontic logic, viz. the "Traditional Definitional Scheme" (McNamara 2010). This scheme does not, in and of itself, exclude moral dilemmas. The impossibility of moral dilemmas arises only if this definitional scheme is combined with further assumptions, e.g. those contained the "Threefold Classification of Deontic Logic" (McNamara 2010) which states that an act is either obligatory, optional, or wrong, and that no act can fall into more than one of these classes. We do not assume this. Thirdly, and most importantly, we have to stipulate how the moral concepts are to be interpreted. We can, of course, do this one way or the other. But we might as well do it in accordance with standard deontic logic since, as McNamara (2010) says, it is still the most cited and studied version of deontic logic and it seems reasonable to suppose, therefore, that it is best suited to model contemporary moral discourse on which, after all, we focus.

[12] Some theorists distinguish between the notion of an *act* and that of an *action*. The distinction is this: What an agent does is called an "act." An act that is done to attain some end is called an "action." As Christine Korsgaard explains, "making a false promise and committing suicide are what I am calling 'acts' (. . .) Making a false promise *in order to* get some ready cash, committing suicide in order to avoid the personal troubles that you see ahead, and committing suicide *in order to* avoid harming others are what I am calling 'actions'." (Korsgaard 2008, 219; emphasis added, NM) We, however, shall not make this distinction.

an action anyway?" (Wellman 1972, 86) If that were true, it would seem as though any contribution to ethics should, as a preliminary exercise, take stock of the latest developments in action theory, as these developments should "have inescapable implications for ethics." (Wellman 1972, 86) It seems, however, that this is not how ethicists have standardly proceeded. Questions concerning the nature of agency, if they are discussed at all, are not discussed in much detail. There are, I believe, two reasons for this.

For one thing, the debate in philosophical action theory is simply too complex. There are innumerable positions as regards human agency. And, worse yet, there is hardly any common ground between them. E.g., there is not even agreement on what kind of a *thing* an act is or whether acts even belong to any unified category of things. In an ethical study, such as this one, it is therefore not possible to take account of (let alone adjudicate upon) the various views.

The second reason is that questions which arise in the philosophy of action do not seem to have much bearing on ethical matters (cf. Rawls 1974–1975, 5–6). To see this, consider the issue of what kinds of entities actions are. There are different accounts. It is commonly supposed, e.g., that actions are a subclass of events, viz. the class of events which, in some way or other, are related to an *agent*. However, there are different accounts of what an event is. Davidson (1963), e.g. takes events to be ontologically fundamental and holds that we cannot, therefore, analyse them into ontological building blocks. Kim (1976), on the other hand, believes that events are property exemplifications of objects and that we can, therefore, analyse them in terms of the object, the property exemplified, and the time at which the event occurs. Georg Henrik von Wright and others (e.g. Bach 1980, Chisholm 1964, Thomann 2010, 146) have refused to regard actions as events. They have proposed to view them instead as relations, viz. "as the bringing about or effecting ('at will') of a change." (von Wright 1963, 36) Throughout the years, much ink has been spilt in discussing the *pros* and *cons* of views like these. This has led to more and more complexity. Whether it brought us closer to an understanding of what actions are is something we cannot assess here. And, as is my point, we do not need to do it either. As Anthony Appiah says, "fortunately, in philosophy, you can sometimes get the conclusion you want without settling a disputed question." (Appiah 2008, 26) For practical purposes, it seems, an intuitive notion of what an act is and of what it means to act is sufficient. Even a small child can understand someone who says "You should not have *done* that!" or "This is something I don't want you to *do*." If she pushes the vase off the table and it breaks, she can understand that the breaking of the vase is something she *did*. This intuitive sense of an act is, as I said, for the most part entirely sufficient for our work in ethics. Hence, I see no reason we should expand on the notion of an act in general. For the most part, it seems, we can use an intuitive idea of an act. There are merely some minor qualifications which seem to be necessary before we can proceed.

The first point to be made concerns the oft-neglected question of whether *not* doing an act – a "negative act" (Vermazen 1985, 93), as it were – is itself an act. Consider, again, the case of the kid breaking the vase. Is this an instance of an act? Yes, it is an act, viz. the act of breaking the vase. Now, if the kid does *not*

break the vase, is this an act, too? In this case, I take it, we would not ordinarily say that this is also an act, viz. the act of *not* breaking the vase. Negative acts are not commonly considered to be acts. However, they should be regarded as acts in the sense in which we employ the term 'act' in the above definition of a moral theory. The question whether this way of speaking is ultimately appropriate shall not concern us here, and we need not discuss the reasons for and against regarding negative acts as acts.[13] It is, however, important to emphasize the motivation for viewing negative acts as acts. There are two reasons why this seems necessary. The first is logical. Above we introduced the terms 'right', 'wrong', 'obligatory', and 'optional'. Moreover, we accepted standard deontic logic which claims that these expressions are definable in terms of one another. This already presupposes that we regard negative acts as acts. On standard deontic logic, recall, an act can be defined as 'obligatory' if and only if it is wrong *not* to do it. The second reason is substantive. If we excluded negative acts from the class of acts, moral theories could not judge morally significant instances of inaction. To see why this would be a problem consider, for the sake of illustration, a famous thought experiment by Peter Singer. Singer says that

> (. . .) if I am walking past a shallow pond and see a child drowning in it, I ought to wade in and pull the child out. This will mean getting my clothes muddy, but this is insignificant, while the death of the child would presumably be a very bad thing. (Singer 1972, 231)

This example may be taken to show that any moral theory which does not condemn *not saving* the child would have something seriously wrong with it. However, if we construe the theoretical component of a moral theory, like we did above, as a device that attaches moral properties to acts and if we understand by 'acts' only acts in the ordinary sense, no moral theory could condemn not saving the child. For these two reasons, then, it makes sense to stipulate, for the purpose of our investigation, that negative acts are also acts.[14] This might seem a bit artificial, but nevermind.

1.2.1.3 Choice Situations

Let us move on, then, to the notion of a choice situation which involves the idea of an act. A choice situation is simply a situation in which an agent faces a choice between different options, where those options comprise negative acts, too.[15] We can specify it by enumerating the possible acts $a_1, \ldots, a_n$ that the agent might do.

[13] On this point, see, e.g., Vermazen (1985, 93) and Mossel (2009). These authors make a case for the view that negative acts are acts. Bach (2010) critically discusses this claim.

[14] Many moral theorists seem to regard the assumption that negative acts are acts as a matter of course. See, e.g., Peterson (2010, 157).

[15] This way of talking is, in fact, imprecise. We should actually say that an agent is faced with a choice between *options* for action. We use a shorthand way of speaking which is intended to mean just that.

Now, important issues arise regarding the relations between acts. The first concerns act identity. When should we regard a_i and a_j as the *same act*? It is important to note that an identity criterion should not be biased in favour or against certain moral theories. To understand what this means, consider Leonard Savage's well-established view on act identity. For the purpose of the choice-theoretical model that he develops he individuates acts by their consequences. He says that

> [i]f two different acts had the same consequences in every state of the world, there would from the present point of view be no point in considering them two different acts at all. An act may therefore be identified with its possible consequences. Or, more formally, an act is a function attaching a consequence to each state of the world. (Savage 1954/1972, 14)

To be sure, whether Savage's view is adequate in the context of rational choice theory is a separate issue which we need not discuss here.[16] It is important to point out, though, that it is indeed misplaced in the context of moral theorizing. It is easy to see why.[17] On Savage's view, we would not be able to make sense of much of the moral-philosophical debate. Let us look at an example which makes this clear. Suppose two acts, a_i and a_j, have the same consequences under all possible empirical circumstances. Suppose, further, that they are identical in all other respects except one. In the view of many philosophers, the fact that a_i consists in telling a lie while a_j involves telling the truth may make the latter right and the former wrong. However, on Savage's view, it does not even make sense to say this. For if a_i and a_j have the same consequences under all empirical circumstances, they are the *same* act. On Savage's view, then, all moral theories, apart from the consequentialist ones, are inconsistent from the standpoint of classic deontic logic. They are self-contradictory because they may categorize one and the same act as both right and wrong. It seems sensible to require, then, that a_i and a_j are seen as different acts for the purpose of moral theorizing if they differ in regards to any feature which might have moral significance. This, of course, does not yield a full-fledged criterion of act identity. In fact, even if it did, it would lack substantive purport since we did not say anything about the factors which may conceivably matter in moral theory. However, I think these issues are not important in the context of our discussion, and so we shall leave them aside. All that counts is that, unlike Savage, we accept that a_i and a_j may not be identical, even though they have the same consequences.

The second issue concerns another important notion, viz. the idea of two acts being alternatives to one another. When are a_i and a_j appropriately called alternatives? According to the most common answer to this question, a_i and a_j are alternatives if and only if they are mutually exclusive. The set of all possible alternatives is then called a *partition*. For the purpose of our investigation, I see no reason to depart from this convention.[18]

[16] For a critique, see Nida-Rümelin (1993).

[17] This point is also made by Dreier (1993, 22).

[18] Note, however, that Carlson (1999a) has offered a different account of alternativeness. He does not require that they be mutually exclusive, but merely that they are two things the agent might do which are not identical. Note that the difference between the standard view and Carlson's

1.2.1.4 Agents

For the sake of completeness, let us also address the notion of an agent. What we said above in regards to acts applies also in regards to agents. The philosophical discussion about the nature of agents is quite complex and not terribly important for us. I shall propose, therefore, to explain the notion by pointing to a paradigm case, viz. a human being in full possession of her mental capacities. This is good enough, I think, for our purposes and the most we can say without entering into controversial territory. If this immediately convinces the reader, she may skip ahead to the section on normative factors.

Now, why is it so hard to define precisely what an agent is? Intuitively, an agent is simply some entity or being which is capable of performing *acts* of some sort. Depending on what we understand by "act," however, certain things may pass as agents which are not, I take it, agents in any philosophically relevant sense. E.g., ask a chemist what an agent is, and she will probably assure you that certain chemicals are agents because they are capable of 'doing' certain things to other chemicals. Oxygen, e.g., is capable of oxidating iron, thereby creating rust. This is, of course, not the kind of agent we have in mind when we engage in moral theorizing. It is not even the kind of agent people typically think of when they use the concept in an everyday sense.

Arguably, the difference between oxygen and a *proper* agent is that when oxygen reacts with iron nothing is being *done* in a relevant sense. Things just *happen*. However, how do we distinguishing this kind of a mere happening from an instance of agency? Of course, we cannot say that, when an oxygen atom does something, there is simply no agent present. For what an agent is, is precisely what we are seeking to explain. Now, one obvious way of drawing the line between the sense in which we can call oxygen an "agent" and the more exacting sense of the word that is relevant in a philosophical tract is by way of looking at the complexity of what the purported agent is capable of doing (cf., e.g., Pears 1971). A chemical agent, such as an oxygen atom, is capable only of doing very simple things, we may say. In contrast, the kind of agent we have in mind when we philosophize is capable of doing more complex things. Now, this may be true. However, the distinction does not get us out of the woods. For there are certain things which are very complex,

---

interpretation only plays a role when it comes to complex acts, viz. acts which are made up of parts, where these parts are themselves acts. If a_i and a_j are not identical and do not contain parts, they are seen as alternatives on both accounts. If, however, a_i and a_j do comprise parts, the two accounts of alternativeness might come to different conclusions. If a_i and a_j have entirely different parts, they are, again, alternatives to one another on both accounts. If they have parts in common, though, they are seen as alternatives only on Carlson's view. This might conceivably matter when we discuss problems for consequentialism that can arise from complex acts (cf., e.g., Feldman 1997, 17–35 and Mukerji 2013c, 306). However, it is not relevant in the present context since our argument does not draw on the issue of complex acts.

but should hardly count as instances of the relevant kind of agency. A volcano, e.g., may 'do' a very complex thing by scattering its ashes over a large area of land. It is not, however, an agent in the sense we are looking for.

Maybe, then, we have to exclude certain items from the list of possible candidates. How about if we allow only living organisms as agents? If this works, it creates, of course, the problem that we have to account for the notion of a living organism. As it turns out, though, it does not work anyhow. For, as Wilson (2009) points out, a person having a seizure is performing rather complex movements, too. But such a person should not be seen as an agent in any sense that might interest us. One striking feature of a person performing movements in a seizure is that these movements are not goal-directed. So maybe we may pick out agents in the sense that is relevant here by the fact that they are doing something in a goal-directed manner? This, too, is problematic. For, as Frankfurt (1978) remarks, a spider moving about on a table seems to control its legs in a goal-directed manner as well. However, it is certainly not an agent in the exacting sense which we are trying to explicate.

Even if we suppose that it is possible to come up with a satisfactory account of the notion of an *individual* agent, further problems arise. For it might be that individual agents are not the only ones there are. There might also be collectives of such agents which constitute a *group agent*. Peter French famously argued that corporations should, under certain conditions, be viewed as autonomous and morally responsible agents (cf. French 1979 and French 1984). More recently, Philip Pettit and Christian List have defended a similar thesis. Since we cannot straightforwardly reduce the choices of some groups to the choices of individuals, they claim that certain groups ought to be seen as autonomous agents (cf. List and Pettit 2011).[19] Philosophers sympathetic to methodological individualism are, in general, critical of such ideas and object to the view that collectives can be agents (cf., e.g., Nida-Rümelin 2011b, 130–141 and Wall 2000). Given the scope and object of our present inquiry, it is certainly impossible to argue for anything like a satisfactory account of what an agent is. As I said above, then, let us simply use the term in its paradigmatic sense, viz. to refer to a human being in full possession of her mental faculties.

1.2.1.5 Normative Factors

Finally, let us talk about normative factors. We can define them as non-moral characteristics of acts that are *relevant* to their moral status. We typically appeal to them when we discuss moral issues with other people. We say, e.g., "Smith should come to the party because, if he does not, Jones will be very disappointed." Or we say: "Jones should not lie to Smith about Suzy coming to the party. That's just not a nice thing to do." Or we say: "Smith should come to the party because he promised Jones to come."

[19] For an argument to that effect, see also Mukerji and Luetge (2014).

These examples illustrate three different kinds of normative factors.[20] The first is an illustration of a *future-related factor*. We say that Smith should come to the party because, if he does not, Jones *will* be disappointed. We point to the relation of Smith's act with a *future* event or state of affairs.[21] The second example illustrates an *intrinsic factor*. When we say that Jones should not lie because lying is not a nice thing to do, we point, it seems, to an intrinsic quality of the act *itself*. The third example instantiates a *past-related factor*. When we say that Smith should go to the party because he *promised* Jones to do that, we mean that Smith's act stands in a particular relation to an event in the *past*, viz. the promise.[22]

The distinction between these three types of normative factors can help us to give a preliminary characterization of consequentialism. We shall consider it below. Before we do, however, let us move on to the practical component of a moral theory.

1.2.2 The Practical Component

Prima facie it might be hard to see what the function of the practical component of a moral theory could potentially be. The theoretical component, after all, deals with principles for right *action*. What could be more practical than that? In fact, however, the theoretical part in and of itself leaves something to be desired in the way of practical guidance. It merely gives us a criterion of rightness that informs us as to the normative factors $e_1, \ldots, e_n$ that make an act right. It does not provide us with any instructions as to how we should go about figuring out whether a given act, a_i, actually possesses the relevant properties $e_1, \ldots, e_n$ and is, hence, right. This is where the practical component comes in. It gives us a method for making moral choices.

In principle, moral theorists can adopt two kinds of decision-making strategies, viz. a Direct Strategy and an Indirect Strategy.[23]

---

[20] This tripartition is suggested by Sher (1983).

[21] I do not speak of a causal relation in this respect, for it seems that certain non-causal relations matter as well. Recall that we allow negative acts to count as acts. If I do not save you from drowning, it seems as though the relation of my act to the relevant future state of affairs in which you are dead is non-causal. Nevertheless, the relation should, it seems, be included on the list of future-related factors.

[22] The particular relation between Smith's promise and his act of going to the party can, of course, be characterized further. Philosophers of language categorize promises as "illocutionary acts" which possess certain "conditions of satisfaction". The relation between Smith's act of promising and Smith's act of going to the party is such that the latter act is a condition of satisfaction for the former (cf. Searle 1983, 10).

[23] The Direct Strategy and the Indirect Strategy are often discussed under different labels. Alan Goldman, e.g., calls the Direct Strategy "complete moral reasoning" (Goldman 2003, 13) and the Indirect Strategy a "second-best strategy." (Goldman 2003, 4) David Gauthier uses a similar distinction between "straightforward" and "constrained" maximization (cf. Gauthier 1987, Ch. VI). The terminology we use is proposed, e.g., in Sumner (1987, 180–181, 1996, 222).

Direct Strategy
Use the criterion of rightness as a decision-making tool.
Indirect Strategy
Use a heuristic choice criterion as a decision-making tool.

On the Direct Strategy, moral agents should use their criterion of rightness as a decision-making tool. That is, they should seek out all possible courses of action, investigate each of the options as to whether they do, in fact, possess $e_1, \ldots, e_n$, and choose accordingly (cf., e.g., Smart 1956, 344 and 1973, 42).[24] Generally, to "choose accordingly" means to do the act that possesses, on balance, the greatest amount of features $e_1, \ldots, e_n$ (cf. Ross 1930/2002, 41). On the Indirect Strategy, moral agents should not use their criterion of rightness to make moral choices. Rather, they should use a heuristic that recommends a simpler choice process and requires less information than the criterion of rightness (cf., e.g., Gigerenzer 2008, 2010).

We can distinguish between two types of heuristics. *Type-1 heuristics* use morally relevant information, but ignore some of it. That is, they use the same information as the criterion of rightness. However, they either ignore some of the information about all available act options or all of the information about some of them. *Type-2 heuristics* use information that is, in itself, morally irrelevant. They have the agent look for certain heuristic qualities, $h_1, \ldots, h_m$, in her choice options that are supposed to track the morally relevant properties $e_1, \ldots, e_n$ to a great extent.[25]

Let us briefly consider the motivation for each strategy. The motivation of the Direct Strategy appears to be clear. It seems as though the safest way to ensure that we act rightly is to ascertain directly whether our acts possess the relevant features $e_1, \ldots, e_n$ that make them right. The motivation for the Indirect Strategy is not immediately clear, however. Why should we use a heuristic that either does not take into account all of the morally relevant information or takes into account information which, by itself, is morally irrelevant? We shall discuss this question in quite a bit of detail in Sect. 4.2.1.2. So a few comments should suffice at this point.

Whether the Indirect Strategy makes sense depends on the content of the criterion of rightness. If the normative factors on which the theory focuses are very easy to ascertain, it is hard to see why the agent should not try to do this in a direct manner. In contrast, if these factors are epistemically more difficult to establish, it may make sense for the agent to use heuristics. To illustrate, consider a typical deontological moral theory whose criterion of rightness says that an act is right if and only if it is not an act of lying, stealing, promise-breaking, and so on. It is easy to apply this criterion directly. We typically know when we lie, steal, and so on. Moreover, we can consciously try to avoid doing these things. However, consider, in contrast, a moral theory which says that an act is right if and only if it maximizes overall well-being.

[24] Pettit (1984, 167) and Feldman (1997, 3) also give a helpful description of the choice procedure involved in the Direct Strategy.

[25] Note that the two types of heuristics do not exhaust logical space. Neither are they mutually exclusive.

Since it is very hard to anticipate how our actions affect people's well-being, it may be helpful to use a Type-2 heuristic – a heuristic "rule of thumb," as Smart (1973, 42) says. We know, e.g., that breaking promises tends to have adverse consequences. So when we make moral choices, we should not attempt to ascertain which act would best promote everyone's well-being. Instead, it seems to make sense to follow certain moral rules, such as the rule not to lie and so on (cf. Hooker 2003, 142). Why? Because there is a fair chance that, if we consciously aim at producing good consequences, we may, e.g., frequently miscalculate, thereby bringing about worse consequences than if had we used a simple heuristic.

1.2.3 Some Distinctions Between Moral Theories

Based on what we said above, we can draw further distinctions between classes of moral theories. Right at the outset, we distinguished theories about acts from theories about different evaluative focal points, and we chose to concentrate exclusively on the former. Among them, we distinguished pluralistic and monistic theories and decided to look only at the latter. Now that we have characterized the idea of a moral theory in a bit more detail, we can distinguish theories according to their *scope*.

The scope of moral theories varies along, at least, two dimensions. They can have limited/universal scope in the sense that they apply to some/all choice situations and insofar as they address themselves to some/all agents. Philosophers commonly regard consequentialist theories as having "universal pretensions" (Goodin 1995, 6) along both dimensions. That is, they are intended to apply to all moral problems and every agent. In Jeremy Bentham's words, they apply to "every action *whatsoever*." (Bentham 1838, 1; emphasis added, NM) This is, then, how we will interpret them in what follows.[26]

As we said above, moral theories can furthermore be distinguished by the normative factors that they accept. We divided these factors into three categories: past-related, intrinsic and future-related. Apparently, a moral theory can either admit all three, either two of the three or either one of the three. There are, hence, seven categories of theories, as Table 1.1 shows.

In this investigation, we are interested only in those theories to which the label 'consequentialist' applies. Intuitively, consequentialist theories should be seen as instantiations of the idea that "the rightness of an act depends only on its

[26]It should be noted that whether or not a moral theory has universal scope in the one sense is logically independent from whether or not it has universal scope in the other sense. A moral theory may apply only to a limited range of moral problems, but to all agents who are faced with these problems. An example of such a theory would be a bio-ethical moral conception which is concerned only with specific moral issues that pertain, in some way or other, to the sphere of biology. Also, a moral theory may apply to all moral problems, but address itself only to specific agents. Professional ethics codes, such as the Hippocratic Oath, are of that type.

Table 1.1 A categorization of moral theories based on normative factors[a]

	Past-related factors	Intrinsic factors	Future-related factors
Consequentialism			x
Absolutist deontology		x	
Antecedentalism	x		
Absolutist deontology	x	x	
Moderate deontology	x	x	x
Moderate deontology		x	x
Relationalism	x		x

[a]In our categorization of moral theories, there are six forms of non-consequentialism. The question what we should call them is certainly a contested matter. One standard way of naming them, however, is this. We describe all moral theories which accept intrinsic factors as forms of deontology. Within deontological theories, we can distinguish between those which entirely ignore the future and those which do not. As Kagan (1998, 79) suggests, they may be called "absolutist" and "moderate," respectively. Moral theories which only regard past-related factors as morally relevant may be called "antecedentalist," as proposed by Sher (1983). Lastly, doctrines which determine the moral status of acts solely based on its relation to events in the past and in the future may be called "relationalist"

consequences." (Sverdlik 1996, 330) That is, we should see them as those theories which allow only future-related factors to count. To put it differently, the idea is that "only future consequences are material to present decisions." (Rawls 1955, 5) This, in fact, seems to capture contemporary usage quite well. The distinction based on normative factors gives us, then, a neat preliminary idea of what consequentialism is. It is the class of all act-focused, monistic moral theories with a universal scope which judge acts only in light of future-regarding factors or consequences. As it will turn out, however, this characterization is problematic because it assumes that there is a clear-cut distinction between an act and its consequences. This is not, in fact, so. We shall say more about this problem in Sect. 3.1.

1.3 Summary

In this chapter, we characterized the object of our inquiry and discussed and clarified some important notions to set the stage for our investigation. We started out by noting that we can partition moral philosophy into normative ethics, applied ethics, and metaethics. Our study, we noted, deals with a general issue in normative ethics, viz. the question whether consequentialism is a tenable moral view. In Sect. 1.1, we homed in on normative ethics. It studies various objects. However, we are interested only in that part of the field which deals with the moral evaluation of *acts*.

In Sect. 1.2, we then discussed moral theories that are used to evaluate acts. We said that they have two components, viz. a theoretical component and a practical component.

We can describe the former as a tool that attaches a particular moral status to the acts available in a given choice situation to a moral agent as a function of certain normative factors. We went on to discuss the notions that are involved in the idea of a moral theory. There are, we said, various moral *stati*, viz. right, wrong, obligatory, and optional, which can be defined in terms of one another if we make certain deontic-logical assumptions. This allows us to interpret the theoretical component of a moral theory as a criterion of rightness. This criterion of rightness tells us which normative factors have to be present in an act for that act to count as right. After that, we clarified various important ideas, starting with that an act. We distinguished acts from negative acts. These are not commonly seen as acts. However, for the purpose of our discussion, we said, we have to regard them as such since our moral theories should be able to evaluate morally significant cases of *in*action. We also discussed the idea of a normative factor which, we said, is a morally significant non-moral property of an act. We distinguished between three normative factors, viz. future-regarding factors (or consequences), intrinsic factors, and past-regarding factors. The object of our study, consequentialism, is often characterized as the class of moral theories which acknowledge only future-regarding factors. We said, however, that this characterization is problematic and announced that we will revisit it in Sect. 3.1 below.

The practical component of a moral theory instructs us as to how we should make moral choices. In this connection, we said, it is useful to distinguish between two practical approaches that we called the Direct Strategy and the Indirect Strategy. According to the former, we should directly apply the theoretical component of a moral theory – i.e. its criterion of rightness – to decide what to do. We should work out whether the normative factors that make an act right are, in fact, present in a given option and decide accordingly. The Indirect Strategy, we said, is the idea that in working out what to do we should not try to ascertain whether an act is, in fact, right. Instead, we should use heuristic criteria.

Chapter 2
Metaethical Foundations

In the previous chapter, we narrowed down what we mean by the term "moral theory," and we developed an understanding (at least a preliminary one) of what it means for a moral theory to be consequentialist. Since it is our goal to criticize all consequentialist theories, we should, in a next step, address the question how we can evaluate them. This is what we shall do in this chapter.

In Sect. 2.1, we will introduce an influential approach to theory evaluation which we will refer to as the Rawlsian Approach. It can be factorized into at least three evaluative criteria, viz. consistency, connectedness, and intuitive fit. On the Rawlsian Approach, we can criticize moral theories by pointing out that they leave something to be desired in regards to at least one of these criteria. Since the primary objections to consequentialism draw on intuitive fit, we shall focus on this sub-criterion alone.

We can distinguish between three interpretations of the criterion of intuitive fit, viz. the Top-Down Approach (TD), the Reflective-Equilibrium Approach (RE), and the Bottom-Up Approach (BU). In Sect. 2.2, we will discuss these three approaches. It will be our aim to establish that we can reject TD and that either RE or BU is justified. This will play a crucial role in our argument. Here is why. To refute consequentialism, we will draw on our moral intuitions about individual cases. Such a procedure is admissible both on RE and on BU, but would be ruled out by TD.

Having clarified the interpretation(s) of intuitive fit on which our argument relies, we will proceed to develop this criterion into a workable method for our investigation. This is necessary for the following reason. Intuitive fit merely states a philosophical ideal, viz. that our moral theories should fit our intuitions. It does not, however, give us a methodic procedure that we can use to test whether a given moral theory does, in fact, live up to this ideal. In Sect. 2.3, we will, hence, discuss how we can apply intuitive fit in moral argumentation. Our answer to this question is the Provisional Fixed Point Approach (PFPA). On PFPA, we evaluate moral theories as follows. We look for very strong intuitions about cases – provisional

N. Mukerji, *The Case Against Consequentialism Reconsidered*,
Theory and Decision Library A: 51, DOI 10.1007/978-3-319-39249-3_2

fixed points – and examine whether a given moral theory can match them. If not, we reject it (subject to the *proviso* that the best moral theory is, in fact, compatible with these provisional fixed points).

PFPA leaves open which kinds of cases we should use. In Sect. 2.4, we will introduce 'trolley cases' which we will use in our argument against consequentialism. We will discuss their characteristics and possible uses. Since there have been many objections to their applications in moral philosophy, we will also discuss their *pro*s and *con*s.

In Sect. 2.5, we will close the chapter with a brief summary of the main points.

2.1 The Rawlsian Approach

How can moral theories be evaluated? When we ask this question, we leave the field of normative ethics and set foot into the area of metaethics and moral epistemology, in particular. This is worth emphasizing. After all, at the beginning of the second chapter we characterized the subject of our inquiry as a matter of *normative* ethics. Our digression, however, seems justified.[1] Our goal is to develop a convincing critique of consequentialism. To do this, we need evaluative criteria. After all, every objection to a moral doctrine is a claim that it falls short of a particular evaluative criterion.

Now, which criteria should we use to evaluate moral theories? One obvious answer is to say that a theory (or, at least, its theoretical component) should be *true* and that we should, hence, adopt *truth* as our evaluative standard. This approach, however, is not terribly fertile. Moral philosophers are very much divided on the issue of whether or not moral theories can be true. *Moral realists* affirm this, while *moral anti-realists* deny it.[2] However, even if there was agreement on the matter of moral truth, it appears that this would not help us much. For the issue of the truth of moral theories – if, in fact, it can be had – might be largely independent of the question whether we should accept them (cf. Railton 1984, 155). To see this, consider the analogous controversy between scientific realists and instrumentalists in the philosophy of science. A stylized picture of their debate looks like this: *Scientific realists* believe that our scientific theories ought to convey the truth about the world, presupposing, of course, that these theories

[1] On this point, see also Pettit (1997/2007).

[2] An instructive discussion of the realist position is offered by Sayre-McCord (2011). For a concise general examination of anti-realism, see Joyce (2009). Moral anti-realists are commonly partitioned into non-cognitivists and error theorists. Non-cognitivists believe that our moral persuasions are not apt for truth or falsity. They suggest that they are, rather, expressions of emotional attitudes (cf., e.g., Barnes 1934, Ayer 1952, 102–120 and Stevenson 1937) or prescriptions (cf., e.g., Hare 1961; Gibbard 1990). Error theorists believe that moral views can have a truth value, but think that all our moral judgements are false. The classic statement of such a view is found in Mackie (1977).

can have a truth value. *Instrumentalists*, on the other hand, deny that scientific theories are apt for truth. They believe that they do not refer to something *real*, but are rather devices for predicting observable phenomena. It seems, then, that scientific realists and instrumentalists are at an impasse when it comes to the issue of theory evaluation. Realists will evaluate theories regarding whether or not they are true, while instrumentalists will assess them in terms of their predictive power. However, this picture is not accurate. It is important to keep apart the issue of theory *acceptance* and matters of *ontological interpretation*. Both realists and instrumentalists, it seems, can *accept* theories based on the same criteria since "[t]he acceptance of a theory involves only the claim that it is empirically adequate, not its truth on the theoretical level." (Niiniluoto 2011) Nida-Rümelin (2002, 45) emphasizes this point, too, and throws a bridge to moral theory. For reasons of space, we shall not go into the reasoning he gives. Rather, we shall simply take it for granted that the distinction between the epistemological issue of theory acceptance and the ontological problem of a theory's aptness for truth, as drawn in the philosophy of science, carries straight over to moral philosophy. In our inquiry, then, we shall put aside questions about moral truth and turn immediately to the evaluative criteria for moral doctrines.

Alas, the field of moral epistemology, which deals with these criteria, is also highly controversial. Hence, any stipulations we might make about the criteria of evaluation for moral theories are bound to be controversial as well. Unfortunately, though, we have to make at least some such stipulations. For, plainly, "[i]f we take up a point of view stripped of all evaluative conviction, we have no basis for evaluation." (Hooker 2003, 11).

There are various contrary viewpoints about the justification of a moral theory. Some theorists suggest that it is justified if it conforms to the *will of God* (e.g. Quinn 1990). Accordingly, the method of evaluation might consist in comparing the content of a moral doctrine with the laws that are laid down in some sacred text. Alternatively, it may consist in personal revelation. Moral naturalists maintain that ethics should be grounded in *empirical facts*. These theoreticians may favour a scientific study as a way of making progress on moral questions.[3] Others, most notably Kant (1785), advocate a rational approach to ethics which regards *pure practical reason* as the ultimate arbitrator on matters of right and wrong. A further approach to assessing moral doctrines is the *intuitionist method*. It assumes that we can know certain moral 'facts' simply by intuiting them (e.g. Ross 1930/2002; Prichard 2002; Crisp 2006). Another view on moral justification is due to Hare (1981). He believes that we can support his moral theory (a version of preference utilitarianism) through a careful analysis of the meaning of moral language and, in particular, the property of universalizability that, as Hare argues, attaches to moral utterances.[4] Nowadays, however, the most common conceptions of moral justification appear to be variants of what may be called the Rawlsian Approach.

[3] For a summary article on moral naturalism, see Lenman (2008).

[4] I am grateful to Julian Nida-Rümelin for pointing out to me that my argument does not cover versions of consequentialism that follow Hare's justificatory approach.

Like the intuitionist method, it also allows intuition to play an important role (cf. Rawls 1951, 1971/1999). Due to its status as the standard view of justification in modern ethics, we shall adopt it. Given the task that lies ahead of us, we shall not, however, attempt to justify it. The remainder of this section shall serve merely to explain and interpret the Rawlsian Approach.[5]

In his famous book *A Theory of Justice* (1971/1999), John Rawls starts his discussion of the issue of justification in moral theory by pointing out that human beings have a remarkable faculty. In the context of his theory of justice he calls it a "sense of justice." More generally, we could call it a "moral sense." Now, this moral sense can be thought of as the capacity to form moral intuitions at various levels of generality. We are capable of having high-level intuitions about abstract moral principles that cover a whole range of moral cases or even all cases. Moreover, we can form low-level intuitions which cover only a narrow variety of scenarios or, in the limiting case, just one particular moral problem.

We may regard our moral sense as analogous to our sense of vision. When we consider a particular moral judgement we can *sense*, as it were, whether it is correct. In a similar way, one may say, we can *see* whether an object has a particular colour (cf. Harrison 1967, 72).[6] This analogy brings out an important point about moral intuition. Sometimes we choose not to believe what we see. Similarly, we may sometimes opt to disbelieve our moral intuitions (cf. Greene 2008, 63 and Sinnott-Armstrong 2008, 48). To make this distinction clearer, let us use a standard example: Consider the well-known Müller-Lyer illusion (cf. Müller-Lyer 1889), as shown below (Illustration 2.1).[7]

It contains two arrows whose shafts are of the same length. However, the fins of the two arrows point in different directions. The fins of the one arrow point outwards, while the fins of the other arrow point inwards. This creates the illusion that the shaft of the first arrow is longer than the shaft of the second though both are of the same length. We can convince ourselves that this is true. We can measure the two shafts with a ruler. Interestingly, this will not persuade our eyes. The one shaft still looks longer than the other. In that case, however, we should choose not to trust appearances. Analogously, an act might *seem* wrong, and the claim that it is right may strike us as inadequate. However, on reflection, we may come to believe,

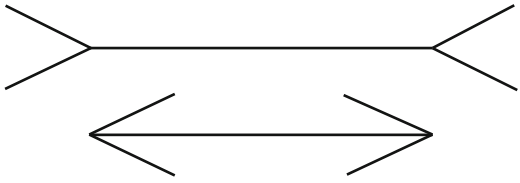

Illustration 2.1 The Müller-Lyer illusion

[5]For a detailed defence, however, see Daniels (1996).

[6]Similarly, Appiah (2008, 113) uses an analogy between reasons for perceptual belief and reasons for action (as opposed to reasons for normative beliefs that we are interested in).

[7]Maria Mukerji suggested this example.

nevertheless, that it is not. That is, we may eventually come to believe that our intuition is unreliable in this instance just like our sense of vision is sometimes inaccurate. It is important to stress, then, that there is a difference between having a moral intuition and adopting this intuition on reflection as a *belief* (cf., e.g., Kagan 2001, 55 and Tännsjö 2011, 307).[8]

This said, we can introduce an approximate criterion of justification for moral doctrines. Rawls says that we may provisionally think of them as the "attempt to describe our moral capacity" (Rawls 1971/1999, 41) and the high-level and low-level moral intuitions that issue from it. This suggests that their acceptability is determined, at least in part, by how well it fits the moral claims which we intuitively endorse. Elsewhere, I called this criterion "intuitive fit" (Mukerji 2013c, 299).

It is evident, though, that this criterion of intuitive fit cannot be the only measure for the acceptability of moral theories. If it were, we would not need them. We should then directly endorse a complex of very specific and unconnected moral principles that just state our moral intuitions. We should, in other words, adopt an "unconnected heap of duties" (McNaughton 1996). Such a construct, however, would not seem very attractive. First of all, it might not even count as a moral theory on the monist interpretation which we have adopted for the purpose of this inquiry. Recall that, according to this interpretation, the theoretical component of a moral theory can be represented as a criterion of rightness. It is doubtful whether a single criterion can be made to fit all our considered moral judgements. Furthermore, a theory like that would not do what we may reasonably expect it to do, viz. "to achieve an acceptable coherence" (Daniels 2011) between our various intuitive judgements that explains and justifies them.

To be sure, by "coherence" we do not mean "coherence with our moral intuitions" (Wood 2008, 47). This requirement is entirely distinct from intuitive fit. It concerns the internal structure of a moral theory, i.e. the relations in which its individual moral claims stand to one another. It does not concern, that is, its external relation to our intuitions. In fact, a fully coherent theory may be one which consists only of highly counter-intuitive claims (cf. Sayre-McCord 1985).

Now, what does the notion of coherence involve?[9] According to a standard interpretation, it requires, first of all, that the claims we endorse not contradict each other. As Rawls points out, there is no reason to suppose that our intuitive moral judgements fulfil this requirement (cf. Rawls 1971/1999, 42). Verifying this is easy. Take two views which seem intuitively appealing. E.g., take the idea that an act which produces the best possible consequences is always right.[10] Moreover, take the view that harming an innocent person is always wrong. Both of these views, I believe, appear, intuitively, quite credible. At any rate, they should appear credible

[8]Note that this sense of "intuition" is different from the sense in which some moral intuitionists have employed the term. They have apparently taken self-evident truth as a defining characteristic of an intuition (cf. Lillehammer 2011, 184).

[9]The notion I work out here has been described as a narrow notion of coherence (cf. Rawls 1974–1975, Daniels 1979).

[10]Even critics of consequentialism have conceded that this idea seems *prima vista* trivially true (cf., e.g., Nida-Rümelin 1993, 1).

to the layman. Then, take a case, e.g., Judith Jarvis Thomson's *Fat Man* case (cf. Thomson 1976, 207–208). Imagine that I am standing on a footbridge over a railway, watching a runaway trolley hurtling down the tracks. I can tell that, if nobody stops the trolley, it will crash into and kill the five people who are working on the tracks. The only way for me to halt the trolley is to push a fat man, who is standing next to me, off the bridge and onto the tracks. Sure enough, this will kill him. However, it will stop the trolley and save the five. Arguably, then, pushing the man has better consequences than not pushing him. It will save a net four lives. According to the first intuition, then, it is right for me to shove the man off the bridge. Yet, since it will also inflict severe harm on an innocent person, it is wrong according to the second intuition. The latter says that it is always wrong to do this. Hence, our pre-theoretical intuitions contradict each other in this case. A moral theory ought to avoid such contradiction. This is the requirement of *consistency*. Insofar as it is a part of the requirement of coherence, it is also a part of the Rawlsian Approach (cf. Kappel 2006, 132).

However, consistency is only a necessary and not a sufficient condition for coherence. The beliefs that $7 + 5 = 12$ and that snow is white are consistent. But there is nothing which *connects* them. Hence, a belief system which contains only those two convictions is not coherent. For coherence requires, secondly, what could be called "systematicity" or "connectedness." (cf. Sayre-McCord 1985) The elements of a moral theory are supposed to systematically link up with one another. This is important to create a sense that the doctrine as a whole is not just an arbitrary collection of randomly assorted components.

Let us consider an example from applied ethics which brings out rather nicely how the criterion of systematicity can be exploited to support a moral claim.[11] Suppose I have the following intuitions about two moral cases.

Intuition 1
If I come across a shallow pond where a child is drowning, and I can save the child at the trivial cost of ruining my best pair of shoes, I ought, morally, to save the child.

Intuition 2
Giving to charity, though it is undoubtedly a good thing to do, is not morally required of me. It is optional.

Singer (2009) points out the following. It is reasonable to assume that, if I give a relatively little amount of money, comparable to the costs of a good pair of shoes, to charity, this suffices to save a child from death by starvation or preventable diseases (e.g. measles, malaria, diarrhoea). Why, then, should it be wrong for me not to save the child in front of me, but permissible not to give to charity? My two intuitions seem to be hard to square. However, what exactly is the problem here? These intuitions are clearly consistent. Apparently, the reason I should be troubled by them lies, then, in their seeming lack of unity. It lies in the fact that my intuitions are entirely disconnected. For this reason, they appear to be arbitrary, and

[11] The example is taken from Singer (1972). We have already used part of it on page 8.

I cannot be sure that they "do not simply express some form of irrational prejudice." (Lillehammer 2011, 176) Evidently, if I would reject Intuition 2 and accept a duty to give to charity instead, I could square this view with my Intuition 1, which says that it is wrong not to save the child. Then, I could bring both my views under, e.g., Singer's proposed *principle of harmprevention*. It says that I ought to prevent a great harm if this costs me comparatively little. I feel a strong inclination, then, to revise my views in the way Singer suggests because I want to make them coherent. This, of course, is precisely the point of the argument. As we can see, then, the systematicity or connectedness of our views can be used as an evaluative criterion besides intuitive fit and consistency.

Let us take stock, then. We have established that the Rawlsian Approach to moral evaluation contains, at least, three sub-criteria. I proposed to call these sub-criteria intuitive fit and coherence. We can factorize the latter into consistency and systematicity or connectedness. In short, then, on the Rawlsian Approach, a moral theory is acceptable to the extent that it is consistent, fits our moral intuitions, and establishes explanatory connections between them.

This idea obviously requires interpretation. Before we proceed by considering various understandings of it, however, let me add a short note on *simplicity* or *economy*. Philosophers often suggest that the acceptability of a theory depends partly on parsimony in its use of fundamental concepts. I propose to disregard this criterion, however – for two reasons. Firstly, the simplicity of a moral doctrine cannot, it seems, make up for its lack of intuitive fit and consistency (cf., e.g., Williams 1973, 137 and 1985, 17; Ross 1930/2002, 23).[12] Secondly, there appears to be a significant correlation between systematicity and simplicity. Theories which contain a dense web of systematic connections between its individual parts tend to possess fewer fundamental concepts than others.[13]

2.2 Interpretations of the Rawlsian Approach

There are various possible interpretations of the Rawlsian Approach. It is a multi-dimensional evaluative criterion. For one thing, then, it is possible to attribute different weights to its distinct sub-criteria. Following Kant's dictum that consistency is a philosopher's greatest duty,[14] this sub-criterion may be seen as a disqualifier. Trade-offs, however, can be made between intuitive fit and connectedness (cf., e.g.,

[12]E.g., Bernard Williams has this to say about simplicity: "If there is such a thing as the truth about the subject matter of ethics (. . .) why is there any expectation that it should be simple? In particular, why should it be conceptually simple, using only one or two ethical concepts, such as *duty* or *good state of affairs*, rather than many? Perhaps we need as many concepts to describe it as we find we need, and no fewer." (Williams 1985, 17; emphasis in the original).

[13]This fact is illustrated, e.g., by Classic Utilitarianism which is often described as both a highly systematic and a rather simple doctrine.

[14]Parfit (2011, xlii) ascribes this *dictum* to Kant.

Kappel 2006, 132). Moral philosophers, of course, differ on the appropriate trade-off ratio.[15] Some theoreticians have vigorously taken the stance that intuitive fit is more important than connectedness. G. E. Moore, e.g., may be interpreted in that way. He says that it is not "the proper business of philosophy, however universally it may have been the practice of philosophers," "[t]o search for 'unity' and 'system' at the expense of truth" (Moore 1903/1959, 222). Tom Nagel has maintained that, "[i]f arguments or systematic theoretical considerations lead to results that seem intuitively not to make sense (. . .), then something is wrong with the argument and more work needs to be done" (Nagel 1991, x). John Stuart Mill and Immanuel Kant famously took the contrary stance. They emphasized the importance of unity and system in moral theory (cf., e.g., Mill 1863, Kant 1785).[16] A further interpretive issue arises in regards to the sub-criterion of intuitive fit. It requires that a moral theory fit the judgements we intuitively endorse. We need to specify this and shall do so in what follows.

There is a consensus, I think, that intuitive fit does not require a moral theory to fit all our intuitions. It merely requires that it match the ones which possess an "initial credibility" (Scheffler 1954, 181). (Ultimately, we may not even demand that it meet all of those since the set of initially credible intuitions may be inconsistent). Rawls says, e.g., that we can discard "those judgments made with hesitation, or in which we have little confidence. Similarly, those given when we are upset or frightened, or when we stand to gain one way or the other can be left aside." It is easy to see why we should dismiss such intuitions as irrelevant. They are dubious from the start. That is, they are not initially credible. We should restrict ourselves, then, to intuitive judgements "rendered under conditions favorable to the exercise of the sense of justice, and therefore in circumstances where the more common excuses and explanations for making a mistake do not obtain" (Rawls 1971/1999, 42). Rawls calls them "considered judgements."

The category of considered judgements is useful to draw attention to the fact that we should not see all moral intuitions as relevant to ethics and that it is important to preselect them before we use them to test moral theories. However, it is by no means clear what makes an intuitive moral judgement a *considered* moral judgement. It is unclear, that is, what distinguishes disliverances of intuition with initial credibility from ones that lack it. Rawls only gives us a few examples. Sure enough, the factors he mentions are plausible. Some emotions are undoubtedly associated with the way we judge moral matters (cf., e.g., Greene et al. 2001; Greene 2008; Huebner, Dwyer, and Hauser 2009), as is self-interest (cf., e.g., Bazerman and Tenbrunsel 2011, 50; Thompson and Loewenstein 1992 and Wright 1996, 13). However, drawing only

[15] To be sure, our talk of a "trade-off" should not be taken too literally. It does not suggest that there are precise, quantitative measures for the overall intuitive fit and systematicity of a moral theory or a single metric on which both can be compared. A moral theory may be evaluated in terms of both its intuitive fit and its systematicity based on a "'seat of the pants' feel." (Putnam 1981, 132) The appropriate trade-off relation, i.e. the overall fit of the theory with our evaluative criteria as a whole, may be determined in the same way.

[16] In fact, both rejected intuitive fit as an evaluative criterion.

on Rawls's ideas, we cannot make progress towards a definite interpretation of the sub-criterion of intuitive fit. So let us look at some views moral philosophers have expressed.

We may distinguish, roughly, between three interpretations of intuitive fit. Each is based on a different second-order theory about the credibility of our moral intuitions. We shall refer to the first as the Top-Down Approach (TD), to the second as the Reflective-Equilibrium Approach (RE) and to the third as the Bottom-Up Approach (BU).[17] To distinguish these three approaches, we need to draw on a differentiation that we made above. Above we discerned low-level intuitions and high-level intuitions. Low-level intuitions, we stipulated, are those intuitions which are less abstract. They concern only one particular case or a very narrow range of cases. In contrast, high-level intuitions are about more abstract moral principles. They cover a broader range of cases or even all possible ones.[18] Now, the distinction is certainly both vague and non-exhaustive.[19] However, I believe that it suffices for the purpose at hand. There are cases in which it seems pretty clear that we are having a low-level or high-level intuition. E.g., if we have an intuitive conviction about whether or not it was wrong for Bill Clinton to lie (or tell a misleading truth) to the American public when asked whether he had sexual relations with Monica Lewinsky, we clearly have a low-level intuition. We make an intuitive judgement which covers only one very specific case. In contrast, if we have the intuition that one ought to act only according to that maxim whereby one can, at the same time, will that it should become a universal law, we clearly have a high-level intuition. Similarly, we certainly have a high-level intuition if we intuitively judge that all sentient beings always deserve to be given equal consideration. Such intuitions cover all possible cases. There are surely moral propositions which lie in between these examples and do not fall clearly on either side of the distinction. This, however, should not be a problem in the present context. With the differentiation between low-level and high-level intuitions in mind, we can define the three interpretations of intuitive fit as follows:

Top-Down Approach (TD)
Only high-level intuitions are initially credible. Hence, moral theories should be judged only by the degree to which they fit our high-level moral intuitions.

[17] There are, of course, innumerable possibilities when it comes to the concrete shape of the respective moral-epistemological theory. For our purposes, however, a rough classification suffices.

[18] Sandberg and Juth (2011) employ a similar distinction between what they call "practical" and "theoretical" intuitions though they draw it in terms of a different criterion. They take them to have different objects. Practical intuitions, they say, are intuitions about cases which is what we call low-level intuitions. Theoretical intuitions are intuitions about moral principles and, apparently, certain metaethical questions too (e.g. the question "what morality is about"). Theoretical intuitions in Sandberg's and Juth's sense should normally be high-level intuitions. There may, however, be instances of intuitions about very specific moral principles which apply only to very few cases. These are, then, theoretical low-level intuitions.

[19] Some authors have distinguished a further category, to wit, *mid*-level principles (cf., e.g., Bell 2007, 71).

Reflective-Equilibrium Approach (RE)
Both high-level and low-level intuitions can be initially credible. Hence, moral theories should be judged in accordance to the overall fit with intuitions both at the high and the low level.

Bottom-Up Approach (BU)
Only our low-level intuitions are initially credible. Moral theories should, hence, be judged only based on the degree to which they fit our low-level intuitions.

To make sense of these approaches, it may be instructive to connect them to the work of some acclaimed philosophers. TD, it seems, can clearly be attributed to the utilitarian philosopher Henry Sidgwick (cf. Singer 1974).[20] In the sixth preface to his legendary book *The Methods of Ethics* (1874/1907), written shortly before his death, Sidgwick provides evidence of this. He explains that he felt forced, at some point, "to recognize the need of a fundamental ethical intuition" (a high-level intuition, as we call it) without which his utilitarian moral philosophy could not "be made coherent and harmonious." (Sidgwick 1907, xvi–xvii)[21]

RE is, as it were, the standard interpretation of intuitive fit. We can attribute it to John Rawls.[22] He describes the process of drawing up a moral theory as a "going back and forth" (Rawls 1971/1999, 18) between the level of the moral principles he seeks to derive, the even more abstract ideas which serve as the premises of this derivation and intuitive convictions about particular cases to which principles are subsequently applied. In doing this, he acknowledges that considerations at all levels of generality – high and low – play a role in the assessment of a moral doctrine.

BU, too, appears to be quite a widespread view. A moral theorist "often starts with intuitions about particular cases and attempts to uncover the general moral principles that underlie these intuitions" (Kahane 2013, 421). Those who favour this approach to theory construction should also hold the corresponding view about theory evaluation. They should hold that a moral theory is acceptable insofar as it implies our low-level intuitions. This view is clearly present, e.g., in works of Phillippa Foot and Judith Jarvis Thomson. They are well-known for their pioneering work on *trolley cases*.[23] These are thought experiments which are designed to trace

[20] Note, however, that this is not the most common reading of Sidgwick. Many philosophers follow Rawls (1971/1999) who, drawing on Schneewind (1963), claims that Sidgwick endorsed RE which is the approach favoured by Rawls himself. I believe that Sidgwick was misinterpreted by Rawls and Schneewind and that the remarks he makes about common-sense morality were falsely taken to represent his own views.

[21] Some may think that Immanuel Kant would also fit the description of TD since he undoubtedly pursued moral philosophy in a top-down fashion. His ambition was to develop a system in which every moral proposition is justified in terms of one supreme principle of morality: the Categorical Imperative. For this reason, some have seen him as a proponent of the TD approach to intuitive fit (cf. Singer 2005). But this would be a mistake since Kant never accepted intuitive fit as an evaluative criterion for moral theories (cf. Nida-Rümelin 2002, 22).

[22] John Rawls explicitly rejects TD to which he refers as the "Cartesian" view. He says that "[t]here is no set of conditions or first principles that can be plausibly claimed to be necessary or definitive of morality and thereby especially suited to carry the burden of justification" (Rawls 1971/1999, 506).

[23] A further major exponent of BU is Frances Kamm. She explicitly states BU in Kamm (2007, 5) and Kamm (1996, 10–12). Interestingly, even the utilitarian philosopher and economist John

out our low-level intuitions to construct high-level moral principles which explain them. We will consider them in more depth below.

We have distinguished the various interpretations of intuitive fit. Now, which one is adequate? Before we address this question, we should note, however, that our argument does not depend on any particular view being *correct*. Rather, the only thing that counts is that one view – viz. TD – is inadequate or, conversely, that either BU or RE is adequate.[24] In what follows, we shall try to establish this by looking at the rationales for each approach.

2.2.1 The Top-Down Approach

Let us look, first of all, at the reasons for TD. Its proponents argue that low-level intuitions are unreliable. They think that we should, hence, rely only on high-level intuitions which they take to be more credible. Their case for TD is based, then, largely on an argument against low-level intuitions. In recent times, proponents of TD have increasingly done this using empirical findings from psychology and related areas. Of course, since this is not a tract in moral epistemology, we can only look at a few examples.[25]

Harsanyi has, at times, made remarks that may be read as expressing a sympathetic attitude towards BU: "Should the axioms of my ethical theory turn out to possess morally unacceptable practical implications," he says, "(. . .) then I must be always willing to revise my axioms" (Harsanyi 1977b, 26).

[24]For an argument to that effect, see Mukerji (2014).

[25]It should be noted that many philosophers regard empirical considerations as beside the point when it comes to the evaluation of moral theories. Drawing on well-known ideas predominantly by Hume (1888/1960) and Moore (1903/1959), they argue that there is a metaphysical divide between the spheres of 'Is' and 'Ought' and that moral facts are distinct from and not definable in terms of natural facts. Ethics, they say, is hence *autonomous* in the sense that no empirical facts could conceivably influence the moral question whether a given action is right or whether a given moral theory is adequate. In reply to such concerns, it should be stressed that empirical arguments do not generally claim that normative propositions follow straightforwardly from empirical propositions (Mukerji 2015). Hume's and Moore's points are usually conceded. It is claimed, however, that certain information about the workings of our moral faculty may be useful when it comes to figuring out which principles are justified. But their justification may itself be independent from empirical matters. An analogy may be useful to drive home the point. Consider our visual sense. It is normally reliable and helps us to figure out what goes on in the world around us. But there are optical illusions (e.g. the Müller-Lyer illusion that we considered on page 20). We should, therefore, be interested in understanding the conditions under which these illusions arise. For, when they obtain, we are, it seems, well advised to put less faith in our visual perceptions than we normally do. Similarly, we may believe that our moral sense is normally reliable. But there may be certain facts about it that cast doubt on its judgements in certain situations.

One obvious requirement for the reliability of an intuition is that it passes Sidgwick's "criterion of consent."[26] (Sidgwick 1879, 108) He argued that

> the denial by another of a proposition that I have affirmed has a tendency to impair my confidence in its validity. (. . .) For if I find any of my judgments, intuitive or inferential, in direct conflict with a judgment of some other mind, there must be error somewhere: and if I have no more reason to suspect error in the other mind than in my own, reflective comparison between the two judgments necessarily reduces me temporarily to a state of neutrality. (Sidgwick 1907, 341–342)[27]

In other words, if we encounter a reasonable person who disagrees with us about some moral question, this should decrease the confidence that we are right. (We need not even go so far as to become completely neutral, as Sidgwick suggests.) Now, the question is whether people seem to disagree comparatively more about low-level matters. There is one difference between low-level and high-level intuitions which may suggest that there must be more disagreement regarding low-level intuitions. One could claim that differences about low-level intuitions are much more likely than opposing intuitive views at the higher level since there are innumerable cases at the low level while there is a limited class of principles which cover all cases. So it is much more likely that we will ever reach agreement on a confined set of high-level judgements than about a potentially infinite amount of convictions about particular cases. A further point one could make is that people disagree regarding their low-level intuitions to quite a large extent.

The second reason one might give preference to high-level intuitions has to do with what psychologists call "framing effects." Let us, first of all, consider what these effects are. In an oft-cited paper, Amos Tversky and Daniel Kahneman report that people's intuitions about how one should act in particular cases may change depending on how the choice is verbally framed.[28] They had two groups of participants face a decision problem between two policies A and B and C and D, respectively. The description of the case was as follows:

> Imagine that the U.S. is preparing for an outbreak of an unusual Asian disease which is expected to kill 600 people. Two alternative programs to fight the disease, A and B, have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows: (Tversky and Kahneman 1981, 453)

The first group got this description of A's and B's consequences.

> If program A is adopted, 200 people will be saved. If program B is adopted, there is a 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved. (Tversky and Kahneman 1981, 453)

[26] Note, however, that the criterion of consent is not universally accepted (cf., e.g., Smart 1956, 346).

[27] This point has been made by other authors, e.g. by Ross (1939, 88).

[28] It should be noted that philosophers have recognized the existence of framing effects before psychologists did. Williams (1970), e.g., describes the phenomenon in the context of the issue of personal identity and suggests that intuitions about a given scenario can differ under two equivalent descriptions.

The second group was given a choice between C and D (instead of A and B). This was the description of their consequences:

> If program C is adopted, 400 people will die. If program D is adopted, there is a 1/3 probability that nobody will die and a 2/3 probability that 600 people will die. (Tversky and Kahneman 1981, 453)

Each group stated their preferences about the programmes.[29] In the first group, 72 % favoured option A and 28 % option B. In the second group, 22 % preferred option C and 78 % option D. Note that the only information participants had about programmes A, B, C, and D was regarding their effects. Programmes A and C and programmes B and D, respectively, had the same effects though these were framed differently (i.e. A and B regarding lives *saved*, C and D relating to lives *lost*). Such a difference in verbal framing, it seems, should not make a difference in the moral evaluation of the respective options. However, as Tversky and Kahneman (1981) showed, it does appear to affect people's judgement.[30]

Now, why should framing effects speak against low-level intuitions and for high-level intuitions? This is because lower-level intuitions seem particularly susceptible to framing effects while higher-level intuitions appear to be comparatively immune to them. This suspicion, one may argue, is supported by the fact that hardly any research shows that framing effects exist in high-level intuitions, whereas there is plenty of evidence which suggests that low-level intuitions change due to framing (e.g. Tversky and Kahneman 1981; Petrinovich and O'Neill 1996; Haidt and Baron 1996).

A third reason to be sceptical of low-level intuitions *vis-a-vis* high-level intuitions is that there might be *debunking explanations* for intuitive judgements at the low level, but not at the high level. In particular, the fact that we have certain low-level intuitions might be because we have genetic or cultural dispositions for emotional responses to specific cases. Take, e.g., our intuitions about killing. The neuropsychologist and philosopher Joshua Greene and his colleagues conducted a series of experiments to study how firmly and under which conditions we disapprove of killing an innocent person. Amongst other things, they compared a pair of cases which they called *Footbridge* and *Remote Footbridge*, respectively. In *Footbridge*, participants were supposed to imagine a situation in which a runaway trolley threatens to kill five people who are working on the tracks of a railway. The only chance

[29] To be sure, participants were not asked specifically for their moral intuitions in these cases. But since their own self-interest was not at stake in either case, it is reasonable to suppose that their favoured choice is based on moral considerations only (cf. Sinnott-Armstrong 2008, 55).

[30] There are various types of framing effects and not all of them depend on the wording of cases. A quite famous sequence-related framing effect is associated with the work of the philosopher Peter Unger. To discredit intuitions about cases Unger (1996) remarks that in moral thought experiments – most notably the "trolley cases" due to Foot (1978, 19–32) and Thomson (1976, 1985) – there are usually only two choice options. And we tend to have strong intuitions for or against, respectively, one of the options. Using his "Method of Several Options," he attempts to show that our intuitions about which option is right change as we add further options to the choice problem.

to save them from the approaching trolley is to push a fat man off the footbridge over the tracks, thus using him as a trolley stopper. This would kill him. However, it would save five lives. The *Remote Footbridge* case involves basically the same scenario, except that this time the only chance to save the five is to hit a switch. This switch will open a trapdoor on which the fat man is standing. He will drop onto the tracks and will stop the trolley. Once more, this will kill him. However, it will also save the five (as does pushing the fat man in the previous case). Most participants judged that pushing the man in *Footbridge* is wrong. Significantly fewer people, however, had the intuition that it is wrong to kill someone by hitting the switch in *Remote Footbridge* (cf. Greene et al. 2009).[31] Greene's explanation for this is that our species has, as a matter of contingent fact, developed emotional "point-and-shoot" responses to types of cases that our ancestors frequently faced. We have an aversion to anything that feels like applying force in an up-close and personal manner. Now, "[t]he thought of pushing the stranger off the footbridge elicits these emotionally based responses" (Singer 2005, 348), while the notion of hitting a switch, which essentially produces the same effect, does not.[32] Given that "these moral intuitions are the biological residue of our evolutionary history," however, "it is not clear why we should regard them as having any normative force."[33] (Singer 2005, 331)

The important point to be added to this is that we apparently cannot say the same about our high-level intuitions. They are abstract and do not elicit the contingent emotional responses which we can explain (away?) by our evolutionary history. They are not, as Singer claims, intuitions in the ordinary sense, but rather "rational intuitions" (Singer 2005, 351) and, hence, more trustworthy and relevant for the assessment of our moral doctrines.

2.2.2 The Reflective-Equilibrium Approach

Now, what can those who support the other approaches say in reply to the above? Both proponents of RE and BU need to show that at least some low-level intuitions seem initially credible and that the justification of a moral theory should, hence, be assessed partly or entirely regarding its fit with low-level intuitions. How can this be done?

[31] Participants were asked to report on a 9-point scale how strongly they approved/disapproved of the killing. Killing the man in the footbridge case received an average rating of 3.89 (standard error 0.22). Killing the man in the remote footbridge case received an average approval of 5.14 (standard error 0.20).

[32] Indeed, the claim that many of our intuitive judgements about particular cases are based on emotional responses is confirmed by a number of recent neuroimaging studies performed by the psychologist and philosopher Joshua Greene (cf. Greene 2008). Greene showed that certain judgements about cases are associated with increased neural activity in emotion-related areas of the brain (e.g. posterior cingulate cortex, medial prefrontal cortex, amygdala).

[33] Crisp (2006, 24) makes essentially the same point.

Let us reconsider the issue of interpersonal variation. Above, we said that there were innumerable cases. Hence, there seems to be a much greater potential for disagreement about low-level intuitions. One may corroborate this impression using findings in psychology and the social sciences which report wide-ranging controversies about cases. Now, it has to be conceded, of course, that whenever we find that other people have different intuitions we are well advised to follow Sidgwick's advice and take a sceptical attitude towards our own intuitive leanings. However, this does not mean that we should be sceptical about all our low-level intuitions. Some philosophers have argued – I believe rightly – that we should take empirical findings regarding intuitive disagreements on cases with a pinch of salt. The reason is that there seems to be a selection process at work. Neuro-psychologists like Joshua Greene purposefully select cases people tend to disagree on because they want to explain the differences in their judgements, e.g. regarding the differences in their brain activities (cf. Greene et al. 2001). The wide range of cases in which our low-level intuitions coincide is not as interesting and is not reported as frequently. Hence, when we read such studies, we get the impression that there is disagreement about almost each and every case. Bernard Gert makes the same point. He says that moral questions "such as whether it is morally acceptable to hurt someone simply because you dislike him are not controversial at all, but because they generate no discussion they tend to be forgotten." (Gert 2004, 14).

Let me add a second point which is surely worth stressing. When we survey the philosophical literature, we come across many cases that apparently exhibit strong disagreement. From this, too, we might conclude that there is probably much disagreement about cases in general and that, therefore, our low-level intuitions are to be doubted as well. However, this would, again, be an inference from a biased sample. Many cases in moral philosophy (and elsewhere in philosophy) have the purpose of drawing out the implications of competing theories and testing them against our intuitions (cf. Dennett 1984, 17–18). They often serve as the basis for *reductio* arguments. One theorist says to another: "Let us assume that your theory is correct. This would mean that in case X it would be right to do Y. But, surely, that view is absurd." To defend her theory against such an objection, the other theorist can either deny that it implies act Y in case X. Or she can simply embrace it and say: "I find doing Y in case X very reasonable." It is easy to explain, then, why there would be such a great deal of disagreement about cases in moral philosophy. The (reported) intuitions of theorists about cases differ, at least partly, because these intuitions have the function to attack competing doctrines and to corroborate one's own theoretical stance.

If this is in fact so, much of the disagreement about cases – at least amongst professional philosophers – is explained by their theoretical disputes over the right moral theory.[34] This, in turn, suggests that there should be a roughly proportional

[34]I am indebted to Martin Rechenauer who suggested to me that disagreement about cases may be seen as the embodiment of a more deeply rooted disagreement about moral principles. This point also seems to be acknowledged by Norcross (2008, 66) who says that he is "all too aware that

amount of disagreement about high-level intuitions since moral doctrines are high-level matters. If cases in moral philosophy do, in fact, mainly serve the purpose of putting a high-level principle to the test, there should be roughly one such principle for each case on which philosophers disagree. We can, of course, multiply these cases by specifying morally insignificant details differently. However, if they serve the same theoretical purpose, I would suggest to view them as essentially the same case.[35]

Now, let us turn to high-level intuitions on which there is allegedly comparatively little disagreement. This is simply not true. There is much disagreement. Take, e.g., Peter Singer's principle of harm prevention which we considered previously. He claims that "if it is in our power to prevent something very bad from happening, without thereby sacrificing anything morally significant, we ought, morally, to do it." (Singer 1972, 231) This principle implies that I ought to save a child who is drowning right before my eyes. However, it also entails that I ought to prevent a child from dying in some remote place in Africa if I can do this at roughly the same costs (e.g. by donating money). The principle contains no reference to physical distance – and rightly so, finds, e.g., Unger (1996). Frances Kamm famously disagreed with this and claimed that "at least intuitively, distance per se matters to what obligations we have." (Kamm 2007, 352) A further example is the high-level intuition that, we ought, *ceteris paribus*, to prevent a greater rather than a smaller harm if we cannot prevent both. E.g., if a flood is threatening the lives of people on both sides of an island and I am the captain of a freight ship who can save people on either side, but not on both sides, I ought, morally, to act so as to rescue more people and prevent the greater harm.[36] This, I take it, sounds plausible to most people. There is, however, no universal agreement about this case. Some philosophers have disputed it (e.g. Taurek 1977; Lübbe 2008). It appears, then, that the seemingly larger disagreement about low-level intuitions does not discredit them to a greater extent than high-level intuitions.[37]

---

non-consequentialists' intuitions diverge radically from [his] own" consequentialist intuitions and by Prinz (2010, 387) who remarks that "[p]hilosophers intuitions are not theory-neutral" which, as he hypothesizes, may be "one reason why philosophers seem to have different intuitions about the same cases."

[35]My view is corroborated by a remark by Shelly Kagan. He makes the point that "typically when we think about cases, we are only thinking about *kinds* of cases." (Kagan 2001, 61–62; emphasis in the original) This suggests that different specifications of a case structure can be seen as the same case. On the same point, see also Appiah (2008, 84–85).

[36]This example is taken from Taurek (1977).

[37]One might, however, draw a generally sceptical lesson from all of this and conclude that neither low-level nor high-level intuitions are reliable (cf. Singer 2005, 349). I find this hardly plausible. In many areas, our intuitions are very unreliable in isolation. But we would not conclude from this that we cannot make progress in these areas. Consider, e.g., probability theory. Many simple card tricks are able to fool us because our intuitions about probabilities are very unreliable. Nevertheless, human beings were able to develop a probability calculus based on intuitive considerations which, over time, got more and more formalized. And this probability calculus is very reliable, e.g., in

Above we said that empirical results concerning our moral intuitions and, notably, findings regarding their evolutionary genesis may cast doubt on them and, in particular, on low-level intuitions. What can we say in reply to this? There seem to be three strategies to defend low-level intuitions against these charges.

- Firstly, there is the strategy of blunt denial. We may say that the empirical findings which adherents of TD use to discredit low-level intuitions are just irrelevant in the context of moral theory.
- Secondly, the empirical results themselves can be challenged.
- Thirdly, empirical findings can be acknowledged, but the link between these findings and the conclusion drawn by proponents of TD can be disputed.[38]

Those who opt for the first strategy may defend their view by pointing towards Hume's (1888/1960) crucial distinction between *Is* and *Ought*.[39] They can claim that, as a matter of principle, it is not possible to draw conclusions for moral theory from factual evidence and that, therefore, the above considerations are fallacious.[40] This would, of course, be an argument against a straw man. No serious thinker would suppose that we can infer normative conclusions straightforwardly from empirical evidence (Mukerji 2015). Rather, those who claim that facts about our low-level intuitions ought to make us suspicious as to their credibility. This is a moral-epistemological claim which does not derive from any fact. It is quite a plausible claim, too! Facts about our moral psychology obviously matter. In this connection, John Rawls may serve as a crown witness. As we saw above, he thinks we should be wary of intuitive judgements made when we are upset, frightened, or stand to gain one way or the other because they are likely to be distorted. In saying this, he plainly acknowledges the relevance of empirical psychology to moral theory.[41]

---

predicting the frequency of future events. Since we are not concerned with the issue of scepticism we can put this issue aside. An instructive overview over sceptical positions on ethics can be found in Sinnott-Armstrong (2006).

[38]These three possible replies are inspired by Timmons (2008, 93) suggestions as to how a deontologist can defend herself against Greene's attack on their theory.

[39]Another possible strategy is to argue that the capacity for ethical intuition is an *a priori* faculty (cf. Lillehammer 2011, 176). It is very unlikely to work. So, for reasons of scope, we shall put it aside.

[40]It should be noted, however, that the precise purport of the Humean thesis is unclear. At least on some interpretations, it is clearly false, as Prior (1960) has shown. Consider a factual proposition E. From E we can derive $E \vee N$, where N is a normative proposition. There are only two possibilities. $E \vee N$ is a normative proposition. In that case, an *Ought*-proposition can be derived from an *Is*-proposition. Or it is factual. In that case, however, it is possible to derive N from $\neg E$ and $E \vee N$ which are factual propositions. So Ought-propositions can be derived from Is-propositions in any case. For a thoroughgoing treatment of the problem identified by Prior, see Schurz (1997).

[41]In recent times, the exponents of a new movement in Philosophy called "Experimental Philosophy" have vigorously defended the relevance of empirical data for philosophical theories (Knobe and Nichols 2008). On the relevance of psychological findings for moral philosophy, see also Driver and Loeb (2008), Greene (2008), and Prinz (2010).

The first strategy is off the table, then. The second strategy would fall under the purview of an empirical scientist. Therefore, it, too, is off the table – at least as far as our present inquiry is concerned. This leaves us with the third strategy, viz. to argue that low-level intuitions are not entirely discredited by empirical findings. How can this be done? Consider, first, framing effects. Does the fact that our low-level intuitions are subject to framing effects show that we should dismiss them *tout court* and trust only high-level intuitions? There are two reasons, I believe, why this would be a hasty conclusion to draw.

First of all, it has not been shown (nor do we have much reason to suspect) that all case-based intuitions are susceptible to these effects. Rather, it has been reported by some researchers that certain experiments could not demonstrate the existence of framing effects. Petrinovic and O'Neill (1996), e.g., failed to detect wording-related framing effects in some cases. To be sure, this does not demonstrate that there were no framing effects. However, it does give us reason to doubt the sweeping conclusion that all our low-level intuitions are susceptible to these effects.

Secondly, even if all low-intuitions were, in fact, affected by framing effects, this would not mean that we have to dismiss them *tout court*. Presumably, framing effects arise from the fact that, e.g., different wordings or different contexts draw our attention to particular features of it. This, in turn, may lead to a well-known problem, to wit, that we neglect other features which may be of equal importance (cf. Brink 1984, 117). If we know that we tend to have such "blind spots," as Bazerman and Tenbrunsel (2011) and Sorensen (1998, 273) call them, we can, it seems, discipline ourselves. We can try to focus on all relevant aspects of a moral problem and carefully consider our intuitive verdicts.

How do we answer the third challenge, viz. that low-level intuitive responses to cases are based on historically contingent emotions? The first line of defence that is possible to launch is to emphasize that the relationship between moral judgements and emotions allows of various interpretations. Even though it might be possible to show that certain emotions accompany certain intuitive judgements, this does not warrant the conclusion that emotions *cause* these judgements. This is "because it by no means follows when two phenomena accompany each other in their variations, that the one is cause and the other effect."[42] (Mill 1882, 496) There are various other possibilities (cf., e.g., Mukerji 2013a, 118–119). To justify the conclusion that variations in some empirical phenomenon x cause changes in another event y, it has to be ruled out, in particular, that

(i) the correlation between these variations is accidental,
(ii) variations in y cause variations in x (rather than *vice versa*),
(iii) variations in some other factor, z, cause variations in both x and y and
(iv) variations in x cause variations in y through some intermediary factor w.

Now, presumably, it can be empirically established that (i) is very unlikely at least when it comes to certain emotions and certain intuitions about cases. Let us

[42] This diagnosis is confirmed, e.g., by Huebner et al. (2009, 4).

assume that empirical scientists did their homework and that they took good care to rule out coincidence using standard methods of statistical testing.[43] However, based on a literature survey, not all of the other possibilities can be ruled out at this stage. There are models of moral reasoning which do not conclude that emotion drives intuitive moral judgements. Instead, they hold that (ii) is true (e.g. Dwyer 1999; Hauser 2008; Mikhail 2007). Scientists who subscribe to such models believe that the direction of causality goes from moral intuition to emotion. According to them, the fact that we feel a certain way about a particular action (e.g. the fact that we are repulsed by the idea of pushing the fat man off the bridge) can be explained by the fact that we have a certain intuition about the wrongness of this act. As reported by Huebner et al. (2009), models of the Piaget/Kohlberg tradition assume what they regard as a Kantian picture of moral judgement. Kant thought, as is well known, that reason gives rise both to the rational emotion of "reverence" for the moral law and the particular moral judgements about cases. Models which adopt this picture support the alternative explanation (iii). They hold, that is, that there is a third factor whose workings determine both variations in emotions and moral intuitions.[44] As I said above, we cannot assess how the respective models, in fact, stand up to empirical evidence. This is a primarily scientific and not a philosophical issue. Scientists have to work out which account is adequate. However, until there is no significant agreement on the issue, we should not make the mistake and listen to just one side of the debate. Hence, we should not jump to the conclusion that our contingently evolved emotional responses to individual cases drive our low-level intuitions, thereby discrediting them.

What is more, it would not even follow that we have to mistrust all our low-level intuitions, even if it did turn out that they are all driven by emotions. To be sure, in many instances the fact that emotion drives an intuition should make us cautious. However, the reason for this seems to lie in the fact that strong emotions have a particular kind of effect. They "cloud" our judgement, one might say.[45] Professional philosophers who have talked with laypeople about moral-philosophical issues can surely confirm this. Sometimes when we ask them to imagine certain abhorrent cases, e.g. cases involving footbridges and fat men, our interlocutors may find the notion of doing a particular act so repulsive that this blinds them to other important factors about the case. Even if this is true, though, it does not follow that we can make the hasty generalization that emotion clouds all our low-level intuitions and that the latter are full of blind spots (cf. Sinnott-Armstrong 2006, 194). To avoid this

[43] Berker (2009), however, has questioned the research reported by Greene (2008) in that way. For a rejoinder, see Greene (2010).

[44] It is, admittedly, quite a stretch to associate this view of moral reasoning with Kant since Kant's moral system is purely based on reason and does not allow moral intuition any role to play (cf. Kant 1785).

[45] For this reason, Sinnott-Armstrong (2006, 194) moots the principle that we need an independent reason to believe intuitions that we have in situations where we are "emotional in a way that clouds judgment." This caveat is important. The principle advises caution only when it comes to emotions which cloud our judgement. And these might not be all emotions.

effect, it seems we just need to make sure that our emotions do not get the better of us and lead us to a judgement that is too brisk. Contrary to inclination, we must ensure that we do not only consider a particular aspect of the case but examine it for all factors which, on reflection, ought to be seen as relevant. If we do this, I see no reason not to trust our low-level intuitions, even if it should turn out that they are laced with emotion. As Tersman (2008) notes, there might even be a reason to think that certain types of emotional involvement might even improve our intuitive verdicts. He argues, e.g., that a "well-founded evaluation of a moral dilemma usually requires information about which interests are at stake, and in order to gather such information it may help if we are capable of some amount of empathy." (Tersman 2008, 393)

As a final note, it may be mentioned that Tersman (2008) also points out that explanations for the genesis of our high-level intuitions are also available. He says that "[a]lready from the start, Christian ethics involved the belief that many differences that had previously been regarded as morally relevant, such as ethnicity or differences in class, are not in fact so." (Tersman 2008, 401) This might explain why Westerners whose societies are coined by a Christian tradition find it so intuitive that all morally relevant subjects deserve the same moral consideration. This is one of the high-level intuitions on which, e.g., Henry Sidgwick bases his utilitarian theory. The same holds for Peter Singer's philosophy. Now, if we can generally regard genetic explanations as casting doubt on our intuitions, they would certainly cast doubt on high-level intuitions, too. It is important to note that we can interpret this reasoning in two ways. One way of interpreting it is as a *Tu Quoque*. In that case, it would not lend support to low-level intuitions. However, coming from a proponent of RE this is not how we should make sense of it. It seems we should rather interpret it as a companionship-in-guilt argument (cf. Mackie 1977, 39). Those who believe in the RE approach believe that high-level intuitions *can* be reliable. When they point out that high-level intuitions may be shaped by tradition, they do not mean to claim, therefore, that this makes them *ipso facto* unreliable. Rather, assuming that high-level intuitions are reliable, they want to point out that proponents of TD are inconsistent when they criticize our low-level intuitions based on their causal history. After all, we could make the same point about the high-level intuitions whose reliability they leave unquestioned. In other words, high-level and low-level intuitions are "companions in guilt." There is no reason to be sceptical about low-level intuitions in particular.

2.2.3 The Bottom-Up Approach

Let us take stock of where we are. In Sect. 2.2.1, we considered the case for TD. We looked at some of the reasons why one might think that a moral theory ought to fit only our high-level intuitions. In Sect. 2.2.2, then, we examined how one could make a case for RE. Proponents of RE, such as John Rawls, think that a moral theory ought to fit our intuitions at both ends. It ought to fit, that is, the relevant intuitions

of high and low degrees of generality and abstractness. Such theorists have to argue that there is no reason to mistrust all our low-level intuitions and to generally give preference to high-level intuitions. As we saw, they can make a persuasive case. For it seems that the arguments presented by proponents of TD, who attack the credibility of low-level intuitions, are rather shaky. Now, supporters of RE reply to the criticisms of the adherents of TD in a rather defensive way. They only seek to establish that there is no reason to discard all low-level intuitions and that moral theories should be tested against them, too. They do not claim, as we just saw, that high-level intuitions are altogether unreliable. For they believe that at least certain high-level intuitions do possess initial credibility. This is where champions of the BU approach come in. They share with those who favour RE the view that we should regard at least some low-level intuitions as initially credible. So they can adopt the case that proponents of RE make in defence of low-level intuitions. They merely need to add to it a criticism of high-level intuitions which shows that the latter do not, in fact, possess initial credibility.

As I said above, nothing in our argument depends on BU being correct. For the purpose of our inquiry, it is sufficient to show that we should reject TD because at least some low-level intuitions possess initial credibility. Both proponents of RE and BU hold this view. Therefore, it is, in fact, unnecessary for us to argue for BU and against RE. Nevertheless, let us, for the sake of sportsmanship, quickly point out why BU might be plausible.

It seems that supporters of BU could say something about high-level moral principles which is similar to what David Hume said about abstract ideas (cf. Hume 1888/1960, 25–33). As an empiricist, Hume believed that all ideas are derived from prior sense impressions. Since every impression is an impression of a concrete object, all ideas, he thought, had to be concrete as well. Thus, Hume reasoned, when we appear to think abstractly, we actually have a concrete idea in mind which we then allow to relate to other objects that are sufficiently similar in its qualities. Something like this may be going on when we think of an abstract principle and form an intuition about it. It may be that we do not consider it in its abstractness, but imagine concrete cases to which it applies and then say "yes" or "no" to it depending on whether its implications in these cases seem intuitively acceptable. It may be, that is, that whenever we think we have a high-level intuition about a principle, we have, in fact, one or more (muddled) low-level intuitions about the anticipated implications of the principle in particular cases that we happen to think up. In support of this thesis, one could, e.g., cite Amartya Sen, who said something remarkable in the context of social choice theory. Social choice theory offers an axiomatic take on moral problems. Most of what happens in it happens on a rather abstract plain, where theorists give much attention to the credibility of the axioms which are more or less formalized versions of high-level moral principles.[46]

[46]Such comparisons are necessary, in particular, when it comes to "impossibility results" (e.g. Arrow 1951/1963), which show that certain axioms cannot logically co-exist. In that case, the theorist has to drop at least one to ensure consistency.

It seems, then, that social choice theory is a paradigm example of TD and that those who practice it should believe that the initial credibility of the axioms plays a great role. Now, curiously, Sen has claimed that "[w]hen we say 'yes' to an axiom we do not think absolutely abstractly. We think of actual cases." (Sen 2009) It seems we can interpret this in the way I just suggested, viz. as saying that the justification of a high-level axiom depends entirely on whether or not its low-level implications intuitively make sense. This, in turn, would suggest that there are, in fact, no high-level intuitions about moral principles. It would mean that they are mere chimeras and should play no role in moral inquiry.

Having said this, allow me, briefly, to draw out what it would mean for the evaluation of moral theories if we adopted the BU approach. It may seem that, on BU, we would always have to talk about cases and would have to eschew any mention of general principles. But this is not so. BU does not suggest that philosophers should entirely disregard the intuitive appeal of principles. It would still allow us to endorse or reject moral theories in light of their compatibility or incompatibility with abstract tenets that we find intuitively plausible. However, it would remind us that when we do this, we must not forget that the intuitive appeal of principles derives from the intuitiveness of its case implications. Since many principles apply potentially to an infinite amount of cases, this suggests that we must always consider the possibility that the intuitive appeal of a principle may, on reflection, turn out to be smaller than it initially appeared. We may discover that a seemingly plausible principle has very counter-intuitive implications in particular circumstances. And this may completely destroy its credentials from the standpoint of intuitive fit on the BU interpretation.

As a final note, it should be stressed that, even if BU is accepted, it does not follow that one should reject high-level principles whenever their implications contradict low-level intuitions. As we worked out above, this is because intuitive fit is merely a *sub*-criterion of the Rawlsian Approach. Within that approach, systematicity may play a great role, too. There can, hence, be a trade-off. Even if principles leave something to be desired regarding their low-level intuitive fit, we may still accept them due to their systematizing strength.

2.3 Provisional Fixed Points

Above we factorized the Rawlsian Approach to theory evaluation into distinct sub-criteria. And we discussed various interpretations of it. Now we need to consider how we can use the approach to develop a workable method for our evaluation of consequentialism. This is necessary since nothing that we have said so far can be straightforwardly applied. This may not be obvious. So let me explain.

The Rawlsian Approach does not, as it were, provide a "pass-or-fail test" for moral doctrines. Hence, we cannot directly apply it to assess consequentialism. The approach offers, rather, a "philosophical ideal" which, presumably, none of our moral doctrines can achieve. We should not, therefore, dismiss a theory, if it does

not attain a perfect fit with the relevant moral intuitions. We should, rather, assess it regarding whether or not it "moves us closer to the philosophical ideal" (Rawls 1971/1999, 43) than its alternatives. However, that would mean that we have to examine not only consequentialism but also its main competitors in what could be called a *comparative study* (cf. Sinnott-Armstrong 2011). This is something which we cannot do here. Given the scope of this inquiry, there is simply no way that we can compare consequentialism even with its most prominent rivals.

There are, however, ways to circumnavigate this problem. We can look for decisive tests that follow from the Rawlsian Approach. As we said above, consistency which is one of its sub-criteria possesses the status of a knock-out criterion. Hence, on the assumption that at least some moral doctrines are actually consistent, we can reject those which are not. For, according to the Rawlsian Approach, they will certainly be inferior to any consistent moral doctrine. That means, if we could show that all consequentialist theories are, indeed, inconsistent, we could conclude that they fail. Certain theorists have, in fact, discussed whether this line of argument can be successful. Some of them have focused on the issue of "complex acts."[47] (e.g. Bergström 1966, Bykvist 2002; Castaneda 1968; Carlson 1999a) Others have brought up charges of self-defeat. They have argued that consequentialist agents fail to achieve aims that are deemed desirable by the lights of their own moral theory. This strategy has been applied, e.g., by Hodgson (1967) and Nida-Rümelin (1993). I shall propose, however, to put it aside here.[48]

Can we try to use coherence? I believe that this would be a bad idea. Such an approach would probably not give us enough to chew on. As some philosophers have pointed out, consequentialist theories are, in fact, rather "unlikely to encounter problems of coherence." (Sumner 1987, 173)

The strategy that seems to fit our present purpose best is based on the sub-criterion of intuitive fit. It uses what John Rawls calls "provisional fixed points" for moral theorizing.[49] I shall, therefore, refer to it as the Provisional Fixed Point Approach (PFPA). The idea behind it is as follows. When we assess a moral theory, we look for intuitive convictions which possess a high degree of initial credibility. They have to be so strong that it seems very reasonable to expect that an acceptable moral theory should fit them. Then, we check whether the doctrine in question does, in fact, match these intuitive judgements. If not, we reject it, no matter how coherent it seems and irrespective of how intuitive it is in other regards.

In the context of our discussion, this approach seems appealing for two reasons. Firstly, we do not need to conduct a comparative study. We do not have to consider the merits and demerits of consequentialism in comparison to its alternatives. PFPA

[47] Elsewhere, I have briefly discussed this strategy (cf. Mukerji 2013c, 306).

[48] It can be argued that at least the second strategy runs into difficulties when certain types of agent-relative consequentialist theories are taken into consideration (cf. Mukerji 2013a, 114–117).

[49] It seems that PFPA is implicitly recognized in many moral-philosophical tracts. In addition, there is a number of authors who have emphasized that the approach plays a great role in the practical application of the Rawlsian Approach. See, e.g., Daniels (1996, 28), Mulgan (2007, 58), Nida-Rümelin (2002, 34–35), Otsuka (2006, 110) and Rawls (1971/1999, 18).

allows us to devote our full attention to the object of our inquiry. Secondly, the approach homes in on what seems to be the most important aspect of the debate about consequentialism. Most critical studies have focused on intuitive fit and have attempted to demonstrate that consequentialism is unacceptably counter-intuitive.

It should be noted, however, that these advantages come at a cost. First of all, since PFPA is based exclusively on one sub-criterion of the Rawlsian Approach, it will miss objections that draw on the other evaluative criteria. Those may turn out to be crucial. Secondly, the assumption assumes that a consistent moral theory can, in fact, fit the respective provisional fixed points. Hence, we have to qualify the conclusions we draw from it with a *proviso*: Should it turn out that it is, in fact, impossible to match the respective provisional fixed points, we have to revoke our verdict.[50] Thirdly, PFPA can only tell us whether we have sufficient reason to *reject* a given moral doctrine. However, it cannot tell us whether we have sufficient reason to *accept* it. It is easy to see why. If we find that a given moral theory violates certain provisional fixed points, we can judge that it ought to be rejected (under the mentioned *proviso*, of course). If we find that a moral theory fits all our provisional fixed points, we cannot judge, however, that it ought, therefore, to be accepted. It might still be possible that the doctrine is, in fact, untenable. All we can say, based on PFPA, is that we do not have reason to think so.

Before we move on, let me briefly address these worries. The first problem naturally arises for any in-depth investigation of a philosophical problem. Granted, in using PFPA, we may lose sight of certain issues which are undoubtedly important in their own right. This seems defensible, however, because we have to confine the scope of the inquiry to ensure its tractability. In reply to the second problem, we can give a similar answer. It is true that, in using PFPA, we do rely on the assumption that it is, in fact, logically possible for a moral doctrine to fit the respective fixed points in question. We cannot ensure that this assumption is justified – at least not within the scope of the present inquiry. However, every philosophical investigation has to take certain things for granted. It cannot address all problems at once.[51] The third problem, I believe, is one we can indeed ignore, given the rather modest aim of the investigation. We are merely interested in developing a case *against* consequentialism. We are not seeking to investigate whether a constructive case *for* consequentialism is possible. With this in mind, it seems that PFPA is adequate for the purpose at hand.

Now that we have, I hope, a clear enough idea about PFPA in the abstract, we should specify it further. In particular, we should answer the question where we

[50] Social choice theory has shown that weak seeming moral judgements may turn out to be logically incompatible. An example which illustrates this is an impossibility theorem proved by Sen (1970b). It is called the "Impossibility of the Paretian Liberal" and shows that a minimal notion of individual rights is incompatible with the Weak Pareto Principle.

[51] In addition, it might be mentioned that the possibility of inconsistency seems to be confined to abstract level fixed points. Fixed points about cases cannot be inconsistent unless they concern the same case. As will become clear in Sect. 6.1, our argument relies entirely on fixed points about cases.

may find provisional fixed points. This gives us a chance to tie up loose ends and to relate what we just said with the points we made in the previous section. Obviously, the answer depends on the particular version of intuitive fit that we accept. As we discussed above, the TD approach holds that the only initially credible intuitions can be found at the high level. Accordingly, a proponent of TD must maintain that the only intuitions suited to figure as provisional fixed points lie at the high level. Theorists who accept RE believe that there may be provisional fixed points at both the high and low level while those who support BU think that they can be found only at the low level. So, in principle, provisional fixed points might be found anywhere, depending on the favoured interpretation of intuitive fit. That is, PFPA can be combined with TD, RE, and BU. Recall the above, however. We made a case against TD. If this case is accepted, we can assume that there are provisional fixed points to be found at the low level, too, as RE and BU purport. In fact, our case against consequentialism will turn entirely on low-level provisional fixed points.

Before we proceed, allow me a brief note of clarification. Some may object to our commitment to PFPA because we premised it on a particular moral-epistemological position. To explain, it is common in general epistemology as well as moral epistemology to distinguish between foundationalist and coherentist approaches to justification. And it may be alleged that our approach falls on the wrong side of this distinction. Such criticism, I think, would be misjudged. Though PFPA, as we have so far characterized it, does fall on the coherentist side, nothing that we will say below depends on an endorsement of coherentism since minor adjustments would allow us to transform PFPA into a foundationalist procedure. Let me explain.

First up, what is the essential difference between foundationalism and coherentism? The former view, I take it, assumes that all of our convictions are justified to the extent that they are either self-justifying or derivable from a self-justifying belief.[52] In contrast, the latter view is based on the idea that justification is a matter of mutual support. We cannot have self-justifying and irreversible beliefs. Rather, our beliefs are justified if and only if they fit into a web of convictions which possesses the highest possible degree of credibility *overall*. Note that PFPA does not assume that any of the intuitive convictions we use are irreversible. That is why it is called the *provisional* fixed point approach. So it falls on the side of coherentism. It is not hard, however, to transform it into a foundationalist methodology. To do that, we simply have to assume that the intuitive judgements we use to test consequentialism are not *provisional* fixed points, but *properly* fixed points. Such a modified version of PFPA (a Fixed Point Approach or FPA, for short) is defended, e.g., by Judith Jarvis Thomson. She says that she accepts PFPA

[52]Foundationalists who are non-sceptics believe, in addition to that, that there are self-justifying moral beliefs. This assumption is necessary to avoid scepticism. It is easy to see why. The foundationalist criterion of justification is perfectly compatible with the sceptical view that no moral belief fulfils it because there might, after all, be no self-justifying moral beliefs.

> with this proviso: on Rawls' account of the matter, everything is provisional, everything is open to revision, whereas I am suggesting that some moral judgements are plausibly viewed as necessary truths and hence not open to revision. (Thomson 1990, 32)

With such a *proviso* in place, our moral-epistemological approach would fall on the foundationalist side. I believe that it is not necessary to engage in any debate here. Both foundationalists and coherentists can accept our case against consequentialism based on the approach outlined above, as long as they find the intuitive judgements that we use in the argument acceptable. The only difference lies in their respective interpretations of these verdicts. Foundationalists may regard them as self-justifying moral views, while coherentists will see them merely as statements that possess a high degree of initial credibility. Who is right about this issue? In the context of our present discussion, this is largely a moot question. For this reason, we can safely put it aside.

2.4 Trolley Cases

In the previous sections, we established that, on the Rawlsian Approach, moral theories are evaluated, at least partly, in terms of their fit with our moral intuitions. There are various interpretations of this evaluative criterion: TD, RE, and BU. We argued that RE and BU, which maintain that moral theories should be evaluated, at least partly, in terms of how well they fit our low-level intuitions about cases, are the most plausible interpretations of the evaluative criterion of intuitive fit. In the previous section, then, we showed how intuitive fit can be translated into a workable, methodic approach, viz. PFPA. PFPA instructs us to look towards cases that elicit strong intuitive convictions, such that it seems reasonable to suppose that any moral theory which contradicts these intuitions seems faulty. At this point, then, it remains to be explained which kinds of cases we will use to set up our case against consequentialism and why.

There are, broadly speaking, two possibilities. We could use realistic cases – cases, that is, which have occurred in real life or are at least quite likely actually to happen. The second option is to use hypothetical scenarios which are very unlikely ever to arise in practice. As many philosophers before, we will opt for the latter. That is, we will use counterfactual and unrealistic cases which commonly go by the name "trolley cases." In what follows, we shall make some preliminary remarks about this particular sort of case. First of all, we will talk about their distinctive characteristics. The most natural way to introduce them is, I think, to look at a typical trolley case and to abstract the respective features from it. So that is what we will do. After that, we will consider two different uses for trolley cases in a moral-philosophical investigation. And we will point out how we will use them. Finally, we will explain why trolley cases seem to be particularly helpful, given the purpose at hand, before we address some worries that critics may raise about them.

2.4.1 Characteristics

Trolley cases involve a story about an agent facing a morally significant choice. This story usually revolves around a runaway trolley – hence the name – which is threatening to do some serious harm to some unlucky people. A typical example is the following scenario due to Judith Jarvis Thomson.

> **Edward's Case**
> Edward is the driver of a trolley, whose brakes have just failed. On the track ahead of him are five people; the banks are so steep that they will not be able to get off the track in time. The track has a spur leading off to the right, and Edward can turn the trolley onto it. Unfortunately there is one person on the right-hand track. Edward can turn the trolley, killing the one; or he can refrain from turning the trolley, killing the five. (Thomson 1976, 206)

Of course, the presence of a trolley is not what makes this case a trolley case. This particular detail serves merely to make for a colourful illustration of a choice situation which possesses certain distinctive features. It is these features rather than the particular story to which they are tied that make a case a trolley case.[53] We can discover them if we pay close attention to the details of *Edward's Case*.

The first thing to note is that the case strikes us, I presume, as a *tragic choice*. The above description does not state this explicitly. However, it is the most natural interpretation. And it is surely the one that is intended. We do not suppose, e.g., that the five men on the main track are "old and suicidal" and that "they'd gathered on the tracks to end their lives." (Appiah 2008, 97) We assume, quite naturally, that each person's life is valuable. And we recognize that there is no way for Edward to avoid ending at least one of these valuable lives. This is what makes his choice tragic.[54]

The second important aspect of the case is that Edward only has two options for acting. He can either do nothing or turn the trolley to the right and onto the spur.

The third feature worth highlighting about *Edward's Case* is the assumption, albeit implicit, that all normative factors that the description does not explicitly mention are absent. Further information, particularly information about the six people on the tracks, might conceivably make a difference in this situation. We assume, however, that there is no such information. For clarity's sake, let us specify what this means. The only facts that matter from the moral point of view in *Edward's Case* are the (relevant descriptions of the) acts that are available to Edward as well as their consequences. The latter, in turn, are fully described by the number of deaths that each option, respectively, will cause. All further factors that might conceivably matter are assumed to be out of the picture (cf. Wood 2011, 73–74). We

[53] We follow a terminological suggestion by Wood (2011) and Fried (2012).

[54] Note that the idea of a tragic choice is often mixed up with that of a moral dilemma. A moral dilemma is (i) a situation in which the agent is confronted with a choice between a number of options all of which are wrong or (ii) a situation in which at least two mutually exclusive options for acting are obligatory (cf. Vallentyne 1989). A tragic choice is "merely" a choice between options that are all bad.

might, e.g., suspect "that a trolley driver is a professional who is plausibly specially responsible for the trajectory of their trolley." (Mendola 2005a, 82) Alternatively, we might conjecture that one of the workers is, say, Edward's brother or friend, such that personal loyalties become relevant. We might also hypothesize that Edward is, perhaps, especially indebted to one of the workers. Though all these considerations might be morally relevant, they can be assumed to be out of the picture in *Edward's Case* because there is no explicit mention of them. I should stress that this holds, in particular, for any historical factors that might play a role. It may matter, e.g., that "one or more of the six potential victims is at fault for the coming about of the situation they now face." (Thomson 2008, 361) However, we are supposed not to make any such assumption about the history of the case.

The fourth feature about *Edward's Case* that is worth stressing is the assumption that the agent's act uniquely determines the outcome. We stipulate that, if Edward does nothing, five workers will get killed. If he steers the trolley to the right and onto the sidetrack, one person dies. There are no contingencies.

Finally, there is a fifth implicit characteristic.[55] It concerns the epistemic situation of the agent. Edward is supposed to know all of the empirical facts that the description of the case mentions. That is, he is expected to know all of his options for acting and all of their consequences.

In the remainder, we shall assume that trolley cases generally have the characteristics we just highlighted in *Edward's Case*. In brief, they can be stated thus:

Characteristic 1 (Tragic Choice)
The agent faces a tragic choice. No matter what she does, at least one person will suffer severe harm (commonly death).
Characteristic 2 (Limited Options)
The agent has a definite, limited range of options for acting.
Characteristic 3 (Absence of Normative Factors)
There are no morally relevant facts except for those explicitly mentioned. These are the options available to the agent and their respective consequences (e.g. the number of deaths).
Characteristic 4 (Determinism)
What the agent does uniquely determines the outcome of the case.
Characteristic 5 (Omniscience)
The agent knows all facts that the description of the case states.

As we noted above, trolley cases are different from real-life cases in that they are inherently unrealistic. Obviously, this has to do with the above assumptions. It may be worth noting, however, that there is a difference in kind between them (cf. Shue 2006, 231). Characteristic 1 is merely an assumption about the nature of the case. A trolley case is always tragic. This, one may say, makes it somewhat unrealistic in the sense that most cases we confront in ordinary life are not that way. Characteristics 2, 3, 4, and 5 also make trolley cases unrealistic, but in a different sense. These assumptions are never satisfied in a real case. Characteristics 2, 3, and

[55]This feature of trolley cases is discussed, however, in an exchange between Gert (1993) and Thomson (1993). See, also, Fried (2012, 2), Rosebury (1995, 499), and Wood (2011, 70) who state it explicitly.

4 are *abstractions* from the intricacies and complexities of real life. By stipulating that trolley cases possess these features, we assume away, as it were, certain morally relevant aspects of ordinary cases (cf. Gigerenzer 2008, 11; Wood 2011, 69). We assume that the agent has only very few options for acting, although it is clear that moral agents always have many options. We assume that facts which typically matter are out of the picture, although it is plain that in real-life there would always be many considerations that might be morally significant. Moreover, we assume that the agent's choice uniquely determines the outcome of the case, although there are always many factors we have to take into account in real life. Characteristic 5 is an *idealization*. We stipulate that the agent knows all relevant facts about the case, although it is clear that real-life actors never possess all the relevant information.

2.4.2 Uses

With the characteristics of trolley cases in mind, let us briefly consider two different uses for trolley cases in a moral-philosophical investigation.[56] To this end, it is useful to introduce a further trolley case. Consider the following scenario that should sound familiar.[57]

> **George's Case**
> George is on a footbridge over the trolley tracks. He knows trolleys, and can see that the one approaching the bridge is out of control. On the track back of the bridge there are five people; the banks are so steep that they will not be able to get off the track in time. George knows that the only way to stop an out-of-control trolley is to drop a very heavy weight into its path. But the only available, sufficiently heavy weight is a fat man, also watching the trolley from the footbridge. George can shove the fat man onto the track in the path of the trolley, killing the fat man; or he can refrain from doing this, letting the five die. (Thomson 1976, 207–208)

Edward's Case and *George's Case* are similar. In each case, the respective agent has two options for acting. And in each case, one of these options leads to the death of five people, while the other leads to the death of only one person. I assume, however, that our intuitions as to the permissibility of the agent's choices differ between the two cases. At any rate, most people believe that it is at least morally permissible for Edward to kill one instead of five. A majority, however, feels that it is impermissible for George to kill the fat man by pushing him off the bridge.[58] Trolleyologists (as philosophers who deal in trolley problems are sometimes called) have commonly used these facts about our intuitions regarding these cases "for

[56] The distinction we use corresponds to Karl Popper's distinction between the different uses of thought experiments in science (and especially in quantum theory), viz. the *apologetic* use and the *critical* use (cf. Popper 1959/2005, 464–480).

[57] We have already come across this scenario on page 22.

[58] These conjectures were made by Thomson (1985). In the meantime, a lot of empirical evidence has been piled up to support them. Important sources can be found in Greene (2008, 42).

the purpose of unearthing principles of permissible harm." (Kamm 2007, 4)[59] The idea is to go through a series of such cases, to consider our intuitive responses to find provisional fixed points, and to formulate moral principles and theories which capture these fixed points. The exercise is analogous to that of an empirical scientist fitting a curve to her data points.

This, however, is not the only possible use of trolley cases. Philosophers also employ them with critical intent, that is, to test moral theories (cf., e.g., Tännsjö 2011, 295). Here, the idea is to look at a given theory and to check whether it matches provisional fixed points in a specific case or series of cases. We may, e.g., test theories regarding their implications in *Edward's Case* and *George's Case*. That is, we may reject all doctrines that do not imply that Edward should steer the trolley to the right, killing the one. And we may reject all doctrines that do imply that George should push the fat man, saving the five. The analogue to this second use of trolleyology is the case of the empirical scientist who critically tests a theory by examining whether it does, in fact, capture all available data points.[60]

The distinction between these two uses is important because some worries about trolley cases appear to relate solely to the first one. It should be noted, then, that we will use trolley cases only in the latter way. That is, we will use them only to test consequentialism. We will not argue for an alternative moral theory.

2.4.3 Pros and Cons

With the various features and uses of trolley cases in plain view, we can address some of the pros and cons of trolleyology, starting with the pro side. There are two main reasons for using trolley cases rather than more realistic scenarios. These relate to their aforementioned features. One is specific to our investigation. The other is more general.

The first reason lies in the desire to reduce complexity. As will get clear below, one of the difficulties about any study of consequentialism is the fact that there are so many versions of it. Given the scope of this inquiry, going through all of them is an unmanageable task. Trolley cases, however, allow us shortcuts. By using them, it is possible to set aside certain varieties of consequentialism *ab ovo*. Here is why. The differences between the various consequentialist doctrines manifest only

[59] See, also, Wood (2011, 67).

[60] Of course, the scientist need not immediately reject the theory if it turns out that it does not capture all data points. There are always ways of accounting for recalcitrant evidence which are compatible with the truth of the theory (cf. Lakatos 1970). Similarly, a moral theorist need not reject a moral principle if it violates one or more out of a number of provisional fixed points. As we explained on page 40, PFPA is subject to a *proviso*. Should it turn out that no moral theory can, in fact, accommodate all of our provisional fixed points, the fact that a given theory violates one of these points cannot, by itself, count as counter-evidence against it.

in particular kinds of cases. Depending on the case at issue, it may, therefore, be unmotivated to distinguish between certain varieties of consequentialism.

An example should help to drive home the point. It is common, e.g., to differentiate between *subjective* and *objective* versions of consequentialism (cf., e.g., Howard-Snyder 1997). The distinction is, roughly, this. On Subjective Consequentialism, the moral status of an act is determined by the consequences that the agent expects. In contrast, Objective Consequentialism turns on the actual result of the agent's choice. To resolve whether her act is right or wrong, it looks towards objective consequences. Note, then, that the difference between these two versions of consequentialism is only relevant in cases where the agent's epistemic situation is imperfect. We have to assume that she cannot know for sure what the consequences of her act will be so that subjective and objective results can, in fact, come apart. If, on the other hand, the agent knows for sure what will happen if she chooses this or that act, subjective and objective consequences will coincide and so will the moral verdicts of Subjective and Objective Consequentialism. Now, trolley cases make, as we know, an idealized assumption about the agent's epistemic situation. She is supposed to have perfect knowledge of all morally relevant facts of the case (Characteristic 5). This, of course, includes the objective consequences of her options. Since she knows this, subjective and objective outcomes coincide, and so do the verdicts of Subjective and Objective Consequentialism. This *must* be the case! Hence, it eliminates the motivation for distinguishing between these two variants of consequentialism. This is, of course, only an example. As we will see in more detail below, trolley cases do not only take away the motivation for a distinction between Subjective and Objective Consequentialism. They also make superfluous the difference between Direct and Indirect Consequentialism and between consequentialist doctrines that subscribe to different theories of individual well-being (e.g. Welfarism Hedonism, Welfare Preferentism, and so on). This will help to reduce the workload considerably.

The second, more general reason why trolley cases seem useful is that they allow us to clarify our intuitions. This has to do with their simple make-up and their comparatively little complexity. Realistic cases, in contrast, can be very fuzzy. There are many options, a lot of normative factors to consider, and other relevant aspects that do not play a role in trolley cases. Under these complexities, our intuitions may give out (cf. Nagel 1986, 180). Moreover, even if we do have an intuition about a case, it is unclear whether it is reliable. For it may be, as we observed previously, that we fall prey to moral "blind spots." That is, we may end up paying too much attention to certain factors, while ignoring others. This, it seems, is not as likely to happen in a trolley case. Here, the relevant facts are reduced to a minimum such that we may assume that anyone can handle the cognitive load.[61]

[61] A further consideration that might be brought up to motivate the use of trolley cases is given by Amartya Sen. He writes that "in many of the common cases, intuitions based on quite different principles tend to run in the same direction, so that it is impossible to be sure of the basis of an overall judgment." (Sen 1982, 14) Therefore, it may be hard for a moral theorist to establish that her favoured theory is the best explanation for our moral intuitions if only common cases are used.

So much for the plus side. It is evident, however, that the advantages of the methodic use of trolley cases (trolleyology, henceforth) come at a cost. Trolley cases have a very simple, idiosyncratic makeup. As a consequence, certain types of ethical problems cannot be addressed in a trolleyological investigation. Take, e.g., the ethics of risk and uncertainty. It is undoubtedly an important theme in moral theory that consequentialists have had interesting things to say about (e.g. Norcross 1998). Now, as we discussed above, trolley cases assume that the decision of the agent necessitates a given outcome (Characteristic 4). Hence, they cannot be used to address moral issues that may arise in the context of risk and uncertainty. By using trolley cases, we will, therefore, inevitably miss important aspects of the moral-philosophical debate that pertain to these phenomena. Philosophers who are especially interested in discussing them may, therefore, regard trolleyology as a flawed method. I believe, however, that it is possible to address their reservations. To do this, we should remind them that we use trolley cases with a specific purpose in mind. We seek to construct an argument that makes plausible the claim that all forms of consequentialism should be rejected. To do this, we do not have to address all ethical issues on which consequentialism may have something to say. All we need to do is to demonstrate that there is at least one serious objection to all forms of consequentialism. If we succeed in doing this by using trolley cases, we may skip over many interesting philosophical questions. However, we will, nevertheless, accomplish what we set out to do.

This said, we should turn to some objections that seem to be more fundamental and more severe. Before we do that, however, allow me to express my discontent with the current state of the debate. The use of hypothetical cases in ethics and trolley cases, in particular, has "become so common that many philosophers hardly notice it and if they do, find it unproblematic." (Elster 2011, 241–242) In fact, the principal exponents of trolleyology usually do not bother to justify their methodology properly. What they say about it hardly ever surpasses the stage of mere explanation, even though objections to trolleyological thinking have been piling up for years. This is a lamentable fact. A systematic and comprehensive investigation of the virtues and limitations of trolleyology is surely in order. However, given the limited scope of this inquiry, it is not a task we can take on here. Nevertheless, we shall try, at least, to make our trolleyological method plausible.

This having said, let us turn to the objections to trolleyology. First up, we should demarcate two sorts of scepticism about it. One kind of worry has to do with the fact that the methodology relies on our low-level intuitions about cases. In Sect. 2.2, we addressed this concern at some length. We concluded that, to the extent that we can trust our intuitions at all, we do not seem to have any reason to distrust low-level intuitions in particular. Hence, we shall set this particular worry aside.

If our intuitions about these cases can be explained by a large number of normative factors, the moral theorist will have a hard time arguing that *her* explanation should be chosen. "In order to do the discrimination," Sen says, "we choose examples such that different principles (. . .) push us in different directions." (Sen 1982, 14) Sen's reasoning, I believe, provides a good motivation for the constructive use of trolley cases, but is less relevant to our destructive use.

Instead, we shall focus on objections that philosophers voice who are otherwise sympathetic to the idea that our intuitions regarding cases can be valid but insist that the particular features of trolley cases disqualify them for the purpose of moral inquiry. These objections mark, as it were, a "family quarrel" (Elster 2011, 242) between philosophers with similar epistemological inclinations (either towards BU or RE).

Objection 1
People disagree about trolley cases

The first objection that we shall address relates to one of the points that we made in our general discussion about the reliability of intuitions.[62] Recall the quote by Henry Sidgwick that we came across above. Sidgwick says that "if I find any of my judgments, intuitive or inferential, in direct conflict with a judgment of some other mind, there must be error somewhere." (Sidgwick 1907, 342) This, in turn, should reduce the confidence in my judgement. Obviously, the same goes for intuitions about trolley cases. If there is genuine disagreement about them, we should be cautious. Now, it may be suggested that the empirics of trolleyology show that people do disagree about trolley cases. This may give rise to something like the following argument.

(P1) If people disagree in their intuitive judgements about a case, this makes everyone's intuitions about that case initially incredible.
(P2) People disagree in their intuitive judgements about trolley cases.
(C1) Everybody's intuitions about trolley cases are initially incredible. (from P1, P2)
(C2) Trolleyology is an invalid method. (from C1)

As it stands, it is unclear what this argument says. For one thing, it contains a concealed quantification – assuming, of course, that it is formally valid. C2 follows from C1 only if we interpret C1 as a universal statement. Trolleyology is an invalid method only if *all* our intuitions about trolley cases are unreliable. Because then it would be impossible to find any trolley case to which it could justifiably be applied. If, however, we can have initially credible intuitions about trolley cases and if we restricted the application of the trolleyological method to these cases, then it would seem to be unobjectionable. Hence, we have to assume that C1 is a statement about *all* our intuitions. Furthermore, this version of C1 follows from P1 and P2 only if we also interpret P2 as a universal assertion. Plainly, if we interpret P2 merely as saying that people have different intuitions about *many* trolley cases, it would not follow that *all* intuitions about trolley cases are initially incredible. In that case, there might be some intuitions that are in fact reliable. And the success of a trolleyological investigation may largely be seen as a matter of finding them. So P2 and C1 have to be interpreted as *universal* statements. This, in turn, means that objectors have to interpret P2 in its strongest and least plausible form.

[62] I am indebted to Michael von Grundherr for making me aware of this objection.

The precise purport of P1 and P2 is also unclear. Both employ the notion of interpersonal disagreement. It is important to spell out what this involves. To this end, let us look at an example. Suppose that we present 1000 people with a case. 999, say, share an intuition about it. One person, however, reports a different intuition. Is it adequate, then, to say that people disagree in their intuitive judgements about this case? I believe that this is not so in any relevant sense. We are interested in cases of disagreement that would cast doubts on our intuitions. Their reliability, given a certain level of disagreement, depends on a number of factors besides the fact of disagreement itself. This is easy to see. Even on the assumption that our intuition about a given case is reliable, we would still expect to find some disagreement on it. When we ask people what they think about this or that case, we would expect some people to misapprehend it. Moreover, we would expect some people to interpolate additional assumptions that are not intended (contrary to Characteristic 3, mind you). And we may even suspect that some people are merely joking about their answer. We have no reason, then, to distrust our intuitions about trolley cases if it is possible to explain the level of disagreement among them by factors such as these. Hence, the notion of disagreement that is relevant here is *substantial* disagreement. P1 should, hence, be read as saying that everybody's intuitions about a case are unreliable if people disagree substantially in their intuitive judgements about that case. And P2 should be interpreted as saying that people disagree substantially about all trolley cases.[63]

So much, then, for the interpretation of the argument. Let us consider now whether the premises, P1 and P2, are plausible. P1 certainly is. Following Sidgwick, it makes sense to take our intuitions with a grain of salt if we find that they are subject to substantial disagreement. P2, however, appears to be false as a matter of empirical fact. It does not seem to be true, that is, that people substantially disagree on *all* trolley cases. Rather, there are some on which they agree and some on which they do not agree. Hauser et al. (2007), e.g., conducted a study of the moral intuitions of over 5000 subjects from 120 countries. They found that in the so-called "loop case" (Thomson 1985, 1402) 56 % of the people asked judged a given act permissible, while 44 % opposed this view.[64] However, Hauser et al. (2007) report a greater measure of agreement, ranging from 72 % to 88 %, in three other cases. These included one scenario that resembles Philippa Foot's original trolley case (*Edward's Case*) and one situation that is fashioned after Judith Jarvis Thomson's Fat Man Case (*George's Case*). This gives us reason to suspect that it

[63] It is hard to pin down, of course, what a reasonable threshold for substantial disagreement is.

[64] The description of the case was as follows: "Ned is walking near the train tracks when he notices a train approaching out of control. Up ahead on the track are 5 people. Ned is standing next to a switch, which he can throw to turn the train onto a side track. If the train hits the object, the object will slow the train down, giving the men time to escape. The heavy object is 1 man, standing on the side track. Ned can throw the switch, preventing the train from killing the 5 people, but killing the 1 man. Or he can refrain from doing this, letting the 5 die." (Hauser et al. 2007, 5) The question that was asked was "Is it morally permissible for Ned to throw the switch?" (Hauser et al. 2007, 5)

may be possible to find cases about which people do not disagree substantially and may suggest that we need not regard our intuitions about these cases as initially incredible.

Though we should, I believe, reject the argument from disagreement against trolleyology, it highlights a methodological point of some importance: When we construct arguments based on our intuitions about trolley cases, we had better check whether there is a substantial disagreement among them. In recent times, some of the foremost trolleyologists have neglected this practice and objections to *their* (ab)use of trolley cases may be raised quite fairly. Frances Kamm, e.g., does not seem to worry at all whether her intuitive judgements are agreeable to others, even if they figure as crucial premises in her argument.[65] It is no surprise, therefore, that other philosophers disagree with her to the extent that they see no common ground with her *at all*. Some even report that they "have found *no one* who agrees with her." (Norcross 2008, 66; emphasis in the original, NM) This sort of embarrassment is one that we shall seek to avoid in our case against consequentialism. Of course, since this is not an empirical study, we have no way of knowing for sure whether there is, in fact, a substantial disagreement on the intuitive judgements that we use to make our case. But let us, at least, be open to that kind of empirical refutation. And let us try to use only intuitive judgements that we may reasonably take to be entirely uncontroversial.

Objection 2
Trolley cases allow no general moral conclusions.

A further common objection to trolley cases consists in saying that they provide an inadequate basis for generalizations. It may be argued that it is illegitimate to draw any substantive lessons from trolley cases because many of the principles that philosophers have derived from them "do not produce the 'right' answer if applied beyond trolley cases." (Fried 2012, 13) At best, one might insist, trolleyologists may conclude that a given principle holds in a particular trolley case. However, a more general conclusion, it may be claimed, is unfounded.

This invective concerns, I think, not trolley cases per se. It concerns only one of the two uses of trolley cases. To be more precise, it concerns only the way in which we will *not* employ them. Here is why. Recall the distinction that we made above between the two primary purposes of trolley cases. We can use them in a constructive way, and we can employ them in a critical or destructive way. The idea behind the former use is to look at a series of trolley cases and to induce from them a general moral principle (cf. Kamm 2007, 4). The other is to use them as critical tests for given moral theories (cf. Tännsjö 2011, 295). As Popper (1959/2005) pointed out, there is, from a logical point of view, an asymmetry between these two enterprises. When moral theorists construct moral principles based on trolley

[65] In fact, much unlike Sidgwick, Kamm explicitly advises her readers to ignore the intuitive judgements of others. She does that since she believes that "much more is accomplished when one person considers her judgments and then tries to analyze and justify their grounds than if we do mere surveys" (Kamm 2007, 5).

cases, they do so, usually, hoping that these principles will match our provisional fixed points not just in those cases, but in *all* cases.[66] However, all they can say for sure is that their conjectured principles match the provisional fixed points in the particular trolley case(s) they have looked at. This is analogous to the case of the empirical scientist who cannot be sure that the curve she has plotted based on given data can accommodate the next data point. It is always possible that the very next trolley case shakes the firm confidence theorists put in their principles. And it is possible, furthermore, that their principles yield an entirely wrong conclusion when applied to real-life cases.

This, I believe, is the point of the objection. There is certainly something to it. It may, in fact, be very problematic to use trolley cases to derive moral principles and to then generalize them. Maybe real-life cases have important features – e.g. risk and uncertainty – which call for moral principles that are entirely different from those which suggest themselves in trolley cases. Note, however, that our critical use of trolley cases is unaffected by this criticism. Here, the aim is not to derive moral theories which match certain fixed points. Rather, the idea is to test specific doctrines and to check whether a case exists in which they give an answer that appears plainly wrong. Once we have established that a given theory does, indeed, give such a highly problematic answer, this is a *fact* we can work with. And it is a fact that does not change. As it turns out, then, the objection does not apply to our use of trolley cases.

Objection 3
Trolley cases suppose that normative factors are additively separable.

Another reason to reject trolleyology is to say that it employs a strategy that "relies on an underlying assumption concerning the role of [normative; NM] factors – an assumption that is questionable and should probably be rejected."[67] (Kagan 1988, 12) To explain, trolleyological inquiries do not, for the most part, rely only on one case. They rely on *pairs* or *series* of cases that are sometimes used to construct what Shelly Kagan calls "contrast arguments." The idea is that we take one case, vary one factor, holding everything else fixed, and compare the contrast case that results from this modification to the original case. If we find that the moral evaluations of the two cases differ, we conclude that this is due to the varied factor. If we find that nothing changes, we conclude that the factor does not play a role. In and of itself, this procedure is not objectionable. But here comes the kicker. Since the conclusion that a factor matters (or does not matter) in a particular case is quite unexciting, we shoot for a bolder claim and generalize our finding, employing what Kagan calls the "ubiquity thesis." The idea is that "if variation in a given factor makes a difference *anywhere*, it makes a difference *everywhere*." (Kagan 1988, 12; emphasis in the original) Hence, if a factor contributes to the normative evaluation of a given case, we conclude that it *always* makes this contribution. Moreover, if

[66] As we shall see below, however, some theorists deny this.

[67] This problem is also acknowledged and responded to by Kamm (2007, 345–367; esp. 348–349). See, also, Kamm (1983).

it does not play a role in that one instance, we take this to indicate that it *never* plays a role. The reason we believe the ubiquity thesis, Kagan argues, is that we think of normative factors as having additively separable weight. That is, we picture their weights like numbers in an addition equation. Each number on the left-hand side of the equation increases the value of the number of the right-hand side by a certain amount. And it does that independently of the values of the other summands. E.g., adding 5 to a sum of numbers always increases the value of that sum by 5. Analogously, normative factors that contribute to the rightness of a given act in one case are thought to make the same contribution to the rightness of acts in other cases.

Now, what is the problem with all of this? The problem is that this way of thinking about normative factors seems to be flat out incompatible with many respectable moral views. Many of us are *holists* about factors. To illustrate, most of us would agree that the fact that an act alleviates suffering is a morally relevant factor that generally counts in its favour. At the same time, however, many of us might believe that there is no reason to do a particular act, even though it alleviates suffering. E.g., when a guilty person is punished, there is perhaps no reason to bring relief to that person because she *deserves* to suffer.[68] This, at any rate, is what many people are inclined to think. Holding such a belief system, however, is inconsistent with the ubiquity thesis. According to the ubiquity thesis, if the fact that the act alleviates suffering counts in its favour once, it always has this effect.[69]

In regards to Fried's objection, we said that it applies only to constructive trolleyological arguments, while our inquiry seeks to establish a critical conclusion. The additive assumption and the ubiquity thesis, however, seem to underlie both the constructive and the critical use of trolleyology. Advocates of moral theories imagine trolley cases which support the significance of the factors which, according to their theory, are important. Critics of moral theories conjure up cases in which these factors seem to be irrelevant. Both generalize their findings using the ubiquity thesis. Advocates conclude that the respective factors are always relevant, while critics infer that they are always irrelevant. Hence, the reply that we gave to Fried's objection will not do here. Instead, we should draw attention to the fact that it is not always necessary for critics of a moral theory to show that the factors that the theory *always* takes to be relevant are, in fact, *never* relevant. It may be enough to show that a theory violates certain provisional fixed points in one instance. In fact, this is precisely what we shall attempt to do in our case against consequentialism. At no

[68]This position is called *retributivism* and is commonly associated with Immanuel Kant, who expressed the view in his *Metaphysics of Morals* (*Die Metaphysik der Sitten*). See, in particular, his infamous thought experiment of the dissolving civil society (cf. Kant 1803, 229).

[69]Note, however, that the chosen illustration is not a conclusive demonstration, as it presupposes a particular model of normative factors on which the property of an act to alleviate suffering is seen as an autonomous factor. This model can be rejected. Alternatively, we may distinguish between acts that alleviate the suffering of an innocent person and acts that alleviate the suffering of a guilty person that results from a just punishment. We can, then, take the former to be a right-making feature and the latter to be a wrong-making feature. This would resolve the difficulty in the present case.

point throughout the inquiry shall we generalize our conclusions beyond the level of the individual case. Hence, the ubiquity thesis has no role to play in our argument. We can allow ourselves to remain agnostic about it. [70]

Objection 4
Trolley cases are outlandish.

A further objection to trolleyology is to say that trolley cases are outlandish (cf., e.g., Kagan 1998, 76–77). In and of itself, this does not seem to be a problem. So what does the objection consist in precisely?

One interpretation is this. Since the scenarios that trolley cases present are so outlandish and unlikely, there is no reason to suppose that we have reliable intuitions about them. We may feel that our immediate judgement is robust. But this is a mistake. Our moral intuitions are not fit to judge cases of that kind. They evolved to help us deal with "normal" scenarios that we are likely to encounter on a daily basis. Hence, we should not trust them in freakish and unusual cases, such as trolley cases (cf. Singer 2005).[71]

This variant of the objection is highly implausible, as Allen Wood points out. "It is extremely rare," he says, "for a man to lure teenage boys into his apartment, then kill, dismember and eat them (. . .). But the rarity of such cases does not lead us to mistrust our moral intuitions about these cases" (Wood 2011, 69).

Another interpretation of the objection is this. Since trolley cases are unlikely ever to arise in practice, it is not fair to use them as tests for moral theories which aim to assist us in making *practical* choices. Those who bring up this objection seem to underestimate the tremendous ambitions that consequentialists have commonly had. They aspire to offer us a *universal* standard of right and wrong which applies, as Jeremy Bentham zealously professed, to "every action *whatsoever*" (Bentham 1838, 1; emphasis added, NM). Therefore, they seem to be in no position to cry foul when their critics invoke cases that are unlikely ever to arise in practice. Given consequentialists' "universal pretensions," their theories are, as Robert Goodin has emphasized, "absolutely fair game for purveyors of such fantasies" (Goodin 1995, 6).

There is an obvious objection that an objector may give to this reply. She can say that we should, perhaps, drop the "universal pretensions" of our moral theories and understand ethics, for once, as a *practical* discipline. Accordingly, we should eschew hypothetical examples and should use realistic scenarios to test doctrines. Or, as Thomas Pogge says, "[w]hat does it matter that our morality is inapplicable to the life context of fictitious Martians or of the ancient Egyptians, so long as it

[70] However, see Sorensen (1998, 272–273) for a critical rejoinder to Kagan's argument.

[71] Hare (1981, Chap. 8) gives a similar justification for Objection 4. As he argues, our intuitions are the product of our moral upbringing. He believes that "however good these may have been, they were designed to prepare [us] to deal with moral situations which are likely to be encountered" and that, therefore, "there is no guarantee at all that they will be appropriate to unusual cases." (Hare 1981, 132).

provides reasonable solutions to our problems."[72] (Pogge 1990, 660) On this view, the testing of theories against surreal scenarios is useless at best, as these situations are irrelevant in practice. Moreover, it may even be positively harmful because the use of hypothetical cases may lead us to reject moral theories that give entirely satisfactory answers to the practical problems they are intended for.[73]

This plea may be a fair point. It is noteworthy, though, that not all moral theorists are in a dialectical position to make it. Thomas Pogge can consistently raise it because he believes that the endorsement of moral principles "is consistent with their limited range."[74] (Pogge 2000, 138) As a pluralist about moral realms, he believes that principles vary across domains, where the domain of *real* or *possible* cases may be one to which specific principles apply – principles that do not apply to *outlandish* ones.[75] Consequentialists, on the other hand, are *monists* in the sense that they claim that there is precisely one moral criterion which applies to all acts and under all circumstances. Hence, they cannot put forward such a reply. In doing so, they would *ipso facto* abandon their moral theory.

Let me state, then, by way of conclusion, that the use of trolley cases is controversial. However, it appears to be rather unobjectionable, given the purpose to which we will put these cases in our subsequent investigation.

2.5 Summary

Let us sum up. The aim of our inquiry is to reject all versions of consequentialism. To develop an argument to this effect, we need to understand how moral theories can be evaluated and criticized. In this chapter, we tried to do just that.

In Sect. 2.1, we investigated the Rawlsian Approach, which seems to be the *modus operandi* in moral philosophy these days. It says, roughly, that a moral theory is acceptable to the extent that it fits our moral intuitions, is consistent, and establishes explanatory connections. As we discussed, this idea can be factorized into two sub-criteria, viz. *intuitive fit* and *coherence* which can, in turn, be factorized into two further criteria, viz. *consistency* and *systematicity* (or *connectedness*). In the debate about consequentialism, intuition-based arguments occupy center stage. Thus, we decided to base our argument on the criterion of intuitive fit.

[72] Similar views can be found, e.g., in Rawls (1951, 182 and 2003, 71), Hare (1981, 47–48), and Miller (2008, 44).

[73] An argument much like that was suggested to me by Andreas Suchanek in personal conversation. I believe that this way of thinking is common amongst scholars whose predominant focus is applied ethics.

[74] See, in particular, sections VIII through XIII in Pogge (2000).

[75] The sense in which the term "pluralist" is used here should not be confused with the sense in which it was used above. In Sect. 1.2.1, we called a moral theory pluralist if it contained more than one foundational moral principle. Here we call it pluralist if it contains different moral principles for different realms. A moral theory can be pluralist in the one sense but not in the other.

In Sect. 2.2, we talked about three interpretations of intuitive fit, viz. the Bottom-Up Approach (BU), the Reflective Equilibrium Approach (RE), and the Top-Down Approach (TD). This differentiation is based on a distinction between two types of moral intuitions, viz. high-level intuitions that concern abstract and principled questions and low-level intuitions which relate to concrete cases. BU is the view that only low-level intuitions are initially credible and that one should evaluate a moral theory according to its fit with them. RE is the more ecumenical view that both high-level and low-level intuitions can be initially credible and that a moral theory should, therefore, be evaluated in light of its overall fit with both of them. TD is the view that only high-level intuitions are initially credible and that we should judge a moral theory according to its fit with them. We argued that TD should be rejected and that either BU or RE is the correct view. This is important because our argument in Chap. 5 will rely on the assumption that intuitions about moral cases are admissible in moral inquiry.

In Sect. 2.3, we then considered how we can develop a workable methodic procedure for our investigation based on the evaluative criterion of intuitive fit. This step was necessary because intuitive fit does not, in and of itself, provide a testing procedure for moral doctrines. It merely gives us a philosophical ideal, viz. that our moral theories should fit our moral intuitions. We introduced and discussed the Provisional Fixed Point Approach (PFPA). The idea behind it is this. To test theories, we check them against provisional fixed points in our thinking. These provisional fixed points are intuitive convictions which are so strong that it seems reasonable to expect that an acceptable moral theory should be able to match them. If it does not, we can justifiably reject it. This conclusion is, of course, provisional in nature. It may turn out that no moral theory can fit all our provisional fixed points. In that case, the conclusion may not hold. Whether that is, in fact, the case is, however, a question we cannot address, given the limited scope of our inquiry. Having explained the basic idea of PFPA, we tried to make it more concrete by linking it with some of the points we had made in the second section of this chapter. We noted that we could, in fact, combine PFPA with BU, RE, and TD. TD, which we had rejected, would rule out provisional fixed points at the low level. However, BU and RE, which we did not exclude, allow them. Hence, we concluded, that PFPA in conjunction with either BU or RE permits us to draw on our intuitions about cases. This, in fact, is how we will proceed in our argument in Chap. 5.

In Sect. 2.4, we noted that, though PFPA allows us to use cases, it does not give us any guidance as to the kinds of cases we should use. We looked at trolley cases and concluded that they are suitable for the task ahead. We started by looking at their characteristics. Then, we went into their possible uses. Finally, we considered some objections to them that critics raised in recent times. Our main point was that valid criticisms of the methodical use of trolley cases do not seem to concern the way in which we will use them. They are directed only at the constructive use, while we are interested in employing them with a critical intent only.

Chapter 3
Methodology

In Chap. 1, we discussed what a moral theory is. Towards the end, we said what it plausibly means for a moral theory to be consequentialist. In Chap. 2, then, we worked out how such theories can be evaluated. We argued that one way of doing this is to set up a number of trolley cases, look for provisional fixed points about them, and check them against the implications of the theories that we seek to test. It seems, then, that there is not much more to say as far as preliminary remarks go. Apparently, we simply have to apply PFPA to the moral theories to which our definition of consequentialism applies.

Unfortunately, however, things are not as easy as that. This Definitional Method (DM), as I would propose to call it, does not work. Here is why. Upon closer inspection, the definition we stated above is very problematic. It is hardly clear which doctrines it includes and which ones it excludes. We have two options. We can either try to repair it somehow. As it will turn out, however, this does not work. The second option is to adopt a different interpretation of consequentialism. As Amartya Sen has pointed out, however, "[i]t is not easy to find any definition of consequentialism that would satisfy all those who have invoked that idea."[1] (Sen 2010, 217) The best we can do, it seems, is to state a very general Core Idea of consequentialism which envelops all doctrines that moral philosophers have taken to be forms of consequentialism (minor exceptions aside).[2] We could simply assume, then, that this Core Idea defines consequentialism and attempt to show that all moral theories that come under it are faulty. If successful, we would show *ipso facto* that all particular interpretations of consequentialism fail as well.

[1]Fred Feldman makes the same point when he says that "[m]oral philosophers have widely divergent views about the essence of consequentialism." (Feldman 1995, 584) A similar remark can be found in Kagan (1998, 309).

[2]To be sure, most moral philosophers would probably argue that the Core Idea is, in fact, too broad. This, however, need not concern a critic of consequentialism. After all, any criticism directed at the Core Idea would also apply to a consequentialism more narrowly defined.

N. Mukerji, *The Case Against Consequentialism Reconsidered*,
Theory and Decision Library A: 51, DOI 10.1007/978-3-319-39249-3_3

In using DM, we face a problem, however. The Core Idea of consequentialism is, it seems, immune to criticism. To formulate objections to it, we have to invoke further assumptions about the nature of consequentialist doctrines. This, however, gives consequentialists the chance to use a defence strategy, which we shall call the Humpty Dumpty Defence (HDD). Consequentialists can dodge every objection that assumes a definition narrower than the Core Idea. They can do this simply by saying that *their* understanding of consequentialism is different and that the criticism, hence, does not apply to the version of consequentialism *they* favour. As we shall argue, to address this problem, it is, in fact, necessary to abandon DM and to adopt an entirely different approach – one that does not assume a definition.

In Sect. 3.1, we shall discuss DM and the problems that it faces. After that, we shall finally introduce an alternative approach in Sect. 3.2. We will call it the Family Resemblance Approach (FRA). It will help us to divide the remainder of our inquiry into a sequence of strategic steps.

3.1 The Definitional Method

Before we look at the Definitional Method and its problems, we should state it explicitly.

Definitional Method (DM)

(i) Find a suitable definition of consequentialism.[3]
(ii) Show that all moral theories that come under this definition seem flawed.[4]

Since DM contains two distinct steps, it may potentially fail in two different ways.

- Firstly, it may turn out to be impossible to find a suitable definition of consequentialism.
- Secondly, though it may be possible to find a suitable definition of consequentialism, advancing substantive criticisms based on that definition may not be feasible.

In the following two sections, we will address both of these problems.

[3]To this end, it may seem useful to consider the use the term "consequentialism" was intended for when it first appeared in the literature. But this is not very helpful. The term originates from an article entitled "Modern Moral Philosophy" (1958) by Elizabeth Anscombe, but is used there in a sense that is quite uncommon today (cf. Kagan 1998, 309). Hence, the notion of consequentialism that Anscombe (1958) uses shall not play any role in our investigation.

[4]Note that this two-step procedure is independent of PFPA and trolleyology. In the second step, a critic of consequentialism might use different approaches.

3.1.1 The Definition of Consequentialism

It seems as though we already completed step (i) of DM in Sect. 1.2.3. We defined consequentialism as the class of moral theories which subscribe to the idea that "the rightness of an act depends only on its consequences." (Sverdlik 1996, 330) On that definition, it is the class of doctrines that accept only future-regarding normative factors.[5] At first glance, this definition seems rather unproblematic and clear. How helpful is it, however?

The first problem about it is that, if we take it too literally, it seems to *exclude* moral doctrines which are commonly viewed as versions of consequentialism. In fact, matters are worse. It excludes moral theories that most philosophers would regard as *paradigm cases* of consequentialism.

To see this, consider Classic Utilitarianism (CU) which is roughly the view that an act is right if and only if it maximizes the happiness sum. This theory is a *maximizing* doctrine in the sense that it requires the agent to choose not only an action which has good enough consequences but the action with the *best* outcome. Now, consider what this means. It means that the agent is required to choose the act whose consequences are at least as good as the consequences of any alternative. According to CU, then, whether an act is right does not depend only on *its* own consequences, but also on the counterfactual outcomes of any option that the agent might have chosen (cf., e.g., Dorsey 2012, 52; Mukerji 2009, 118 and Sumner 1987, 178).[6] Hence, the above definition would exclude CU.

Walter Sinnott-Armstrong proposes a quick fix for this problem. He identifies consequentialism with the family of moral doctrines which claim that "whether an act is morally right depends only on *consequences*." (Sinnott-Armstrong 2011; emphasis in the original) This definition leaves out the word "its." Admittedly, this solves the above problem since it classes CU as a consequentialist doctrine. For, even though CU does not judge the moral status of an act only by *its* consequences, it seems to determine whether an act is right only based *consequences*.

However, the amended definition is, nevertheless, unsatisfactory. Consider a doctrine which says that the rightness of *my* act depends on the results of what

[5] Similar definitions can be found elsewhere in the literature. Christine Korsgaard, e.g., defines consequentialism rather straightforwardly as the theory "that what makes an action right is its consequences." (Korsgaard 2008, 194) Pettit (1993, xiii) offers a slightly more roundabout characterization when he says that "[r]oughly speaking, consequentialism is the theory that the way to tell whether a particular choice is the right choice for an agent to have made is to look at the relevant consequences of the decision." This very broad definition is sometimes criticized because it makes the paradigmatically deontological rule "'Thou shalt not kill' a consequentialist rule." (Howard-Snyder 1996, 112) Note, however, that this does not constitute an objection to someone who criticizes consequentialism. If she starts from a definition that is wider and succeeds in making a case against consequentialism, this case only gets stronger.

[6] Vallentyne (1987, 25) refers to moral theories like classic utilitarianism "for which the permissibility of a given action depends not only on its features but also on features of the actions that are alternatives to it" as "comparative theories."

you do. Unless your act can be seen as a consequence of my act,[7] this doctrine is certainly not one that consequentialists would recognize as a version of their creed. On Sinnott-Armstrong's proposed definition, however, it is a form of consequentialism because it judges whether an act is right only based on consequences, viz. the consequences of *your* act.

Presumably, we should say, instead, that a moral doctrine is consequentialist if and only if it holds that the rightness of an act depends only on its consequences as compared to the outcomes of its alternatives. But this does not work either since there are apparently versions of consequentialism which do judge an act only by *its* consequences. In Sect. 4.2.2, we will come to talk about Satisficing Consequentialism. Certain versions of this kind of consequentialism judge that the rightness of an act depends only on whether *its* consequences are good enough, where the expression "good enough" is defined in absolute terms and not in reference to the consequences of other acts.[8] Technically, then, the definition would have to say that consequentialism is the family of doctrines which judge whether an act is right either solely based on its consequences or based on its effects as compared to the effects of alternative actions the agent might have performed. As this is extremely cumbersome, let us simply stipulate that it is what we mean when we say that the rightness of an act depends only on its consequences.

It seems, then, that we can overcome the problem we noted. Here, however, comes a problem which is more severe. It is the fact that it seems quite hard to pin down what our definition actually says. It presupposes that we can draw a line between the *intrinsic nature* of the act and its consequences .[9] The trouble is, as many scholars have recognized, that this line "is notoriously difficult to locate."[10] (Sumner 1987, 166) This might surprise the layman. After all, we use this distinction on a daily basis and treat it as unproblematic. On reflection, however, it becomes apparent that there is a problem. When we observe a person do something, we can usually give different adequate descriptions of what happens.

To illustrate this, let us look at an example. Imagine a college student who says: "Last semester I studied very intensely. As a consequence, my grades improved." In this example, the intense studying is identified as the student's act, while the improvement of her grades is regarded as its consequence. Obviously, though, the student might, instead, have said something like this: "Through intense studying, I was able to improve my grades last semester." This, too, is an adequate description

[7] As Nida-Rümelin (1993, 12) points out, in some instances we might view what you do as an outcome of what I do. Some acts may, after all, have other acts as their consequences.

[8] Slote (1985a, 50) and Hurka (1990, 107 and 2004, 71) mention such a form of satisficing consequentialism. Below, we will discuss it under the label "Non-Comparative Satisficing Consequentialism."

[9] This assumption has received a book-long treatment by Bennett (1998).

[10] This problem has been discussed by many authors, e.g. Allen (1967), Atwell (1969), Broome (1991, 3), Grisez (1978, 24), Hörster (1973), Macklin (1967a, b), Oldenquist (1966), Rachels (1997, 139–141), Rechenauer (2003, 12–13), Scarre (1996, 11), Schroth (2009), and Trapp (1988, 53).

of what happened. However, under this description, the improvement of the grades is picked out as the student's act and not as part of her act's consequences. To put it in more general terms, the problem with the notion of a consequence is this: There are, as we just remarked, usually many adequate ways of describing what a person does. By choosing among them, we can push the boundary between the act and its consequences back and forth.

Now, of course, there seem to be limits to this. It seems reasonable to assume that there are certain "basic actions" (Danto 1965, 1969) that we cannot analyse further into an act-component and a consequence-component. And, of course, if a person could not *know* and, therefore, could not have *intended* that a particular event would result from what she did, this event cannot adequately be incorporated into the description of her act (cf. Nida-Rümelin 1993, 14; Schroth 2009, 73). But this does not change the fact that the line between the act and its consequences is not clearly defined. This, in turn, makes it hard to understand what it means to say that we can define consequentialism as the set of doctrines that judge whether an act is right entirely based on its *consequences*.

Let us look at two responses to this problem. The first claims that we can draw the line between the act and its consequences clearly. To do this, we have to define "consequence" in the most encompassing way. We have to stipulate that "act" refers only to basic acts and that everything which we can see as the outcome of a basic act has to be called a "consequence" in the technical sense of the term.[11] This would contradict ordinary usage, but nevermind. It would solve the problem of indeterminacy.

Many moral theorists would not like this proposal. For if we posit an all-encompassing idea of a consequence, the definition seems to trivialize consequentialism.[12] If we drew the line between the act and its consequences in this way, we would count as consequences aspects of the act that are standardly regarded as intrinsic factors. Therefore, even doctrines which determine the rightness of an act partly based on what seems to be its intrinsic nature would pass as forms of consequentialism. Actually, it is worse than that. Even doctrines which appear *only* to look at the intrinsic nature of the act would pass as versions of consequentialism (cf. Portmore 2007). Consider, e.g., Kant's injunction against lying. Kant thought that lying was morally wrong, no matter what (cf., e.g., Kant 1799). On an all-encompassing notion of consequence, it becomes possible to interpret this view as a form of consequentialism. We can say "that lying is the *consequence* of speaking

[11] Many philosophers go even further and claim that the doing of an act will have the consequence that the act was done. On such a view, every act could be seen as a consequence, viz. as a consequence of itself (cf., e.g., Broome 1991, 4; Howard-Snyder 1994, 107; Sen 2010, 215–217; Slote 1985a, 35; Shaw 2006, 6). And this would hold for basic acts as well. The distinction between moral theories which judge the rightness of an act solely based on its consequences and those which do not would then collapse entirely. For a defence of this distinction, see Nida-Rümelin (1993, §14).

[12] Nida-Rümelin (1993, 51) already points to this problem, though in a different context.

in a certain way."[13] (Oldenquist 1966, 181; emphasis added, NM) Hence, we can hold that Kant's idea is that we should judge acts regarding a particular kind of consequence, viz. the consequence that a lie is told.[14] To sum up then, if we fix the notion of an act's consequences in a way that is sufficiently clear, our definition of consequentialism appears to become too inclusive. It includes doctrines which philosophers have traditionally conceived of as arch rivals of it.[15]

A second solution has recently been proposed by Schroth (2009). He, too, seeks to show that our definition can be rescued. In order to make sense of it, he claims, it is not necessary, as the previous proposal assumed, to explain where the line between the act and its consequences should be. There is no need to impose such a delimitation from *without* since consequentialist theories, he thinks, possess the resources to draw the line from *within*. Here is a sketch of Schroth's reasoning. The theoretical component of every consequentialist theory contains two sub-components, viz. a theory of the *right* and a theory of the *good* (cf. Rawls 1971/1999, 21). According to Schroth, consequentialist theories are defined in terms of a shared characteristic of their theory of the right. All of them judge whether an act is right only based on its consequences. This idea, however, is indeterminate, as we have seen, because there are always multiple descriptions which partition what happens into an act and its consequences. However, as Schroth claims, a criterion by which the *relevant* specifications can be picked out (for normative-ethical purposes) is at hand. It is provided by the theories of the good that the various forms of consequentialism may adopt.

The Distinction between Act and Consequence (*DAC*)
A given event which can adequately be described both as part of an act and as part of its consequences should be regarded as a the latter (for the purpose of normative ethics) if it could, on some remotely plausible theory of the good, be called "good" or "bad."

Two examples should make clear what DAC purports.

Flowers
Smith buys Suzy flowers, which makes Suzy happy.

[13] See, also, van Roojen (2004, 162). Essentially the same point is made by Stocker (1969, 280–282). He uses the example of promise-keeping.

[14] This kind of "Kantian Consequentialism" differs from the one proposed by Cummiskey (1990, 1996). What we just showed is that it is possible to reconstruct one of Kant's substantive moral views as belonging to a consequentialist moral doctrine if one interprets the notion of a consequence in an all-encompassing way. In contrast, Cummiskey argues that Kant's substantive deontological views do not follow from the basic tenet of his doctrine, viz. the Categorical Imperative. He thinks that we can make a case for consequentialism based on Kantian premises.

[15] It should be noted that not all authors see the trivialisation of consequentialism as a bad thing. Dreier (1993) and Portmore (2011), in particular, have proposed that consequentialists adopt a strategy of "consequentializing" other moral doctrines. They aim at reformulating the moral criteria of non-consequentialist doctrines in terms of consequences. The reason why one would want to do this is, as Portmore (2007) explains, that non-consequentialist doctrines may have more intuitive implications than consequentialism, while consequentialism is unified by a compelling idea, viz. that it is always morally permissible to bring about the best consequences. For a critique of the "consequentialization project," see, e.g., Schroeder (2006) and Brown (2011).

In this situation, recall, we can say that Smith's act was to buy flowers for Suzy and that Suzy was happy *as a consequence*. Alternatively, we can say that Smith *made* Suzy happy by buying her flowers. (The latter description seems to be adequate, too, since Smith presumably *intended* to make Suzy happy.) However, on DAC, it seems as though only the former description is relevant for our present purpose. Happiness is a paradigmatic case of something that we call "good." Hence, it should be seen as a consequence, according to DAC.

Jones's Lie
Jones tells Smith a lie. But Smith forgets it immediately and does not act on a false belief. So no harm is done.

In this situation, we can say, as we mentioned above, that Jones spoke to Smith in a certain way and that this had the consequence that he thereby lied to Smith. Alternatively, we can say – as I presume most of us would be inclined to say – that Jones lied to Smith. According to DAC, whether the first or the second way of putting it is relevant here, depends, as we have just noted, on whether the fact that somebody tells a lie pertains in any way to the good. Most moral theorists would agree that the fact that somebody tells a lie is, in and of itself, neither a good thing nor a bad thing. Hence, it should not be seen as the consequence of an act, but as (belonging to) the act itself.

Having clarified what DAC means, there are, I believe, two points we should make about it. Firstly, DAC has two merits that are worth highlighting. Its first merit is that it does give us a meaningful and non-arbitrary interpretation of the notion of a consequence. This will make our investigation easier because we can still talk about an act's consequences.[16] DAC's second merit is that it can explain why certain things that are commonly not conceptualized as consequences should, in fact, not be regarded in that way. According to many philosophers, the telling of a lie, e.g., should be seen as an act and not as an act's consequence. To be sure, even they would agree that it is perfectly *adequate* to say that the consequence of speaking in a certain way is that somebody tells a lie. What they deny, however, is that, from the standpoint of normative ethics, it is a *relevant* one. DAC supports this. The fact that a lie is told is not commonly seen as a good or bad thing in itself and should, hence, not be conceptualized as an act's consequence (cf. Schroth 2009, 74).[17]

Secondly, it may seem, as Schroth claims, that DAC can save our initial definition of consequentialism. But that is not the case. Since DAC specifies the notion of an act's consequence in terms of the more fundamental concept of the good, it, in fact,

[16]It should be noted that this way of using the notion of a consequence is not at all foreign to ethicists. See, e.g., Scheffler (1982/1994, 1–2) and Kagan (1998, 28).

[17]Note that, since rightness/wrongness and goodness/badness are logically distinct concepts, this does not exclude the possibility that a lie is *wrong* in itself. In fact, this is quite a common view to hold amongst deontologists. Unfortunately, John Broome seems to confuse these concepts (cf. Broome 1991, 4).

replaces the idea of a consequence by the concept of the good. Calling something the "consequence" of an act is, then, shorthand for referring to any aspect of a situation which one may, on some plausible theory of the good, see as good or bad. We can interpret consequentialism, then, as the class of doctrines that judge whether an act is right solely based on the goodness it produces.[18]

The Core Idea of Consequentialism
Whether an act is right depends only on the goodness that it produces.

This Core Idea, it appears, does, in fact, capture a feature that all versions of consequentialism share in common.[19] To be sure, most moral theorists would reject it as too broad. They would allow it only as a *minimal condition* for consequentialism and would say that the class of consequentialist doctrines consists only of a subset of all theories that conform to the Core Idea.[20] But we could, nevertheless, treat it as if it were an agreed-upon definition. For if we succeeded in showing that all moral theories which accord with the Core Idea seem flawed, we would have shown *ipso facto* that we can reasonably reject consequentialism on *any* interpretation of it. As it turns out, DM does not fail, then, because we cannot find a suitable definition of it.[21]

[18]This point is also made by Sumner (1987, 173).

[19]I borrow the notion of a "Core Idea" of consequentialism from Vallentyne (2006, 22). It is very similar to the ideas that Henson (1971) and Smart (1973), e.g., have put forward (under the heading "utilitarianism"). It also seems to resonate in Mendola (2006, 2) as well as in Oddie and Menzies (1992, esp. 512). Note that our Core Idea excludes what has been called "non-evaluative consequentialism" (cf. Sinnott-Armstrong 2011), such as "self-other utilitarianism" (Sider 1993) and "dual-ranking act-consequentialism." (Portmore 2008) These do not determine the moral status of the act solely based on the good that it produces. Rather, the idea is to use two evaluative measures of goodness – goodness for the agent and moral goodness – to compute a supplementary, non-evaluative ranking of act options which determines the moral status of the act. The reason we cannot include these doctrines lies, of course, in the limited scope of the inquiry. But it seems that one should be sceptical towards them anyway, at least insofar as they purport to represent variants of consequentialism. Many authors have confirmed my suspicion, e.g. Leonard Sumner, who writes that a consequentialist moral theory has to have an "operation for combining (..) separate goods into a *single* global value." (Sumner 1987, 172; emphasis added, NM) Scheffler (1982/1994), who has put forward a conception which is similar to Sider's and Portmore's (insofar as it regards rightness as depending on both moral goodness and goodness for the agent), refers to it as a "hybrid theory" rather than a consequentialist theory. In short, though I am willing to make significant concessions when it comes to the idea of consequentialism, I believe that those who hold that there are non-evaluative forms of it go one step too far.

[20]I am indebted to Lisa Herzog, who made me aware of the possibility that we might characterize consequentialism in terms of a minimal condition, even if an agreed-upon definition is not available.

[21]Note that our conception of consequentialism is similar to the conception of teleology proposed by Broome (1991, 3–4).

3.1.2 *The Humpty Dumpty Defence*

Step (ii) of DM consists in formulating critiques based on step (i). Let me be clear what this means. Criticisms should not aim at individual consequentialist doctrines. It is well-known that there are many examples of consequentialist theories which are highly objectionable. This does not demonstrate that we can reject consequentialism *as a whole*. The latter, however, is what we are trying to show. To establish our conclusion, then, DM requires us to show that consequentialist doctrines are objectionable due to the Core Idea and not because they possess some accidental feature that consequentialists can reject.[22] For the sake of illustration, let us briefly consider two possible kinds of criticism that an objector might formulate based on PFPA.[23]

As we pointed out in Sect. 2.3, PFPA can be combined with any interpretation of intuitive fit, that is, with TD, RE, or BU. Depending on the understanding of intuitive fit that we accept, it allows us to level two kinds of critique against consequentialism.[24] We can formulate case-based low-level critiques and principled high-level objections. Let us consider one example of each type, starting with an example of a common low-level criticism.

As we established above, the Core Idea of consequentialism is the idea that the moral status of an act depends solely on the good that it produces. It seems, then, as though every consequentialist doctrine demands that I always promote the good – whatever it consists in – to the greatest possible extent. And it seems to condemn my behaviour if I fail to advance it maximally. Now, one could say that, plainly, this is an outrageous demand. This becomes apparent as we apply this idea to a case. Suppose I want to go to the cinema to watch a movie.[25] This seems to be a rather innocent wish and acting on it surely cannot be morally wrong. But consequentialism demands that I abandon my plan and pursue a different course of action instead. It may require, say, that I volunteer at a soup kitchen to help the less fortunate. And it may urge that I donate the money, which I had intended to spend on the movie ticket, to help someone who needs it more than I. Suppose, then, that I do just what consequentialism demands. Once I have done that, can I see that movie? On consequentialism, it seems that the answer must be no. Apparently, consequentialism does not demand that I forgo my movie night just once or twice.

[22] Sen (1979), too, makes this point when he criticizes utilitarianism. He argues that consequentialism should not be rejected on the count that utilitarianism is objectionable since it is, according to him, not the consequentialist characteristic of the doctrine, but another feature (viz. welfarism), which is problematic.

[23] It might be that the problems about the Definitional Method can be addressed if another approach is chosen in the second step (e.g. one which focuses on coherence rather than intuitive fit). We shall, however, not pursue this possibility.

[24] On TD, recall, criticisms must be based on high-level fixed points, while, on BU, they must be based on low-level fixed points. The RE interpretation allows both kinds of critique.

[25] This example is borrowed from Kagan (1998, 154).

As we pointed out previously, it has a universal scope and applies to all my acts. Presumably, there is always something I could do which would do more good than going to the cinema. So it seems to follow that I should never see that movie. Now, this most certainly violates a low-level provisional fixed point, one might insist. Surely, any reasonable moral doctrine would allow me to see a movie, at least every once in a while. A theory which embraces the Core Idea of consequentialism, however, would always disallow me to do it. At any rate, so it seems. Hence, it should be rejected and, with it, all consequentialist doctrines.

At this point, we do not care whether the objection is, in fact, sound.[26] We are interested, rather, in whether or not it is a critique of the Core Idea of consequentialism and not a critique of an accidental feature of some consequentialist doctrines. *Prima facie*, it seems that it is. The objection, after all, starts at the Core Idea. Note, however, that an additional idea is smuggled in at some point in the reasoning. It is the notion that I should do what *most* promotes the good. To be sure, to most of us this may seem to be an innocent assumption. But it is not warranted by the Core Idea of consequentialism, as we stated it above. The Core Idea only claims that the rightness of an act depends solely on the good that it produces. This is different from saying that whether an act is right depends only on whether it *maximally* promotes the good. Though a subset of consequentialist doctrines does, of course, embraces the latter claim too, not every form of consequentialism needs to do that. Hence, it is not true that consequentialism can be rejected as a whole based on the objection that we just considered. All that its proponents need to do to dodge the criticism is to insist that, according to their understanding of consequentialism, it need not include a demand for the *maximization* of the good.

Maybe we just looked in the wrong place. Maybe we should leave case-based objections aside and look towards principled objections at the high level? If we would believe that RE is an adequate interpretation of intuitive fit, we could do this.[27] To see where this would lead us to, let us look at another example. To show that consequentialism is unacceptable, Julian Nida-Rümelin (1993, Ch. 10) draws on a result in social choice theory by Amartya Sen. Sen christened the "Impossibility of a Paretian Liberal." (Sen 1970b) The result suggests, roughly, that a moral doctrine cannot, at once, satisfy two high-level moral principles, viz. Liberalism (Condition *L*) and the Weak Pareto Principle (Condition *P*). *L* is basically the idea that every individual should be free to determine what happens in at least one area of her life. E.g., everybody should have a choice between sleeping on their back and sleeping on their belly.[28] *P* is a principle for ranking states of affairs which says what appears to be a sensible, minimal condition for the betterness of one state *vis-a-vis* another. It says that one state of affairs, *A*, is better than another, *B*,

[26] For a comprehensive rebuttal, see Kagan (1989, 231–270).

[27] Note, however, that, in doing this, we would abandon trolleyology, which is essentially a low-level methodology.

[28] This example is taken from Sen (1970b, 52).

if all morally relevant subjects are better off in *A* than in *B*.[29] Now, the relevance of this result may not be immediately obvious. So let us examine how one might formulate a high-level critique of consequentialism based on it. We start from the Core Idea of consequentialism, to wit, that the moral status of an act derives solely from the goodness that it produces. It suggests that every consequentialist doctrine must contain a theory of the good. As *P* seems to be a weak principle, every theory of the good available to consequentialists has to fulfil *P*. That is, every consequentialist has to judge an act permissible if it leads to a state of affairs where everyone is better off than under the *status quo*. Now, Sen's result suggests that the following follows: In certain situations, *every* consequentialist doctrine has to judge that it is permissible to violate condition *L* because *P* is incompatible with *L*. If we assume that *L* is a provisional fixed point which every reasonable moral theory must fit, we can reject all variants of consequentialism as unreasonable.

Again, let us put aside the issue whether this objection is sound and ask, merely, whether it is, in fact, a critique that applies to the Core Idea. Once more, we have to negate this. The objection does not only use the Core Idea, as we stated it above. It makes further assumptions. Most importantly, it makes the assumption that every theory of the good has to fulfil *P*. It does not matter whether *P* is a seemingly weak premise. What counts is that it is an *additional* assumption that the Core Idea does not contain. Hence, the argument cannot be seen as a knock-down objection. If any, it is a knock-down objection against consequentialist doctrines that endorse *P*. To dodge it, consequentialists can once more question the adequacy of the assumptions that their non-consequentialist rival makes about their moral view.[30]

At this point, let me state the problem about DM more generally. Though it is possible to delimit consequentialism in terms of a minimal condition (i.e. the Core Idea), it seems to be impossible to critique it based on this idea alone.[31] Apparently, we can formulate objections only if we assume a more narrowly defined interpretation of consequentialism as a target.[32] This, however, gives

[29] In fact, the Pareto Principle is usually cast as a principle that operates on individual preferences. We use it in a broader sense here. On this point, see also footnote 54 in Chapter 4.

[30] Sen (1982) has challenged the assumption that *P* is a condition to be imposed on the theory of the good and has proposed what he calls a "goal-rights system" which incorporates rights into the theory of the good. All consequentialists who believe either in hedonistic or objective accounts of the good deny *P* as well. After all, *P* seems to presuppose that goodness is a matter of preference satisfaction which both hedonists and objectivists about goodness deny.

[31] We could, of course, assume the negation of the Core Idea as a provisional fixed point. But, for one thing, this would simply beg the question. In addition, as I explained above, I am highly sceptical when it comes to the reliability of intuitions on such a plane of abstraction.

[32] E.g., some have suggested that (a) consequentialism (or teleology) refers only to those doctrines which judge whether an act is right in terms of whether or not it *maximizes* the good (cf. Frankena 1963/1973, 13; Nida-Rümelin 1993, 87; Rawls 1971/1999, 26; Vallentyne 1988, 89). The problem with this definition is that it excludes a form of consequentialism called Satisficing Consequentialism. We will consider it in Sect. 4.2.2. It judges the moral status of an act in terms of whether it does enough good. Other theorists have put forward definitions according to which (b) consequentialism refers to those doctrines which judge whether an act is right in an agent-neutral

consequentialists a strategy to defend themselves (cf. Ridge 2005, McNaughton and Rawling 1991, 168–169).[33] To apply it, they do not even need to argue. They only need to point out that they do not accept the definition on which opponents premise their objections.

Note that this is not only a theoretical possibility. In fact, consequentialists seem to use this strategy all the time. John Broome, e.g., uses it when he concedes that "[m]any serious doubts have been raised about consequentialism," before hastening to add that "they are not about consequentialism *as I defined it*." (Broome 2004, 42; emphasis added, NM) Walter Sinnott-Armstrong does so too, when he says: "Even if other philosophers mean something else by 'consequentialism,' I will be satisfied if my argument supports the view *that I labelled* '*consequentialism.*'" (Sinnott-Armstrong 2001, 345; emphasis added, NM) We shall call this strategy of argumentation the Humpty Dumpty Defence (HDD).[34]

HDD is, I think, the reason some philosophers have remarked that "[a]rguing with a consequentialist can be frustrating." (Brown 2011, 749) Consequentialists who apply this strategy to defend their position do not engage in substantive argument. They just engage in semantics, as Humpty Dumpty does in C. L. Dodgson's novel *Through the Looking-Glass*. He explains to a surprised Alice that "[w]hen *I* use a word (. . .) it means just what I choose it to mean – neither more nor less." (Carroll 1871/1990, 103; emphasis in the original)

If our diagnosis is correct, any case against consequentialism that we construct along the lines of DM can easily be dodged by consequentialists who use HDD. The question arises, therefore, whether there is an alternative method which is immune to this problem. I believe that there is such a method. In what follows, we shall discuss it.

---

way, i.e. independently of the identity of the agent (e.g. McNaughton and Rawling 1991, 1995). Many theorists, however, have disagreed with this definition and have proposed agent-relative versions of consequentialism (e.g. Broome 1991; Portmore 2011; Sen 1982). In fact, Egoism which we will consider in Sect. 4.2.3. is a clear instance of an agent-relative consequentialist theory (cf., e.g., Dreier 1993, 22–23; Österberg 1988, 129; Portmore 1998, 2, 2001, 371; Sinnott-Armstrong 2011). Yet others have suggested that (c) consequentialism should be defined as the family of moral theories which reject the notion that there are moral constraints which forbid acts of certain *types* (e.g., murders). Crisp (2005), however, has pointed out that "[i]t can be said to be a constraint on our acting in any way that we must maximize the good." This would mean that, e.g., classic utilitarianism can be seen as a non-consequentialist doctrine. Vallentyne (1987, 22) makes the same point.

[33]McNaughton and Rawling aptly call this strategy the "consequentialist vacuum cleaner." A description of it is also given by Jamie Dreier. "Whenever the opponent manages to make it plausible that something of value is lost when an agent maximizes good consequences," he says, "the consequentialist just slurps up that value and tosses it into his basket of goods-to-be-maximized." (Dreier 2004, 143)

[34]I first introduced and analysed the Humpty Dumpty Defence in Mukerji (2013b). In that text, I refer to it as the strategy of "interpretive divergence."

3.2 The Family Resemblance Approach

Apparently, we cannot premise our investigation on the assumption that consequentialist doctrines share a (number of) definitional characteristic(s). This would, after all, lead to the problems that we identified above. In what follows, we shall, therefore, look at a method for criticizing consequentialism which does not start with a definition. Rather, it is based on the idea that consequentialist doctrines share a *family resemblance*. For this reason, we will call it the Family Resemblance Approach (FRA).

Our starting point will be a tentative formulation of FRA. It will turn out to be problematic, however, since it makes an unrealistic assumption about the way in which objections to consequentialism are linked to the logical components of consequentialist doctrines. We need to drop this premise to remedy the faults of the method. This is what we will attempt to do in the second formulation.

3.2.1 The First Version

It may be hard to understand how a philosophical investigation can get off the ground if the object to be investigated is not defined.[35] However, this resistance is rooted in a warped view of language. On this view, there are only two kinds of general terms, viz. "basic terms" and "composite terms" (cf. Sluga 2006, 4). The former are used to pick out observable characteristics, while the latter refer to more complex objects and are defined in terms of the former. On this view, it seems to be inexplicable how we can make sense of the term "consequentialism" if it does not fall into either category.

We do not need to buy into this naïve ontology of language, however. As Ludwig Wittgenstein famously suggested, many general terms may fall into a third category. They may be "family resemblance terms." Recently, some philosophers have suggested that "consequentialism" should be interpreted as such a "family resemblance term" (e.g. Portmore 2007, 46; Sinnott-Armstrong 2011).[36] In what follows, we shall explore this possibility and examine how we can use it methodologically.

To start, we should get clear on the idea of family resemblance.[37] As Wittgenstein explains, family resemblance obtains between the objects of a given class if

[35] This view goes back at least to Plato. See, e.g., the dialogue *Euthyphro* where Socrates insists that his interlocutor produce a definition of piety. See, also, Woodruff (2010), esp. his remarks on Socratic definition and the priority of definition (sec. 3–4).

[36] Scarre (1996, 4) has suggested that the idea of family resemblance may be used to characterize the class of utilitarian doctrines.

[37] The notion of family resemblance is usually seen as originating in Ludwig Wittgenstein's work. The first mention is in his *Blue Book* (1933–1934/1960, 17–18). An oft-cited passage on family resemblance is found in the *Philosophical Investigations* (1953/1986, §§66–67).

Table 3.1 An illustration of family resemblance: criss-crossing

a	*b*	*c*
ABC	*ADE*	*BDF*

Table 3.2 An illustration of family resemblance: overlapping

a	*b*	*c*	*d*
ABC	*BCD*	*CDE*	*DEF*

they form "a complicated network of similarities overlapping and criss-crossing" (Wittgenstein 1953/1986, §66), while there is *no single feature* they all share in common which could serve as the basis for a definition.[38]

To be sure, overlapping and criss-crossing are distinct ideas.[39] Consider three objects *a*, *b*, and *c*. Each of them possesses three out of six components *A*, *B*, *C*, *D*, *E,* and *F*, as Table 3.1 shows.

The similarities that are characteristic of the family *abc* criss-cross in this case. That is, the components that the objects share are different throughout pairs. *a* and *b* share component *A*. *b* and *c* share *D*. And *a* and *c* share component *B*. Overlapping, on the other hand, is illustrated by Table 3.2.[40]

Here, the similarities between objects extend not only throughout pairs. They overlap – at least in the case of components *C* and *D*. These run through *a*, *b*, *c* and *b*, *c*, *d*, respectively.

Let us ask, then, whether it is adequate to interpret the class of consequentialist doctrines as a family. Given what we said so far, there may be doubts. In Sect. 3.1.1, recall, we concluded that there is, in fact, one idea – viz. the Core Idea – which lies behind *all* variants of consequentialism. This seems to rule out that we can interpret the term "consequentialism" as a family resemblance term. For if the Core Idea lies behind all consequentialist moral doctrines, they apparently *do* share one characteristic in common.

At this point, it seems to be instructive to introduce the distinction between the substantive *content* of a moral theory and its *structure*, as it is drawn, e.g., by Hurka (1992, 71).[41] To this end, let us go back to the first example of family resemblance which illustrates criss-crossing. We can interpret *a*, *b*, and *c* as moral theories and *A*, *B*, *C*, *D*, *E*, and *F* as their logical components. Consider the statement:

1. _ contains components *A*, *B*, and *C*.

[38] Wittgenstein also includes further characteristics. He suggests, e.g., that family resemblance terms are also vague, i.e. their extensions are indeterminate. As Michael Forster argues, however, this conflation of family resemblance and vagueness is a mistake. He says that "it would in principle be quite consistent with Wittgenstein's core model of family resemblance concepts (...) that it leaves the extension of such a concept perfectly determinate." (Forster 2010, 67) In what follows, we shall follow Forster's view.

[39] Both of the following examples are adapted versions of examples used by Forster (2010, 67).

[40] We take this example from Forster (2010, 67).

[41] This distinction also seems to be implicit in Brink (2006, esp. 381).

The placeholder _ stands for a moral doctrine, e.g. *a*. If we plug in *a* for _ in (1), we get a true statement about the *content* of *a*, viz.:

2. *a* contains components *A*, *B*, and *C*.

If we plug in *b* for _ in (1), however, we get a false statement, viz.

3. *b* contains components *A*, *B*, and *C*.

In contrast, consider

4. _ contains three out of six components *A*, *B*, *C*, *D*, *E*, and *F*.

If we substitute *a* for _, we get a true statement about *a* again, viz. a true statement about the *structure* of *a*.

5. *a* contains three out of six components *A*, *B*, *C*, *D*, *E*, and *F*.

However, we can also plug *b* in for _ and get a true statement about the structure of *b*.

6. *b* contains three out of six components *A*, *B*, *C*, *D*, *E*, and *F*.

The fact that (2) and (3) are true and false, respectively, though (5) and (6) are both true, suggests that the members of a family of moral doctrines *a*, *b*, *c*, ... may share a given *structural* feature, though they may share no *substantive* feature. If, then, the Core Idea behind consequentialism relates to the structure of consequentialist doctrines rather than to their content, this does not rule out that we can construe consequentialism as a family of theories. This, indeed, seems to be the case. The Core Idea concerns the theoretical part of moral theories. It is the notion that the rightness of an act depends only on its goodness. What this says is that every consequentialist doctrine possesses certain *kinds* of components – a conception of the right and a conception of the good (*plus* further subsidiary components) – which are connected in a particular way. The Core Idea, hence, does not relate to the substantive content of consequentialist doctrines. It only concerns their structure, which is compatible with the assumption that the versions of consequentialism form a family.

It seems, therefore, that we can characterize consequentialism as a family of theories. In and of itself, however, this result does not help us much. We need, of course, a clearer idea as to how this family can be delimited. As a first step, it will be useful to conduct an inquiry into the logical structure of consequentialist doctrines. To this end, we should look towards a theory that is undoubtedly a version of consequentialism. Fortunately, it is not hard to find. Proponents, as well as critics of consequentialism, unanimously agree that CU is a *paradigmatic member* of the consequentialist family (cf. Sinnott-Armstrong 2011). We can, hence, use it as a starting point and factorize CU into logically independent components, $C_{11}, \ldots, C_{n1}$. This first step will reveal a number of claims, $C_{11}, \ldots, C_{n1}$, that are *typically* involved in a consequentialist doctrine. It will, *ipso facto*, expose the logical structure that is common to all forms of consequentialism. That is, it will tell us the precise number *n* of components that are involved in a consequentialist theory. Moreover, it will tell

Table 3.3 Construction kit for consequentialist doctrines

C_{11}	C_{12}	...	...	C_{1a}
C_{21}	C_{22}	...	...	C_{2b}
⋮	⋮	...	...	⋮
C_{i1}	...	C_{ij}	...	⋮
⋮	⋮	...	...	⋮
C_{n1}	C_{n2}	...	...	

us something about the nature of the elements that consequentialist theories contain. They must be of the same kind as $C_{11}, \ldots, C_{n1}$, respectively.

In the next step of our inquiry, we can make use of a logical consequence of our assumption that the class of consequentialist doctrines is taken to be a family. This assumption implies that there have to be alternatives to every paradigmatic component.[42] That is, for each logical element of CU, C_{i1}, there has to be at least one alternative, C_{i2}. To understand the range of possibilities that consequentialism allows, we should, therefore, investigate which non-standard alternatives there are to each of the paradigmatic components $C_{11}, \ldots, C_{n1}$. That is, we should, in a second step, take stock of the alternatives, $C_{i2}, \ldots, C_{im}$, to each of the paradigmatic components $C_{i1}, i = 1, \ldots, n$.

Upon completing the first two steps, we end up with a "construction kit" for consequentialist doctrines, as shown in Table 3.3. It contains all logical building blocks that consequentialists can use to construct a theory. The first column shows all the paradigmatic elements, i.e. those of CU. Each row contains one paradigmatic component and all its alternatives. To construct any consequentialist doctrine from the kit, we simply choose one option from each row. If the components are, in fact, logically independent, then it seems to be possible for consequentialists to combine every two components, C_{ij} and C_{kl}, from two different rows i and k in a consequentialist moral theory.[43]

Our goal, recall, is to show that we can reject consequentialism as a whole. To this end, we have to show that every member of the consequentialist family is subject to a reasonable objection. To judge whether this enterprise is at all feasible, we should try to understand, at least roughly, how many members this family has. Before we do the maths, it is useful, however, to introduce some technical verbiage.

We should distinguish between two important concepts: that of a *determinate component*, C_{ij}, and that of a *determinable component*, $C_{i(-)}$.[44] The determinate component, C_{ij}, is the $(j-1)^{\text{th}}$ alternative to the i^{th}paradigmatic element, C_{i1}.

[42] If there were not, all consequentialist doctrines would share one characteristic in common. This characteristic could be used to define consequentialism which, in turn, contradicts the assumption that they form a family.

[43] That, in fact, is only roughly correct. Towards the end of the section, we will qualify this idea.

[44] We borrow this piece of conceptual apparatus from Oddie (2001, 314).

In contrast, the determinable component $C_{i(-)}$ is not a component at all, but a placeholder that stands for C_{i1} or its alternatives. (Below we will, for ease of use, frequently employ the term "component" rather than "determinate component" when it is clear enough what we mean.) Let $D(C,\ C_{i(-)})$ stand for the determinate component that the consequentialist doctrine C embraces in place of the determinable component $C_{i(-)}$. We can represent C, then, as the following logical conjunction.

Logical Structure of a Consequentialist Theory

$$D\left(C, C_{1(-)}\right) \wedge D\left(C, C_{2(-)}\right) \wedge \cdots \wedge D\left(C, C_{n(-)}\right)$$

A further important notion is the *domain* of a determinable component, $C_{i(-)}$. It is simply the set of alternatives for which $C_{i(-)}$ is a placeholder.

Now, how many different combinations $D(C, C_{1(-)}) \wedge D(C, C_{2(-)}) \wedge \ \ldots \ \wedge D(C, C_{n(-)})$ that form a consequentialist doctrine C can there be? This will depend on the number n and the number of alternatives, $C_{i2}, C_{i3}, \ \ldots$ to each paradigmatic component C_{i1}. To skip ahead, in Sect. 4.1 it will turn out that the theoretical and practical parts of CU comprise eight logically distinct components in total. On the assumption that there is, at least, one alternative to each paradigmatic component, there would be at least $2^8 = 256$ possible combinations. The actual number is likely to be much greater since there are, plausibly, more than two alternatives to every paradigmatic component. However, even if there were "only" 256 consequentialist theories, the task of going through all of them would be insurmountable. We should, therefore, seek a methodological shortcut which can give us the desired result much faster.

An initially attractive idea starts from the observation that every consequentialist doctrine, C, has to endorse exactly one determinate component, C_{ij}, for each determinable component, $C_{i(-)}$. Hence, if we could show that every determinate component in the domain of $C_{i(-)}$ is subject to a reasonable and convincing objection, we would show *ipso facto* that all versions of consequentialism deserve to be rejected. It seems, therefore, that we should, in a next step, survey objections O_1, $O_2, \ \ldots O_o$ to consequentialism and correlate them with determinate components, C_{ij}. The fourth and final step suggests itself, then. It would just consist of putting together a comprehensive case against consequentialism. That is to say, it would consist of putting together a set of objections $O{=}(O_1, \ \ldots, \ O_m)$ such that there would be, at least, one objection, O_l, for all determinate components, $C_{i1}, \ \ldots, C_{im}$, of at least one determinable component, $C_{i(-)}$, $l = 1, \ \ldots, m$.

In summary, the methodic procedure that we just outlined can be stated thus:

Family Resemblance Approach – Version 1 ($\mathbf{FRA_1}$)

(i) Factorize CU into logically independent components, $C_{11}, \ \ldots, C_{n1}$.
(ii) Take stock of all alternatives, $C_{i2}, \ldots, C_{im}$, to each of the paradigmatic components C_{i1}, $i = 1, \ \ldots, n$.
(iii) Survey objections $O_1, \ O_2, \ \ldots O_o$ to consequentialism and correlate them with determinate components, C_{ij}.
(iv) Put together a set of objections $O{=}(O_1, \ \ldots, \ O_m)$, such that there is at least one objection, O_l, for all determinate components, $C_{i1}, \ \ldots, C_{im}$, of, at least, one determinable component, $C_{i(\cdot)}$, $l = 1, \ \ldots, m$.

Having stated FRA_1 explicitly, let us briefly discuss whether it fulfils its purpose. In Sect. 3.1, recall, we found that DM leads to the problem that consequentialists can use the Humpty Dumpty Defence (HDD). That is, we saw that when we employ this method consequentialists will be able to say that our argument merely shows that some forms of consequentialism are problematic. However, it does not pertain to their favoured version of it. Our motivation for developing FRA_1 was to work out a method for criticizing consequentialism that is not plagued by HDD. Let us ask, then, whether we have succeeded in doing that.

Evidently, consequentialists cannot dodge a case based on FRA_1. Here is why. Consequentialists can only use HDD if they can retreat to a variant of their creed which is not covered by the argument. They will not be able to do this, however, when they face a set of objections, $O_1, \ldots, O_m$, such that there is an objection to each determinate component, $C_{i1}, \ldots, C_{im,}$ of at least one determinable component, $C_{i(-)}$. As we said above, every consequentialist doctrine C can be represented as a conjunction $D(C, C_{1(-)}) \wedge D(C, C_{2(-)}) \wedge \ldots \wedge D(C, C_{n(-)})$. Consequentialists must, hence, embrace either C_{i1} or one of its alternatives $C_{i2}, \ldots, C_{im}$. It does not matter which version of the doctrine they embrace. If, therefore, an FRA_1-based argument is successful, it shows that each of these alternatives is subject to *some* serious objection. Hence, when consequentialists use HDD, they jump out of the frying pan and into the fire. No matter which version of consequentialism they retreat to, there is always a decisive objection to it. It seems, therefore, that FRA_1 does, in fact, fulfil the purpose for which we designed it.

Let us conclude this section with a few clarificatory remarks that pertain to the construction kit for consequentialist doctrines. Above, we said that we can construct a consequentialist theory from the kit by choosing exactly one component from each row. And we said that, if the components are, in fact, logically independent of one another, then it should be possible for consequentialists to combine every two components, C_{ij} and C_{kl}, from two different rows i and k in a consequentialist moral theory. Though this is roughly correct, we should, I think, make a distinction between two types of logical independence that may obtain within the kit, viz. logical independence between determinate components and logical independence between determinable components.

When we say that logical independence holds between the determinate components C_{ij} and C_{kl}, from two different rows i and k we mean that the fact that one endorses (or rejects) C_{ij} does not necessitate that one accepts (or rejects) C_{kl} and vice versa. As we shall see in Sects. 4.1.1 and 4.1.2, all components of CU are independent of one another in this sense.

The second kind of logical independence is one which may obtain between the rows or determinable components of the kit. To understand what this means, we need to draw on a notion that we did not introduce above, viz. the idea of a *proper subdomain* of a determinable component. As we said above, a determinable component is a placeholder for determinate components. And the domain of a determinable component is the set of determinate components that a consequentialist moral theorist can adopt in its place. A *sub*domain of a determinable component is

simply a subset of this set. And a proper subdomain of a determinable component is simply a proper subset of this set. Now we possess the conceptual arsenal to define logical independence between determinable components or rows. Logical independence obtains between two determinable components, $C_{i(-)}$ and $C_{j(-)}$, if and only if there is no proper subdomain of $C_{i(-)}$, such that the choice of a determinate component from this proper subdomain of $C_{i(-)}$ would logically commit one to the choice of a determinate component from a proper subdomain of $C_{j(-)}$.

The important thing to note about these two forms of logical independence is that the latter implies the former, but not vice versa. That is, the fact that the determinable components $C_{i(-)}$ and $C_{k(-)}$ are logically independent implies that any two determinate components C_{ij} and C_{kl} within their respective domains are logically independent of one another. However, the fact that two determinate components $D(C, C_{i(-)})$ and $D(C, C_{k(-)})$ of a consequentialist doctrine C are logically independent does not imply that the determinable components $C_{i(-)}$ and $C_{k(-)}$ to which they belong are also logically independent. CU provides an illustration of this. As we shall see in Sects. 4.1.1 and 4.1.2, its determinate components are logically independent. But it seems as though certain CU-components impose *domain restrictions* on the determinable components of other determinate components. That is, it is apparently not possible to combine them with all other determinate components in the other rows of the kit. To be more concrete, one of the elements of CU is, as we shall see below, Summation. Summation says, roughly, that the goodness that attaches to an act is measured by how it affects the *sum* of well-being in the world. Though the choice of Summation does not commit a proponent of CU to any particular determinate component in the domain of the other determinable components, it obviously commits her to a given subdomain in the case of at least one determinable component. Plainly, a moral theorist who accepts Summation cannot choose just about any conception of well-being. For Summation presupposes that well-being can be quantitatively measured and compared across individuals. The choice among conceptions of well-being is restricted, then, by the choice of Summation as a determinate component since the latter logically requires one to choose a measure of well-being which allows interpersonal (unit) comparability.

The point we just established has an interesting consequence. If logical dependencies do exist to a great extent between determinable components, this implies that certain combinations should be ruled out on grounds of coherence. In fact, this may be an attractive strategy to pursue for critics of consequentialism. However, we shall not make use of it. As we said above, we will focus on objections to consequentialism based on intuitive fit. That is, for the purpose of our investigation, we will simply ignore the possibility that logical dependencies may exist between determinable components. We will assume for all i and for all k that it is possible to combine all determinate components in the domain of $C_{i(-)}$ with all determinate components in the domain of $C_{k(-)}$. Should it turn out that this is not in fact so, so much the worse for consequentialism.

I should, perhaps, mention a further kind of dependency between the components in the construction kit. The choice of a particular determinate component from one of the rows may not only impose a logical domain restriction on the choice of

other determinate components. In addition, it may impose *plausibility restrictions*. Given that a consequentialist moral theorist has committed herself to determinate component C_{ij}, it may be rather implausible for her to commit to C_{kl} because the two seem in tension with one another. Here are two examples.

In Sect. 4.1.1, we will discuss the idea of Hedonism, which is one of the determinate components of CU. It is the view that we can measure individual well-being by the balance of pleasures over pains. Tim Scanlon believes that a theorist who subscribes to this idea commits herself to the idea of Summation. This is not because Hedonism logically entails Summation, but because it would not make sense to embrace anything other than Summation in conjunction with Hedonism. In Scanlon's words:

> Although hedonism does not entail sum-ranking [=Summation; NM], it is plausible that if pleasure and the absence of pain is the sole ultimate value, then since more of it is better than less, the value of states of affairs should be determined by the amount of happiness they contain. (Scanlon 2001, 40)

The second example has to do with the idea of Perfectionism. Perfectionism is a competitor of Hedonism. That is, it is an alternative view about the value of a human life. It involves the idea that humans essentially possess certain human properties and that an individual human life is better the more these are brought to fruition. Perfectionism can be combined with a maximizing moral conception. But, as Thomas Hurka points out, this "is not the only possible structure, for perfectionism could take a *satisficing* form." (Hurka 1993, 56; emphasis in the original) That is, "[i]t could care that humans develop their nature to some reasonable degree but be indifferent about what they do beyond that." (Hurka 1993, 56) As Hurka goes on to explain, though, the combination of Perfectionism and Satisficing is not very attractive.

> Satisficing may be attractive for some values, for example, pleasure or desire-fulfilment, but not for an ideal of perfection. If we are attracted by this ideal, it is as something to be maximized and pursued to the highest degree. (Hurka 1993, 56)

Hence, if a consequentialist theorist subscribes to Perfectionism, this seems to impose a plausibility restriction on her. She should also be a maximizer, not because this a logical dictate of Perfectionism but, rather, because it is philosophically more plausible.

As it turns out, then, various possible combinations of determinate components may be ruled out in advance, not only on grounds of coherence, but also on substantive philosophical grounds. It may be an interesting undertaking for a critic of consequentialism to fathom which combinations of components do or do not make sense. It is hard to see, however, how this could be done based on the trolleyological approach to which we have committed ourselves in Sect. 2.4. For our purposes, we will ignore the possibility that certain combinations of determinate components can be ruled out as philosophically implausible. Again, should this turn out to be false (which is likely), so much the worse for consequentialism.

3.2.2 The Second Version

It looks like we are done now as far as issues of methodology and procedure are concerned. But, unfortunately, we are not. As it turns out, FRA_1 is not tenable. We made an assumption in one of the steps which jeopardizes our enterprise. Step (iii) of FRA_1 requires that we consider objections to consequentialism and work out, in each case, which determinate component seems to be the culprit. This presupposes, rather naïvely, that objections to consequentialism correlate with the individual components of consequentialist doctrines.

To be sure, we can certainly reject individual components based on one decisive objection alone. E.g., it is quite obvious that any consequentialist theory which endorses the idea that it is right to *minimize* value is blatantly absurd since it would give moral agents the instruction to cause as much mayhem and destruction as possible on any plausible theory of the good.[45] However, "regular" objections to consequentialism target doctrines that do not, on the face of it, appear completely unreasonable. They work differently. Typically, it is not possible to attribute these objections to a single determinant component.

To illustrate, let us consider the well-known objection from moral freedom against consequentialism. In a nutshell, it claims that consequentialist doctrines leave us no "wiggle room" (Dreier 2006, xi), as it were. They usually designate one and only one option in each choice situation as morally permissible, viz. the act with the best consequences. Hence, consequentialist moral theories constrain our freedom of choice to an extent that is counter-intuitive. Intuitively, a moral agent is "free to lead her life as she chooses in any of a wide variety of ways."[46] (Arneson 2004, 34).

We might argue that the objection from moral freedom applies in virtue of the principle of Maximization, which is characteristic of typical variants of consequentialism, including CU. It claims that an act is right if and only if it produces at least as much goodness as any alternative. Since, normally, there is one uniquely best act, Maximization narrows down the range of permissible choices to a single act – i.e. the best one.

As it turns out, however, Maximization is neither a necessary nor a sufficient condition for this objection. It is not sufficient because not all maximizing consequentialist doctrines are overly freedom-constraining. Some of them accept a comparatively "coarse-grained" theory of the good (cf. Vallentyne 2006, 26). This allows many acts to tie regarding goodness, such that there will normally be many best acts. All of them are permissible, then, on a maximizing consequentialist view.

[45] Interestingly, just like the multiplication of two negative numbers yields a positive number, the combination of two rather abstruse components can yield an apparently plausible consequentialist doctrine. E.g., when we combine the claim that goodness should be minimized with the claim that pain is good and pleasure bad, we get a doctrine that issues the same deontic verdicts as CU.

[46] As we will discuss in Sect. 4.2.2.1, Bernard Williams famously made the point that a moral theory should, in particular, allow us to pursue *ground projects* around which we build our lives. See, e.g., Williams (1973, 108–118).

We can achieve the same result when a certain degree of value incommensurability is allowed into the theory of the good. When two act options are incommensurable, Maximization permits us to choose either one. It, too, will, hence, allow us some degree of freedom of choice.[47]

Maximization is not a necessary condition for the objection either since non-maximizing forms of consequentialism may also restrict moral freedom to an implausible extent. As we shall see in Sect. 4.2.2, Satisficing is an alternative to Maximization. It is the principle that an act is right if and only if it brings about enough goodness. Evidently, a version of Satisficing Consequentialism which sets the bar for enoughness sufficiently high may severely constrain moral freedom, too.

It is, hence, not possible, as step (iii) of FRA_1 requires, to correlate the objection from moral freedom with exactly one determinate component of a consequentialist theory. Since, for all we know, many objections to consequentialism might function like this,[48] it seems that our chances of putting together a successful case against consequentialism are slim if we proceed along the lines of FRA_1. We had better rework it.

How do we do this? Our goal, recall, is to show that all members of the family of consequentialist moral theories are subject to a reasonable and decisive objection. However, as we discovered in the previous section, the consequentialist family has very many members. Given the scope of this inquiry, we cannot go through all of them. We need, as we emphasized above, a methodological shortcut that enables us to get the desired result much faster. As we just learnt, the first potential shortcut leads us into a dead end. But, perhaps, there are other options.

For a start, it would surely help if we could somehow show that it is not necessary for us to consider all members of the consequentialist family. This would reduce the scope of our inquiry. However, what justification could there be to exclude certain variants of consequentialism? There are, I believe, two.

- Firstly, we can argue that we may justifiably ignore particular determinable components because the distinction between the alternative determinate components in its domain is unmotivated.
- Secondly, we can argue that it makes sense to focus only on those determinate components which appear to possess certain credentials.

The first point links up with an aspect of our trolleyological methodology that we briefly discussed in Sect. 2.4.3. We noted that one of the advantages of trolley cases is that they permit us to ignore certain varieties of consequentialism. The reason for this, we said, is that the particular design of trolley cases takes away the motivation for considering certain distinctions. As we discussed, the differentia-

[47] We should, therefore, not confuse Maximization and Optimization. For the distinction, see Sen (1997, 746). On this point, see also Sect. 4.1.1.

[48] As we will see below, our argument against consequentialism is a further case in point. It illustrates that individual components of consequentialist moral theories are usually neither necessary nor sufficient conditions for an objection to apply.

tion between Subjective Consequentialism and Objective Consequentialism, e.g., becomes pointless. On the former, whether an act is right depends, roughly, on the consequences that the agent expects. On the latter, the actual outcomes matter. Plainly, then, if expected and actual consequences coincide, as they are assumed to do in trolley cases, the moral implications of objective and subjective variants of consequentialism will be the same if they are otherwise identical. Therefore, any valid trolleyological objection to the one form will also be effective against the other and vice versa, such that consequentialists cannot get any mileage out of the distinction. This means that we can eliminate from our investigation the respective determinable component in whose domain it lies. As I mentioned above, further distinctions may also lose their motivation in trolley cases. It seems, therefore, that we are well advised to investigate which distinctions appear unmotivated in trolley cases and which variants of consequentialism are, hence, irrelevant to our inquiry. This, I hope, will enable us to curtail the scope of our investigation vastly.

If successful, our first measure will allow us to eliminate some of the determinable components from our investigation. Our second move focuses on the alternative components that lie in the domains of the remaining determinable components. Logically speaking, there are innumerable alternatives to each of the CU-components. But, presumably, not all of these alternatives deserve our attention to the same extent. First of all, there will be some which we can ignore in an investigation that uses trolley cases. We can put them aside. Secondly, we can confine our attention to those determinate components that seem to have something convincing to be said in their favour. It is reasonable to assume that almost all variants of consequentialism (and certain other moral theories, too) are motivated by problems that attach to CU and can be seen as attempts to remedy these faults (cf., e.g., Tännsjö 2002, 15). If this much is true, we should be able to detect the most interesting variants of consequentialism, as we go through significant problems that have traditionally been associated with CU and look for alternatives to the CU-components that would solve these problems.

Once we have completed these two steps, we can start putting together our case against consequentialism. Presumably, however, we will still not be able to look at all possible combinations of components one by one. Therefore, I propose that we work, as it were, *from the inside out*. We start by refuting CU, which is at the center of the consequentialist family. We do this by introducing an initial case, $Case_0$, which shows that CU sins against some of our provisional moral fixed points. After that, we work our way outwards. As we shall discuss below, this method is not without problems. Before we turn to that issue, however, let us examine the procedure in more detail.

What does it mean to make a case against consequentialism from the inside out? As I just said, we start by formulating an initial case, $Case_0$, which shows that CU clashes with a provisional fixed point in our moral mindset and is, hence, unacceptable. After that, we formulate an objection against consequentialism that generalizes our finding. We suggest that our initial $Case_0$ does not only show that CU is flawed. Since consequentialist theories share a family resemblance with one another and since CU is the paradigmatic member of the consequentialist family, it

can be expected, we reason, that $Case_0$ discredits *all* consequentialist theories. To be sure, this allegation is most certainly false. However, it is useful. For it raises the question how consequentialists can defend themselves against it. Obviously, they have to explain *why* our allegation is false. To this end, they have to demonstrate that it is possible to get rid of the problem that $Case_0$ illustrates. To do this, they have to modify CU. They have to point to an alternative component to one of the CU-components which, when combined with the other logical building blocks of CU, leads to an appropriate answer in the initial case. There are presumably many ways to achieve this goal. In a second step of the argument, we have to find out which options there are. Once we have completed this step, we end up with a number of possible replies that consequentialists can give to rebut our objection. One reply may be to drop the CU-component C_{i1} in favour of C_{ij}. Another reply may be to drop C_{k1} in favour of C_{kl} and so on. That means branches will 'grow' out of our initial case, $Case_0$, such that a tree-like structure unfolds.

To show that we should reasonably reject consequentialism, we have to show that each of the possible replies $1, \ldots, m$ to our initial case is problematic on its own. We show this, once again, by working from the inside out. Suppose, e.g., that one possible reply to our initial objection is that consequentialists should endorse C_{ij} rather than C_{i1}. To show that this particular response is unsatisfactory, we look at the paradigm case of a doctrine which contains C_{ij}, viz. the conjunction of the components $C_{11}, \ldots, C_{ij}, \ldots, C_{n1}$. Then, we introduce a case which shows that this doctrine also collides with a provisional fixed point. And we generalize our finding, as we did with the initial objection. We suggest that the complaint affects all consequentialist theories which comprise C_{ij}. Consequentialists can perhaps rebut this objection. Maybe they can show that it is possible to modify $C_{11}, \ldots, C_{ij}, \ldots, C_{n1}$ by dropping one of its components in favour of another component. One such possibility may be to retreat from C_{p1} to C_{pq}. Another may be to drop C_{r1} in favour of C_{rs} and so on. In other words, there may be subsidiary branches which grow out of each branch. In each of these sub-branches, we proceed in the same way, i.e. from the inside out. We formulate an objection to the respective paradigmatic doctrine. Then, we generalize the objection and look for possible replies on the part of the consequentialist.

Now, it may seem like this process may carry on indefinitely. However, that is not the case. At each node in the tree, the noose tightens around the consequentialist's neck. This is because every time the consequentialist gives an answer, she incurs a commitment to a particular determinate component in the kit. To illustrate, let us suppose, once again, that one possible reply to our initial case, $Case_0$, is to retreat from C_{i1} to C_{ij}. In that case, we said, we proceed by formulating a case-based objection which discredits $C_{11}, \ldots, C_{ij}, \ldots, C_{n1}$. Then, we generalize that objection by suggesting that all consequentialist doctrines which contain C_{ij} as a component are affected by the same problem. And we look for possible replies that consequentialists might give. Let us suppose, then, that it is possible for consequentialists to rebut our case against that doctrine by dropping C_{p1} in favour of

C_{pq}. In that case, we have to proceed by arguing against $C_{11}, \ldots, C_{ij}, \ldots, C_{pq}, \ldots, C_{n1}$ and so on. Now, should it turn out that, somewhere down the road, we get to a point where consequentialists cannot give a reply to an objection or the only possible response to a complaint involves dropping C_{ij} (or C_{pq}), we have accomplished our goal. That is, we have shown that dropping C_{i1} in favour of C_{ij} (or C_{p1} in favour of C_{pq}) was, in fact, never a good reply. In that case, we can check off the respective branch or sub-branch that we are in and go to the next. We keep doing this until we have checked off all branches and sub-branches.

In summary, our methodic procedure can be stated thus:

Family Resemblance Approach – Version 2 ($\mathbf{FRA_2}$)

(i) Factorize CU into its logical components, $C_{11}, \ldots, C_{n1}$.
(ii) Work out which determinable components $C_{i(\text{-})}$, $C_{j(\text{-})}$, ... can be put aside in a trolleyological investigation.
(iii) Examine the typical motivations for adopting an alternative to the paradigmatic component of each of the remaining determinable components and take stock of the options that are available to consequentialists in light of these motivations.
(iv) Make a case against consequentialism *from the inside out*.

Before we conclude the exposition of our methodic approach, let me add a final note on the usage of cases within the inside-out procedure in step (iv). I said above that we start out by putting together *one* scenario which shows that CU violates certain provisional fixed points. It may prove useful, however, to augment the initial case, $Case_0$, against CU by further cases – $Case_{0*}$, $Case_{0**}$ and so on – that also target CU. The reason is simply that this may help us to save time. It may, after all, turn out that, though consequentialists can rebut $Case_0$ by diverging from a paradigmatic component, C_{i1}, to an alternative component, C_{ij}, they are unable to refute both $Case_0$ and $Case_{0*}$ simultaneously by doing so. In that case, we can discount the possibility that a consequentialist doctrine which contains component C_{ij} can be a satisfactory moral theory.

This said, we can address some criticisms that may be put forward against FRA_2. As I said earlier, it may be doubted that we can develop a convincing argument against consequentialism based on FRA_2. One may say that we are bound to miss certain determinable components. And it may be objected that it is simply impossible to take into account all alternative determinate components that might matter. This is certainly true. No matter what we will say below, it is bound to be incomplete. But this does not mean that our argument cannot achieve anything at all. I believe that a thorough investigation along the lines of FRA_2 can, at least, shift the burden of proof to the consequentialist. If we can show that it is not possible for consequentialists to dodge all objections by making modifications to their paradigmatic doctrine CU, we have at least established that it is reasonable to suspect that no consequentialist theory can do the trick. We can justifiably suspect this until consequentialists have come forward with a doctrine that demonstrably clears all the hurdles that we have set up.

A further criticism of FRA_2 is that an FRA_2-based argument will be, in a sense, weaker than an FRA_1-based argument, if the latter, in fact, worked. Here is why. Let us assume that our construction kit is complete in that we, in fact, identified all relevant determinate components. An FRA_1-based argument would show that a decisive objection exists against every determinate component in the domain of a given determinable component. Since every consequentialist theory must endorse at least one of these components, the argument would, hence, show that there cannot be a consequentialist doctrine which is not subject to severe criticism. Working through the steps of FRA_1 is as good as working through all versions of consequentialism. In contrast, the firm conclusion of an FRA_1-based argument is not warranted by an FRA_2-based argument. An FRA_2-based argument takes CU as a starting point. It attacks CU and then looks for possible replies on the part of the consequentialist that lie near CU. In each step of the argument, we consider only partial modifications of CU. E.g., we allow the consequentialist to retreat from CU, i.e. $C_{11}, \ldots, C_{i1}, \ldots, C_{n1}$, to $C_{11}, \ldots, C_{ij}, \ldots, C_{n1}$. We do not allow consequentialists to make two (or more) modifications at a time. We do not permit them, e.g., to suggest the doctrine $C_{11}, \ldots, C_{ij}, \ldots, C_{pq}, \ldots, C_{n1}$ in response to the initial objection that we make against CU. If we did that, FRA_2 would also be as good as going through all versions of consequentialism. However, this would come at the price of us actually having to go through all of them. This was precisely what we wanted to avoid to ensure tractability. Now, there are, of course, reasons why our discussion centers around CU. As we shall discuss in Sect. 4.1.4, all of the parts of the paradigmatic consequentialist doctrine CU seem initially attractive. It is not clear why a consequentialist would not want to stay true, as far as possible, to these elements of the classic doctrine. That is, it is not clear why she would want to abandon more of the paradigmatic components than she has to in each step of our argument. However, it may be, nevertheless, that there *is* a non-standard consequentialist doctrine which can avoid all our objections. Hence, intellectual honesty requires us to qualify our conclusion. An FRA_2-based argument can, by its very nature, do no more than shift the burden of proof to the consequentialist. It can merely establish that, by all appearances, there does not seem to be a consequentialist theory which can avoid our objections.

Nevertheless, I believe that FRA_2 is the best procedure available. FRA_2 is certainly better suited to address the Humpty Dumpty Defence (HDD) than DM. Like FRA_1 and unlike DM, FRA_2 acknowledges that consequentialists can diverge to alternative components to dodge objections. It advises us to examine the various alternatives to the CU-components and allows for case differentiations to take them into account. This is certainly a step in the right direction. Critical studies of consequentialism traditionally tended to focus almost exclusively on individual aspects of consequentialist doctrines, e.g., Maximization. They ignored, for the most part, important distinctions and options that are open to consequentialists, presumably because "it is rather fatiguing always to keep the full range of alternatives in play." (Sumner 1987, 175) Consequently, their arguments beg important questions against consequentialism. Though our case may also beg certain questions, it will, I hope, do so to a lesser extent than previous studies.

3.3 Summary

The aim of this chapter was to develop a method for criticizing consequentialism and to divide the remainder of the inquiry into a sensible sequence of steps. At the beginning of the chapter, we explained why this was necessary. Previously, we had already worked out an approach for criticizing moral theories, viz. PFPA. And it seemed as though we simply had to apply it to the consequentialist ones. But this reasoning overlooks the problem that delineating the class of consequentialist theories is no trivial matter. In the course of this chapter, we looked at two methods for doing this.

In the first section, we considered what seems to be the conventional approach. We called it the Definitional Method (DM). DM has two steps. In step (i), we look for a suitable definition of consequentialism. In step (ii), we show (by way of PFPA or otherwise) that all doctrines which come under that definition are objectionable and that we should, hence, reject them.

This approach, we noted, may fail in one of two ways. Firstly, it may be impossible to find a suitable definition of consequentialism. Secondly, it may be impossible to advance substantive criticisms based on that definition. As we saw, the first problem can be overcome. It is, in fact, possible to state a comprehensive Core Idea that includes all versions of consequentialism (minor exceptions aside). If it were possible to show that all doctrines which conform to this Core Idea are subject to a reasonable objection, this would *ipso facto* show that consequentialism should be rejected no matter how one defines it. But, as it turned out, this seems not to be possible. The Core Idea alone appears not to offer enough substance for criticism. To formulate substantive objections to consequentialism, we need to introduce additional assumptions. This, in turn, leads to the problem that consequentialists get the chance to defend themselves using a strategy that we called the Humpty Dumpty Defence (HDD). To apply it, they merely need to claim that the criticisms which objectors have formulated against their view rely on assumptions about the nature of consequentialism that they reject.

In the second section, we set out to develop a methodic approach which is immune to HDD. We called it the Family Resemblance Approach (FRA). The idea is that we can avoid the problem of HDD if we drop the assumption that consequentialist doctrines share a definitional feature and assume, instead, that the term "consequentialism" is a family resemblance term. The mark of such a family resemblance term, we explained, is that it applies to a class of objects which are unified by shared characteristics, though there is no *single* feature that all of the objects share in common. In a family, the relevant features are allowed, rather, to overlap and criss-cross. We noted, however, that the fact that we had previously identified a basic Core Idea, which lies behind all versions of consequentialism, seemed to rule out that we cold regard the class of consequentialist moral theories as a family. For it appeared to suggest that there was, in fact, one feature which lies behind all of these theories. As it turned out, though, the Core Idea relates not to a substantive element of consequentialist doctrines but, rather, to a structural trait

that these theories share. This, we found, is consistent with the assumption that they form a family. We concluded, therefore, that there seems to be no reason we should not view consequentialism as a family. Hence, we proceeded to investigate the methodological consequences of this idea for our case against consequentialism

To show that we should reasonably reject consequentialism *as a whole*, we reasoned that we would have to show that all members of the consequentialist family are defective. The first step of FRA should, therefore, consist in the delineation of the consequentialist family. To this end, we said, we should, first of all, investigate the logical structure of a paradigmatic consequentialist theory, viz. CU. We should, that is, factorize it into its logical components. As we recognized, the assumption that consequentialist theories form a family implies that there *must* be at least one alternative component to each of the paradigmatic components. So we said that we should, in a next step, investigate which alternative components there are to each of the paradigmatic components.

Upon completing the first two steps, we end up with a construction kit for consequentialist doctrines, as shown in Table 3.3 on page 72. The n paradigmatic components $C_{11}, \ldots, C_{n1}$ are shown in the first column. Alternative components $C_{i2}, C_{i3}, \ldots$ are shown to the right of the respective paradigmatic component C_{i1}. Our task is to demonstrate that all possible combinations of the elements in this kit give rise to objectionable doctrines. We noted, however, that going through all doctrines one by one will not be possible since there are simply too many of them. So be looked for potential shortcuts.

An initially attractive idea, we noticed, starts from the following observation. In the abstract, a consequentialist theory is the conjunction of n determinable components $C_{1(-)}, \ldots, C_{n(-)}$. That means every consequentialist theory has to adopt at least one determinate component, C_{ij}, for each of these determinable components, $C_{i(-)}$. If, therefore, it was possible to show that all alternatives $C_{i1}, C_{i2}, \ldots$ of a given row i were subject to at least one decisive objection $O_1, O_2, \ldots$, then this would show that all versions of consequentialism should be rejected. We reasoned, hence, that we should, in a next step, survey objections to consequentialism and attempt to put together a set of criticisms, such that there is at least one for each alternative of a given row in the construction kit. This was the logic behind the first version of FRA, which we called FRA_1. In concise form, it is stated on page 73.

As it turned out, however, FRA_1 is defective, as it naïvely assumes that individual objections to a moral theory can be attributed uniquely to its individual parts. Generally speaking, this does not seem to be the case. For this reason, we tried to formulate a second version of FRA which does not rely on that shaky assumption. We called it FRA_2. In concise form, it is stated on page 81. It contains a number of crucial steps. Like on FRA_1, step (i) consists in factorizing CU into logically independent components. The aim of step (ii) is to reduce the scope of the inquiry by investigating which determinable components we can ignore for the purpose of our trolleyological investigation. In step (iii), FRA_2 instructs us to consider the rationales for alternatives to the paradigmatic components and to take into account only those that seem motivated on that basis. Step (iv) of FRA_2 recommends, finally, to make a case against consequentialism *from the inside out*, that is, from a paradigm

case to the peripheral cases. To apply this procedure, we start by raising an objection against CU. Then, we generalize it. We claim that it does not only affect CU, but consequentialism more generally. After that, we investigate how a consequentialist might defend herself against this objection. That is, we sift through the construction kit to find alternatives to the paradigmatic components, $C_{11}, \ldots, C_{n1,}$ which consequentialists can endorse to rebut our assertion. We record the various possible replies. In a next step, we show that all of them are, in turn, objectionable. To this end, we also proceed from the inside out. E.g., consequentialists may say that they reject C_{i1} in favour of C_{ij}. In that case, we attempt to formulate an objection against $C_{11}, \ldots, C_{ij}, \ldots, C_{n1}$. After that, we generalize it. That is, we claim that all consequentialist doctrines which endorse C_{ij} are subject to that objection. Then, we look for possible answers. It may turn out that the challenge we raised against $C_{11}, \ldots, C_{ij}, \ldots, C_{n1}$ does not apply if C_{p1} is dropped in favour of C_{pq}. If that is the case, we formulate an objection to $C_{11}, \ldots, C_{ij}, \ldots, C_{pq}, \ldots, C_{n1}$. We claim, on that basis, that every consequentialist doctrine which endorses C_{ij} and C_{pq} is subject to that objection. We keep doing this until, finally, there is either no further reply available to the consequentialist or the only possible response involves dropping C_{ij} or C_{pq} to which she is already committed based on her previous answers. We work through all the branches and sub-branches that are opened up by the various replies until we have shown that all of them are ultimately unsatisfactory.

The bottom line of this chapter is, then, that our investigation can be divided up into steps (i) – (iv) of FRA_2. We will complete steps (i), (ii), and (iii) in Chap. 4. In Chap. 5, we will take on step (iv), which connects the dots.

Chapter 4
Consequentialism and Its Variants

In the previous chapter, we introduced FRA_2, which lays out four methodic steps for our case against consequentialism. In this chapter, we shall take on the first three.

In Sect. 4.1, we will focus on step (i). It requires that we factorize CU into a set of fundamental moral claims. If we are successful, we will thereby achieve two aims. First of all, we will uncover the logical structure which, as we assume, unites all consequentialist doctrines. Secondly, we will lay bare the fundamental moral claims that consequentialists paradigmatically endorse. This will give us a chance to explain, briefly, why the consequentialist perspective on morality is so attractive.[1] To see this, we will go through each of the claims that CU asserts and consider what one may say to support them.

In Sect. 4.2, we will first take on step (ii). We will identify determinable components that we can eliminate from our investigation. To this end, we will examine a number of CU-components and investigate under which conditions it is motivated to distinguish versions of consequentialism which depart from them. In each case, we will show that these circumstances do not obtain in trolley cases. This, in turn, will allow us to conclude that we may ignore, for the purpose of our investigation, the determinable components to which the respective CU-components belong. After that, we will tackle step (iii) of our methodical procedure. That is, we will investigate which variants of consequentialism we should consider in our argument. To this end, we will look at the CU-components of those determinable components that we did not eliminate in step (ii). In each case, we will ask, first, what would motivate consequentialists to depart from them and, then, which alternative components are eligible in light of these motivations.

In Sect. 4.3, we will sum up the main findings of the chapter.

[1] Even critics admit that consequentialism initially appears to be a very attractive moral view. Julian Nida-Rümelin (1993, 1), e.g., remarks that consequentialism appears to be "trivially true." Foot (1985, 198) acknowledges its "spellbinding force."

N. Mukerji, *The Case Against Consequentialism Reconsidered*,
Theory and Decision Library A: 51, DOI 10.1007/978-3-319-39249-3_4

4.1 Classic Utilitarianism

Before we begin our examination of CU it is important, first of all, to note what CU is and, perhaps even more importantly, what it is *not*. Its suggestive name notwithstanding, CU is not the doctrine held by classic utilitarian thinkers, such as Jeremy Bentham, John Stuart Mill, Henry Sidgwick, G.E. Moore, J.J.C. Smart, and others.[2] In their writings, these authors (implicitly or explicitly) describe their views in ways that are to a large extent inconsistent with our assumptions about the nature of moral theories.[3] E.g., we assumed that a moral doctrine is a function which attaches the moral properties *right*, *wrong*, *obligatory*, and *optional* to acts that are available to a moral agent. And we assumed that each of these properties is either present in or absent from it. In contrast, Jeremy Bentham and John Stuart Mill, e.g., seemed to understand rightness and wrongness as matters of degree. Bentham famously said about his *principle of utility* that it "approves or disapproves of every action whatsoever, according to the *tendency* which it appears to have to augment or diminish (. . .) happiness (. . .)." (Bentham 1838, 1; emphasis added, NM) And Mill maintained that "actions are right *in proportion as they tend to promote happiness* (. . .)." (Mill 1863, 9; emphasis added, NM) To both Bentham and Mill, moral properties were, hence, not all-or-nothing affairs. That means we could not represent their doctrines as "moral theories" in the sense in which we have introduced the term.

What is CU, then? And how far is it justified to call it by that name if it is not the doctrine held by classic utilitarians? In fact, the answer to the first question already gives us the answer to the second question. CU is a stylized theory which contains ideas that were widely embraced by classic utilitarian thinkers (though not necessarily in that particular combination) and whose structure is consistent with the assumptions about the nature of moral theories that we made above. The latter implies that we can represent CU's theoretical component as a criterion of rightness. This is precisely the reason we use CU instead of the theories that classic utilitarians have put forward.

4.1.1 The Theoretical Component

With this caveat in mind, we can introduce CU. Recall that we previously made a distinction between the theoretical component of a moral theory – i.e. its *criterion of rightness* – and its practical component, which is a decision-making procedure for moral choices. Let us first turn to CU's criterion of rightness.

[2] Rawls (1955, 9) also includes Hobbes and Hume amongst the classic utilitarians.

[3] For this reason, Fred Feldman remarks that "the doctrine discussed by Mill, Moore and Smart is, strictly speaking, incoherent." (Feldman 1986, 4)

Classic Utilitarian Criterion of Rightness
An act is right if and only if it maximizes the sum total of happiness of all sentient creatures.

Step (i) of our inquiry requires that we decompose this criterion into logically distinct components. There exist numerous attempts to do this in the literature. Since they were devised for different purposes, however, most of them are not well suited to meet our needs. Proposals, such as Sinnott-Armstrong's (2011), lack the required logical rigour.[4] Others are logically rigorous, but not sufficiently differentiated (e.g. Sen 1979). In what follows, we will draw, in large part, on a proposal that I have made elsewhere (cf. Mukerji 2013c).

It is common to divide the theoretical part of moral theories into two determinable components, viz. a conception of the right and a theory of the good (cf. Rawls 1971/1999, 21). We shall follow that convention and do the same. Before we start, however, it is necessary to make some qualifications.

Firstly, it is important to emphasize that a *conception* of the right is not the same as a *criterion* of rightness. The former merely tells us how the notions of right and good are connected. Without any knowledge of what is good, we cannot use it as a criterion for distinguishing between right and wrong acts. A criterion of rightness, in contrast, combines a conception of the right with a conception of the good and states necessary and sufficient conditions for the rightness of an act.[5]

Secondly, when we talk about CU's conception of the good, we should bear in mind that there is a distinction to be made between two senses of the term 'good.' The idea of the good can be related to the moral point of view. Alternatively, it may refer to the personal point of view of an individual. In the former case, the idea is that of the *moral* good.[6] In the latter case, it is that of

[4]Sinnott-Armstrong (2011) offers a characterisation of classic utilitarianism. He regards it as a combination of 11 claims that he calls: Consequentialism, Actual Consequentialism, Direct Consequentialism, Evaluative Consequentialism, Hedonism, Maximizing Consequentialism, Aggregative Consequentialism, Total Consequentialism, Universal Consequentialism, Equal Consideration, Agent-neutrality. The trouble with this scheme of claims is that they are not all logically independent, as Sinnott-Armstrong himself concedes.

[5]Note that certain moral theories do not possess a conception of the good (cf. Nida-Rümelin 1993, 65). They are exceptional in that their criteria of rightness do not establish a relation between the right and the good.

[6]Note that it is important to distinguish between at least three different notions that have all been referred to as the "moral good." William K. Frankena has said that the "sorts of things that may be morally good or bad are persons, groups of persons, traits of character, dispositions, emotions, motives, and intentions," whereas "[a]ll sorts of things (. . .) may be nonmorally good or bad, for example: physical objects like cars and paintings; experiences like pleasure, pain, knowledge, and freedom; and forms of government like democracy." (Frankena 1963/1973, 62) In saying this, Frankena emphasizes that *moral* goodness is connected with agency. Since things, unlike persons, are incapable of agency, they should not, according to Frankena, be categorized in moral terms (Essentially the same point is made by Harrod 1936, 140.). However, Frankena also acknowledges that it may make sense to call certain *things* morally good or bad when "we mean that it is morally right or wrong to pursue them." (Frankena 1963/1973, 62) Something can also be called "morally good" if the sense in which it is good is, as it were, "laced" with ideas that pertain to the sphere of

the *prudential* good.[7] In what follows, we shall often use the idea of the good without any qualification. When we do that, we shall mean the idea of the *moral* good.

Thirdly, it is important to distinguish between what is *intrinsically* (or ultimately) good and what is *extrinsically* (or instrumentally) good. Things that are intrinsically good are valuable in themselves, while things that are extrinsically good are good insofar as they are means to the achievement of certain intrinsic goods. When we use the word 'good' without qualification, we shall mean to refer to the good in the intrinsic sense.

With these clarificatory remarks in mind, let us turn to CU's conception of the right. CU is commonly said to embrace a *maximizing* conception.

Maximization
An act is right if and only if it maximizes the good.

Some points about Maximization are worth stressing.

Firstly, Maximization – in the sense in which *we* are using the term – is a specification or, as it were, a particular interpretation of what we have called the Core Idea of consequentialism, viz. that the rightness of an act depends only on its goodness. It specifies that whether an act is right or wrong depends only on whether it is *maximally* good.

Secondly, as we noted in Sect. 3.1.1, the status of being maximally good is not an internal property of an act. It is, rather, a fact that is true (or false) of an act in virtue of its relation to *other* acts. An act is maximally good if there is no alternative which is better.

Thirdly, as Sen (1997a, 746) emphasizes, Maximization should not be confused with Optimization.[8]

Optimization
An act is right if and only if it is at least as good as every alternative act.

The difference between Maximization and Optimization is subtle and often ignored. To understand it, it is necessary to look a bit more closely at the logic of betterness relations. To this end, consider a choice situation, S, which is constituted by n alternatives $a_1, \ldots, a_n \in A$, where A is the set of acts that are available to the agent. Let R represent the weak betterness relation "at least as good as" and P the strict betterness relation "better than." We can rephrase the claims of Maximization and Optimization using these formalisms. Maximization says that

---

the right, such as the idea of distributive justice (cf. Rawls 1971/1999, 22). The sense in which we use the term is the second.

[7] The relation between these two ideas of the good has been discussed in the context of the debate about the "overridingness thesis" concerning morality. The latter is roughly the claim that an act which is morally mandated is also rationally mandated. See, e.g., Stroud (1998).

[8] Gaertner (2006, 6–7) offers an accessible explanation of the difference between the "maximal set" and the "choice set" which essentially parallels our distinction between Maximization and Optimization.

an act, a_i, is morally permissible if and only if $\neg a_j\ Pa_i$, for all $a_{j\neq i}$. Optimization says that an act is morally permissible if and only if $a_i\ Ra_j$, for all $a_{j\neq i}$. Now, if R is *complete*, that is, if it is true for any two alternatives a_i, $a_j \in A$ that either $a_i\ Ra_j$ or $a_j\ Ra_i$, $\neg a_j\ Pa_i$ implies $a_i\ Ra_j$. In other words, given completeness of R, Maximization implies Optimization. If, however, it is not the case that either $a_i\ Ra_j$ or $a_j\ Ra_i$ – if, that is, R is incomplete – then $\neg a_j\ Pa_i$ does *not* imply $a_i\ Ra_j$ and the implications of Maximization and Optimization may come apart. Unlike Optimization, Maximization may permit an act, even if it is not the case that it is at least as good as any other. That is to say, it is a less stringent requirement for choice than Optimization. It allows for justified choices even when there is no perfect comparability of values.

Fourthly, Maximization establishes a connection with the idea of the *good*. However, it is logically independent of any particular account of the good. That is, one can accept Maximization and adopt or reject any particular conception of the good. The reverse is also true. No matter which conception of the good one endorses, one is free to accept or reject the view that rightness is exclusively an issue of maximizing goodness. The right and the good are logically independent.[9]

Fifthly, Maximization seems often to be conflated with consequentialism.[10] So it appears to be necessary to clarify the relationship between the two. The obvious point to be made is that Maximization and consequentialism are not identical, first and foremost, because they are different kinds of *things*. The former is a component of moral theories, while the latter is, as we construed it, a family of doctrines.[11] Furthermore, it is not the case that a moral theory is a member of the consequentialist family if and only if it is maximizing. Maximization implies consequentialism.[12] However, the fact that a moral theory is not maximizing does not allow us to conclude that it is not a form of consequentialism. That is, one can accept the Core Idea of consequentialism, viz. that rightness is only a matter of goodness, but reject the notion that an act is right only if it *maximizes* the good. In that case, one would

[9]It should be noted, however, that the conception of the right imposes a domain restriction on admissible conceptions of the good. E.g., we have to exclude a conception of the good which says that it is good to do what is not right.

[10]Maximization is commonly seen as an essential feature of all utilitarian doctrines, in particular, and consequentialist doctrines in general. It also figures prominently in most attempts to define consequentialism. See, e.g., Arneson (2004, 34), Nida-Rümelin (1993, 87), Scheffler (1982/1994, 1), and Williams (1973, 85).

[11]Consequentialism may, however, also be seen as a component of a moral theory. See Mukerji (2013c, 298).

[12]It may be objected that non-consequentialists can hold maximizing views as well (cf. Lawlor 2009b, 24; Nida-Rümelin 2005). In some sense, this is true. Consider, e.g., the non-consequentialist doctrine put forward by William David Ross. He held that rightness is a matter of maximizing the balance of right-making features ("*prima facie* duties") over wrong making-features. However, the *value* that ought to be maximized, according to theorists such as Ross, is not goodness but something else (we may call it "prima facie" duty). Hence, they would come out as non-maximizing moral theories on our conceptual stipulations.

be a non-maximizing consequentialist. We will examine some variants of this kind of consequentialism in Sect. 4.2.2. below.

Having made these points about Maximization, let us move on to CU's theory of goodness.

Classic Utilitarian Theory of Goodness (CUG)
The goodness of an act is measured by its impact on the sum total of happiness of all sentient creatures.

CUG, in turn, can be factorized into further subsidiary components. Note that it makes the goodness of an act depend only on individual well-being, viz. happiness. I propose to call this claim Welfarism. A more precise formulation is as follows.

Welfarism (Informal Version)
Goodness is identified with a functional that takes as its arguments only the appropriately weighted numerical representations of the well-being of the morally relevant individuals.

This sounds rather cumbersome. However, it is, in fact, quite a straightforward idea. This gets clear once we avail ourselves of some simple formalisms. Let G represent intrinsic goodness. And let W stand for a functional[13] that takes as its arguments only $u_1, \ldots, u_n$ and $w_1, \ldots, w_n$. $u_1, \ldots, u_n$ refer to functions that numerically represent the well-being of the morally relevant individuals $1, \ldots, n$.[14] $w_1, \ldots, w_n$ represent weights that are attached to the respective functions. Welfarism, as we just stated it, can then be expressed thus:

Welfarism (Formal Version)

$$G = W = f(w_1, \ldots, w_n; u_1, \ldots, u_n)$$

In the interest of clarity, we should qualify what this claim means and how it relates to some of the claims that have been called "welfarism" (with a lowercase "w").

Note, firstly, that welfarism was not originally intended as a philosophical idea. Initially, it was used in economics. Only after that was it introduced into the moral-philosophical debate by Sen (1979). Unsurprisingly, therefore, the term has come to be used in a myriad of ways.

Furthermore, it is often used rather loosely. Various ideas are run together under its name.[15] In what follows, we will try to disentangle them. Before that, however,

[13] Functionals are functions that take functions as their arguments.

[14] A more precise interpretation of $u_1, \ldots, u_n$ is as follows. We assume that there is an ordering R_i for every individual $1, \ldots, n$. R_i ranks acts any two acts, a_j and a_k, from a set of acts, A, according to the well-being that they produce for individual i. This ordering is a *complete* (i.e. either $a_j\ R_i\ a_k$ or $a_k\ R_i\ a_j$), *transitive* (i.e. if $a_j\ R_i\ a_k$ and $a_k\ R_i\ a_l$, then $a_j\ R_i\ a_l$), and *reflexive* (i.e. $a_j\ R_i\ a_j$) relation and can, hence, be represented numerically by a function u_i which attaches greater numbers to acts that are ranked higher in the ordering and lower numbers to acts that are ranked lower in the ordering. That is, $u_i(a_j) \geq u_i(a_k)$ if and only if $a_j\ R_i\ a_k$.

[15] Holtug (2003) and Rechenauer (2003), e.g., have emphasized this.

let us look at what seems to be a common denominator of all notions of welfarism (including Welfarism).

On all interpretations, welfarism is *not* the view that the well-being of individuals is morally relevant. It is, rather, the considerably stronger idea that it is the *only* thing that matters. Moore and Crisp (1996, 598) call this the "Exclusiveness Thesis" of welfarism. While no serious thinker would disagree with the former idea – after all, well-being is of course important! – the Exclusiveness Thesis that all forms of welfarism (including Welfarism) imply is a contestable normative claim and has, in fact, been contested by various authors.[16]

Secondly, a distinction should be made between two broad senses in which we may use the term "welfarism." When Sen (1979) introduced it to the moral-philosophical discussion, he intended it as an axiological idea. That is, he employed the term "welfarism" to refer to an idea about *goodness*. This has also been the practice of a number of authors since (e.g. Blackorby et al. 1984, Holtug 2003, Moore and Crisp 1996, Ng 1981, 1990, Rechenauer 2003, and Sumner 1996). Other authors, such as Keller (2009) and Shaver (1999), have, however, preferred to use the predicate "welfarist" in a deontic sense. That is, they have employed it to label (the theoretical component of) whole moral theories. We should distinguish, then, between the following two notions.

Axiological Welfarism
Goodness is exclusively a function of individual well-being.

Deontic Welfarism
The moral status of an act depends exclusively on its effects on the well-being of individuals.

As we defined Welfarism, it is an idea about goodness and falls, hence, on the side of Axiological Welfarism.

Thirdly, a distinction should be made between Philosophical Welfarism and Technical Welfarism, as Rechenauer (2003) points out.[17] Philosophical Welfarism, I take it, can be seen as the conjunction of two theoretical commitments. The first commitment is normative. It is the idea that goodness can be represented by a functional f which takes as its arguments only the representations of the well-being $u_1, \ldots, u_n$ of the morally relevant individuals $1, \ldots, n$. The second commitment is one of "descriptive adequacy."[18] (Sumner 1996, 10) It forbids us to spell out the

[16] For influential critiques see, e.g., Sen (1979) and Taylor (1995, 127–145).

[17] Rechenauer (2003) credits Fleurbaey (1996) with this distinction. In our characterization of it, we depart from Rechenauer's ideas. Related differentiations can be found in the literature, e.g. in Tungodden and Vallentyne (2007). In the context of justice theory, they distinguish between welfarism and benefitism. Benefitism is, roughly, the view that the justice of a distribution depends only on the benefits of the individuals, where the idea of a benefit is apparently intended to be broader than that of well-being. The distinction between welfarism and benefitism should, hence, roughly match up with our distinction between Philosophical and Technical Welfarism.

[18] In fact, the concept of well-being does not seem to be purely descriptive either. Rather, it seems to be one of the many thick ethical concepts that "express a union of fact and value." (Williams 1985, 129) The reason why it contains normative and descriptive elements is that we would not judge that a person has well-being if she is in a state that does not seem to be worth having.

notion of well-being in just about any way we like and demands that we stay true to the vernacular use of the term "well-being." In and of itself, it is not clear what this criterion entails.[19] But there are obvious candidates for welfarist theories in the philosophical sense, viz. doctrines which contain notions of well-being that center around sensory pleasure or accounts that regard desire-fulfilment as central to well-being.

Philosophical Welfarism

(1) Goodness can be represented by a functional f that takes as its arguments only the representations of well-being $u_1, \ldots, u_n$ of the morally relevant individuals $1, \ldots, n$.
(2) The idea of well-being has to be spelled out in a way that is *descriptively adequate*.

Technical Welfarism, in contrast, is a "term of art" that, unlike Philosophical Welfarism, is not meant to capture a vernacular concept. It drops (2) and endorses only (1).

Technical Welfarism (Informal Version)
Goodness can be represented by a functional f that takes as its arguments only the representations of well-being $u_1, \ldots, u_n$ of the morally relevant individuals $1, \ldots, n$.

This characterization of the distinction makes Technical Welfarism the more general and encompassing idea. On this account, all forms of Philosophical Welfarism are, therefore, forms of Technical Welfarism, but not *vice versa*. That is, one can be a technical welfarist without being a philosophical welfarist, but not the other way around.[20] Formally, Technical Welfarism can be expressed as the following idea.

Technical Welfarism (Formal Version)

$$G = W = f(u_1, \ldots, u_n)$$

This formalization of Technical Welfarism looks very similar to the formalization of Welfarism that we introduced above. The only difference is that, on Welfarism, f has some additional arguments – viz. $w_1, \ldots, w_n$ – that are missing from f on Technical Welfarism. We will come back to that point further below. At this stage, let us focus on the consequences that the similarity to Technical Welfarism has for the character of Welfarism. Note that, like Technical Welfarism, Welfarism does not impose any restrictions on the idea of individual well-being. It is, hence, compatible with many views we commonly regard as contrary to Philosophical Welfarism. It is possible, e.g., to combine it with a resource-based conception of welfare like the

[19] For an extensive treatment of this question, see Sumner (1996).

[20] On Rechenauer's account, this does not hold. Though he emphasizes that an adherent of Technical Welfarism need not be a proponent of Philosophical Welfarism, he does not say that it is impossible to be a philosophical welfarist without being a technical welfarist. He merely says that "if you are a philosophical welfarist, you'd better be a technical welfarist as well." (Rechenauer 2003, 9) The reason for this is that Rechenauer understands by Technical Welfarism what we shall call Narrow Technical Welfarism (NTW). NTW is a narrower version of Technical Welfarism that a philosophical welfarist, in fact, does not need to accept.

primary goods index that is proposed by Rawls (1971/1999). And it is possible to combine it with a capability approach *à la* Sen (1985).[21]

Fourthly, let us focus on the nature of the functional *f*. Technical Welfarism is usually interpreted as imposing a set of specific restrictions on *f*. Sen (1979) mentions that the function has to be *increasing* in its arguments. That is, it has to be such that any increase in the well-being of a single individual will *ceteris paribus* increase the value of the function. As Nils Holtug hypothesizes, Sen's reason for imposing this condition may have been his suspicion that any version of Technical Welfarism whose *f* is not increasing in its arguments would be implausible. However, as Holtug points out, "the issue of whether some view is plausible and the issue of whether it is a form of [technical; NM] welfarism are distinct." And he goes on to say that "lack of plausibility does not explain why [technical; NM] welfarism should rule out non-increasing functions." (Holtug 2003, 160) I think Holtug is right about this. In and of itself, it is not clear why Technical Welfarism should rule out a non-increasing *f*. I believe that when one wants to talk about Technical Welfarism in this confined sense, one should make clear that one speaks about a version of the following view.

Narrow Technical Welfarism (NTW)
f is subject to a set of specific conditions *C*, *D*, *E*, ...[22]

In contrast, Wide Technical Welfarism can be characterized as the view that *f* is not subject to these constraints. It does not impose all of the conditions *C, D, E,* ... on the nature of the functional *f*.[23]

Wide Technical Welfarism (WTW)
f is not subject to any conditions.[24]

Our Welfarism is more similar to Wide Technical Welfarism in that it does not impose any conditions on *f* either.

Sixthly, and lastly, let us investigate which kinds of morally relevant goods Welfarism can acknowledge and how it differs in that respect from what usually goes by the name 'welfarism'. On welfarism, it is often said, individual well-being is the only good (cf. Sumner 1996, 185). However, this is incorrect. Welfarism can

[21] These points are also made by Rechenauer (2003, 8).

[22] On these conditions, see, in particular, Sen (1977, 1552–1553). Characterizations of Narrow Technical Welfarism can also be found in Blackorby et al. (1984, 329–331), Gaertner (2006, 29), and Rechenauer (2003, 6–7).

[23] This wide view of Technical Welfarism is not at all uncommon. In various works, Amartya Sen has defined welfarism in a way that is congenial to it. He says, e.g., that welfarism is the "general approach of making no use of any information about the social states other than that of personal welfares generated in them." (Sen 1977, 1559) Elsewhere he says that welfarism is the view "that the goodness of a state of affairs be a function only of the utility information regarding that state." (Sen 1987/2004, 39)

[24] I should point out that it is not certain that a representing function exists if the respective conditions *C, D, E* ... are not imposed on *f* (cf. D'Aspremont and Gevers 1977). I am indebted to Martin Rechenauer, who made me aware of this fact.

acknowledge other goods over and above individual well-being.[25] The belief that it cannot is due in large part, I believe, to the false assumption that welfarist theories have to accept the following claim.

Person-Affecting Restriction
Nothing can be good (or bad) without being good (or bad) for someone.[26]

If a theory accepts the Person-Affecting Restriction, it accepts Welfarism. However, this does not hold *vice versa*. While certain welfarist theories of the good endorse this restriction (call them *person-affecting* theories), some reject it (call them *impersonal* theories). The former do, in fact, rule out that there are any goods over and above individual well-being. The latter, however, are compatible with the idea that there are further goods, viz. "pattern goods ," as Broome (2004, 44) calls them. As their name suggests, pattern goods are generated by the pattern of a well-being distribution throughout individuals (and/or time). One such good is, e.g., equality. Those who believe in the good of equality think that it is in itself good if well-being is distributed equally throughout individuals. Note, however, that this violates the Person-Affecting Restriction since equality of well-being is not good for anyone in particular. (What is good (or bad) for an individual is, rather, her well-being (or lack thereof)). Therefore, person-affecting welfarist theories rule it out as a good. However, impersonal views can accept it because they are not constrained by the Person-Affecting Restriction. They can acknowledge goods that do not benefit anyone in particular, e.g. equality. As it turns out, then, welfarist theories can recognize that there are other goods over and above individual well-being. This holds, in particular, for moral views that adopt Welfarism.

In fact, Welfarism, as we defined it, is not only compatible with pattern goods. It can also acknowledge further goods that are commonly regarded as incompatible

[25]Two obvious points can, of course, be made in this context. Firstly, welfarism can acknowledge instrumental goods that are good insofar as they are conducive to well-being. Consider the value of friendship. It is valuable, on any welfarist theory of the good, because it usually produces well-being for the parties that are involved in it. Secondly, welfarism is compatible with the existence of further intrinsic goods besides well-being. As Sen (1980–1981) points out, on some accounts, well-being has multiple component parts. On any such plural account of well-being, welfarism does not rule out, then, that there are other values besides well-being, viz. its constituent parts.

[26]To my knowledge, the basic idea behind the Person-Affecting Restriction was first formulated by Jan Narveson. He uses it to characterize utilitarianism. In particular, he says that "[i]n deciding what we are to do, the only consideration which is morally relevant, according to utilitarianism, is how others would be affected." (Narveson 1967, 63) The Person-Affecting Restricting subsequently received its name from Glover (1977, 66). Influential discussions can be found, e.g., in Broome (2004, 136, 145), Parfit (1986, Ch.18), and Temkin (1993, Ch.9). Temkin refers to it as "The Slogan." It should be noted that there are various interpretations of the Person-Affecting Restriction (cf. Roberts and Wasserman 2009). Most importantly, we should keep apart a deontic version of the idea which says that "what makes an act or choice morally impermissible, or wrong, must be connected in some central way with a person's having been made *worse off*" (Roberts and Wasserman 2009, xiv) and an axiological version which says that "one outcome is better than another only if there is some person for whom it is better." (Roberts and Wasserman 2009, xxxv) We use the idea in the second way.

with welfarism, e.g. the value of moral desert (cf. Holtug 2003, 158). The idea behind moral desert is, roughly, that it is good if those who deserve well-being do, in fact, fare better than those who do not. An obvious way to model such a concern is to attach greater weight to the well-being of those who are deserving and a smaller (or perhaps even a negative) weight to those who are not. This is incompatible with welfarism because, on welfarism, intrinsic goodness is a functional f only of $u_1, \ldots, u_n$. The idea that weights $w_1, \ldots, w_n$ play a part is, therefore, ruled out.[27] Note, however, that Welfarism, as we stated it above, says that intrinsic goodness is a functional f of $u_1, \ldots u_n$ and of $w_1, \ldots, w_n$. That is, it allows weights and is, hence, even broader than welfarism is generally taken to be. In addition to pattern goods, it can, hence, take on board goods which manifest themselves in differential weights. This, of course, makes our idea of Welfarism very encompassing. And it raises the question what it actually excludes. So let us turn to that question now.

Welfarism, recall, is the view that intrinsic goodness is equal to overall welfare; formally, $G = W$, where $W = f(u_1, \ldots, u_n; w_1, \ldots, w_n)$. The negation of Welfarism, Non-Welfarism, just negates the identity between G and W.

Non-Welfarism

$$G \neq W$$

There are obviously two kinds of Non-Welfarism, then. One kind asserts that G is a functional h of W and further factors $X, Y, \ldots$. This view may be called "Extra-Welfarism" (Culyer 1989, 36).

Extra-Welfarism

$$G = h(W, X, Y, \ldots)$$

The other kind of Non-Welfarism asserts that G does not depend on W at all, but is exclusively a functional of other factors $X, Y, \ldots$. It may, therefore, be called "Anti-Welfarism." (Rechenauer 2003, 13)

Anti-Welfarism

$$G = h(X, Y, \ldots)$$

As for the factors $X, Y, \ldots$, which plausible candidates are there? Throughout the ages, many philosophers have proposed objects of intrinsic value. Perhaps the most comprehensive list of factors is given by William Frankena. He mentions

> life, consciousness, and activity; health and strength; pleasures and satisfactions of all or certain kinds; happiness, beatitude, contentment, etc.; truth; knowledge and true opinions of various kinds, understanding, wisdom; beauty, harmony, proportion in objects contem-

[27] Note, however, that weights $w_1, \ldots, w_n$ may play a part if they can be construed as a functional g of $u_1, \ldots, u_n$. In Sect. 4.2.3.2, we will come across a view called Prioritarianism which determines weights on that basis.

> plated; aesthetic experience; morally good dispositions or virtues; mutual affection, love, friendship, cooperation; just distribution of goods and evils; harmony and proportion in one's own life; power and experiences of achievement; self-expression; freedom; peace, security; adventure and novelty; and good reputation, honor, esteem, etc. (Frankena 1963/1973, 87–88)

In Sect. 4.2.4, we will come back to this list when we talk about alternatives to Welfarism.

Before we move on, it is important to note that Welfarism, as we understand it, is logically independent of the particular nature of W. We can combine it with any W, as can both versions of Non-Welfarism.[28] Likewise, any particular version of W can be coupled with either Welfarism or Non-Welfarism. That means we can study the character of the classic utilitarian interpretation of W independently of the doctrine's property of being a welfarist moral theory.

Let us ask, then, what the classic utilitarian version of W contains? As we said above, W is computed from a functional f that takes the appropriately weighted well-being levels of individuals, which are associated with a given act, as arguments and spits out a value of overall well-being. Or, formally

$$W = f(u_1, \ldots, u_n; \; w_1, \; \ldots, \; w_n) \tag{1}$$

We can, hence, characterize CU's interpretation of W by answering four questions (cf. Mukerji 2013c, 298).

- Firstly, what is the nature of the functional f that combines $u_1, \ldots, u_n$ and $w_1, \ldots, w_n$ into a single value?
- Secondly, which weights $w_1, \ldots, w_n$ should be attached to the well-being $u_1, \ldots, u_n$ of individuals $1, \ldots, n$ *vis-a-vis* each other?
- Thirdly, who are the morally relevant creatures $1, \ldots, n$? That is, whose well-being counts?
- Fourthly, how should the idea of individual well-being be interpreted? That is, what is represented by $u_1, \ldots, u_n$?

CU's answer to the first question is the principle of Summation.

Summation
Overall well-being is measured by the sum of the appropriately weighted levels of well-being of all morally relevant individuals.[29]

[28] Even anti-welfarists can adopt a version of W. Since they hold, however, that W does not affect G, the nature of W does not have any impact within an anti-welfarist moral conception.

[29] Summation implies that acts can be ranked *vis-a-vis* each other regarding their goodness. According to Summation, an act, a_i, is better than another, a_j, if and only if a_i brings about a greater sum total of happiness than a_j. Amartya Sen refers to this idea as "sum-ranking." (Sen 1979, 468)

Given Summation, W gets the following shape.

$$W = w_1u_1 + \cdots + w_nu_n \tag{2}$$

CU's answer to the second question, viz. which weights should be attached to the well-being of the morally relevant individuals, is the idea of Equal Treatment.

Equal Treatment
The well-being of all morally relevant individuals is to be weighted equally.

Formally, Equal Treatment can be expressed as the claim that $w_1 = \ldots = w_n = 1$. When applied to (2), the weights simply drop out.

$$W = u_1 + \cdots + \ u_n \tag{3}$$

CU's answers to the third and fourth question do not alter the shape of (3). Rather, they give us an interpretation of it. The third question asks whose well-being matters. Or, to put it more formally, it asks what the indices $1, \ldots, n$ refer to. CU's answer is Universalism.

Universalism
The well-being of all individuals capable of enjoying well-being (in the relevant sense) should be taken into account.[30]

The fourth question, finally, asks what u_1, u_2, ... and u_n refer to. CU's criterion of rightness, as we stated it above, says that an act is right if and only if it maximizes *happiness*. This suggests that well-being should be interpreted as happiness. Now, there are two interpretations of the term "happiness" that are in frequent philosophical use (cf. Haybron 2011). An individual may be happy in the sense that she has pleasurable sensations and no unpleasant (or painful) sensations. Her happiness may then be called *sensory* happiness. She may also be happy in a more encompassing sense, viz. in so far as her life as a whole is going well. In that case, her happiness should, I believe, be called *life* happiness.[31] The happiness

[30]We should make two points about Universalism. Firstly, different moral claims carry the label "Universalism." If a moral doctrine subscribes to Universalism in an agent-related sense, it says that moral commands are addressed to all agents in a particular class (cf., e.g., Pettit 2000, 177). If a doctrine is universalistic in a patient-related sense, it says that all individuals in a particular class matter morally. Finally, if a doctrine holds Universalism in a principle-related sense, it holds that the moral principles hold universally, i.e. with regard to all moral choice problems. The intended sense here is the patient-related one. Secondly, Universalism (in this patient-related sense) presupposes that the idea of a morally relevant subject is discrete rather than graded. If a being has moral standing, it has it to the same extent as every other morally relevant subject. (This, to be sure, is not to say that its well-being is attached the same weight.) It might be more fruitful, however, to think of moral standing as coming in degrees (cf. Vallentyne 2007), particularly when it comes to animal ethics. However, since we do not address this issue in this inquiry, the assumption that moral standing is discrete is, I believe, rather unproblematic.

[31]It is easy to see that the two senses of happiness are independent to a large extent. A person can be happy in the sense that she feels good, even though we would not be inclined to say that her

concept that is relevant in the context of CU is the former, viz. sensory happiness. Traditionally, utilitarians have interpreted well-being as the experience of pleasure. This idea may be called "Welfare Hedonism." (cf. Kagan 1998, 31)

Welfare Hedonism
The well-being of an individual is measured by her sensory happiness, i.e. the balance of her pleasures over her pains.

Obviously, all components of CU's version of *W* are logically independent. That is, we can combine any three with the negation of the fourth one and get a non-contradictory combination. Firstly, one can accept Equal Treatment, Universalism, and Welfare Hedonism and drop Summation. One might, e.g., adopt the following view.

Multiplication
Overall well-being is the *product* (rather than the sum) of the appropriately weighted levels of well-being of all morally relevant individuals.

If we apply Multiplication to (1), we get

$$W = w_1u_1 \cdot \cdots \cdot w_nu_n \tag{4}$$

which is a logical possibility.

Secondly, we can accept Summation, Universalism, and Welfare Hedonism and drop Equal Treatment in favour of some version of Unequal Treatment.

Unequal Treatment
It is not the case that the well-being of all morally relevant individuals is to be weighted equally.

When we combine this assumption with (2) we get

$$W = w\prime_1u_1 + \ \ldots \ + w\prime_nu_n \tag{5}$$

where $\neg(w'_1 = \ldots = w'_n)$.

Thirdly, if we accept Summation, Equal Treatment, and Welfare Hedonism, nothing keeps us from rejecting Universalism in favour of some form of Partialism.

Partialism
It is not the case that the well-being of all individuals capable of enjoying well-being in the relevant sense should be taken into account.

---

life is going well for her. A severe drug addict who has just gotten her daily fix is an example of that. Similarly, a person whose life seems to be going well overall need not be happy all the time. Such a person may not always be happy in the sense that she feels good. So happiness in the latter sense does not seem to entail happiness in the former sense. It seems to be true, though, that on any plausible account of life happiness, a person cannot be happy unless she is to some extent happy in the sensory sense. Life happiness seems to entail a certain degree of sensory happiness, then, though the reverse is not true.

I may, e.g., adopt Egocentrism, viz. the view that my well-being alone determines the value of W.[32] Combining this idea with (2) yields

$$W = \text{my well-being} \tag{6}$$

Finally, since Welfare Hedonism concerns merely the interpretation of $u_1, \ldots, u_n$ we can, of course, combine any non-hedonistic view about well-being with Summation, Equal Treatment, and Universalism. The four claims are, hence, logically independent.

Having said this, a few clarifying comments are again in order. Firstly, we should re-emphasize what we said above, viz. that the fact that these four components of CU's version of W are logically independent does not imply that the determinable components to which they belong are also logically independent. As we said above, the choice of the functional f places a domain restriction on conceptions of individual well-being. If one accepts Summation, e.g., this requires that one take a view of well-being that allows it to be measurable in units and comparable throughout individuals since Summation makes the presupposition that there *exists* something that can be summed. Conversely, the choice of a conception of well-being may place a domain restriction on admissible functions. Evidently, if one accepts a view of individual well-being which implies that interpersonal comparisons are not meaningful, one cannot subscribe to Summation as a component.

Secondly, it should be noted that Universalism has no determinate content unless we combine it with a conception of well-being. It says that the well-being of all individuals who are capable of enjoying it *in the relevant sense* matters. The italicized phrase points to the conception of well-being. If we combine Universalism with Welfare Hedonism, as CU does, we get

Hedonistic Universalism
The well-being of all sentient creatures should be taken into account.

Thirdly, it is important not to confuse the claims that are comprised in the classic utilitarian interpretation of W with similar looking views. E.g., many views have been called by the name "hedonism." One example is the following idea.

Psychological Hedonism
As a matter of empirical fact, people only value happiness.

This is a *factual* claim about people's priorities in life. It may be true or false. At any rate, it is commonly stressed that its truth or falsehood does not, in and of itself,

[32]Note that Egocentrism is not the same as Egoism. Egoism is a moral theory, according to which an act is right if and only if it maximally promotes the well-being of the agent. Egocentrism is merely a component of a moral theory. Note, also, that not every moral theory that contains Egocentrism as a component is a form of Egoism. A non-welfarist, egocentric moral theory which accepts certain other factors X, Y, ... besides happiness may be non-egoistic.

have any bearing on the acceptability of Welfare Hedonism as a *normative* view.[33] Furthermore, Welfare Hedonism should not be confused with the stronger claim of "Value Hedonism." (Kagan 1998, 31)

Value Hedonism
Pleasure is the only good (and pain is the only bad).

Value Hedonism is a thesis about the good. Welfare Hedonism, in contrast, is merely an interpretation of the idea of well-being.[34] The former is a logical consequence of combining Welfarism – the view that overall well-being W is the only good – with Welfare Hedonism – the view that well-being is pleasure and the avoidance of pain. Hence, proponents of CU must subscribe to Value Hedonism as well. Other moralists, however, can accept Welfare Hedonism and reject Value Hedonism.[35] They can affirm, that is, that well-being is pleasure, but deny that well-being is the only good.

It is easy to go into reverse and show that Maximization, Welfarism, Welfare Hedonism, Summation, Equal Treatment, and Universalism yield the classic utilitarian criterion of rightness. Maximization says that an act is right if and only if it maximizes goodness. Welfarism specifies that goodness, G, is equal to overall welfare, W, which, in turn, is taken to be a functional f of individual well-being $u_1, \ldots u_n$ of the morally relevant individuals $1, \ldots, n$ that is weighted by $w_1, \ldots, w_n$. In conjunction, these two yield the claim that an act is right if and only if it maximizes a function W, which takes individual well-being as its arguments. Welfare Hedonism adds to this that well-being should be interpreted as sensory happiness. And Summation, Equal Treatment, and Universalism specify that the good is the equally weighted sum total of sensory happiness of all sentient beings. All these claims logically add up to the view that an act is right if and only if it brings about the greatest sum total of happiness for all.

We have factorized CU into six logically independent claims, then. An important *addendum* is, however, in order. CU's criterion of rightness is usually interpreted in an *objective* rather than a subjective sense (cf. Sinnott-Armstrong 2011). That is, according to CU, whether an act is right depends only on what actually happens. In other words, it does not depend, e.g., on what the agent expects to happen as the result of a given act.

Objectivism
The goodness of an act depends only on its *objective* consequences.[36]

[33] Note, however, that one may use Psychological Hedonism as a premise in an argument that seeks to establish Welfare Hedonism. One author who does that is Mill (1863). For this point see, e.g., Mukerji (2013c, 303) and Sumner (1996, 187).

[34] Furthermore, Welfare Hedonism is not a claim about the prudential good of the individual. Logically speaking, one can accept that well-being is pleasure minus pain and reject the idea that well-being is the only thing that gives the life of an individual value.

[35] This point is also made by Darwall (2004, 38).

[36] A very similar claim is put forward by Oddie and Menzies (1992, 516) under the name "actual-outcome consequentialism." Howard-Snyder (2005, 265) examines a claim that she calls

This component, too, is logically independent of the rest of the doctrine. We can combine Maximization, Welfarism, Welfare Hedonism, Summation, Equal Treatment, and Universalism with a claim that is contrary to Objectivism.

Subjectivism
The goodness of an act depends only on its *subjective* consequences.

Subjectivism allows of various interpretations. In Sect. 4.2.1.1, we will talk about some of them in more detail. To anticipate, one version of Subjectivism is the idea that the goodness of an act depends only on the consequences that the agent expects. Combine this with Maximization, Welfarism, Welfare Hedonism, Summation, Equal Treatment, and Universalism and you get the claim that an act is right if and only if the agent *expects* that it will maximize the sum total of happiness. Likewise, Objectivism does not require Maximization, Welfarism, Welfare Hedonism, Summation, Equal Treatment, and Universalism. So Objectivism is logically independent of the other components of CU that we considered so far.

4.1.2 The Practical Component

In Sect. 1.2, we distinguished between the two purposes of moral theories. The first purpose, we said, is to state general principles for the moral evaluation of acts. It is addressed by its *theoretical* component, which gives us a criterion of rightness that states conditions for the rightness of an act. The second purpose of a moral theory is to give us practical instructions in order to help us with our moral choices. It is the focus of the theory's *practical* component. In the previous section, we covered the theoretical component of CU. In this section, we shall briefly address its practical part.

To this end, let us remind ourselves of a distinction that we drew earlier in Sect. 1.2.2. As we said, there are two practical strategies that moral theorists may recommend. The first is a Direct Strategy. The second is an Indirect Strategy. For convenience, let us restate them here.

Direct Strategy
Use the criterion of rightness as a decision-making tool.

Indirect Strategy
Use a heuristic choice criterion as a decision-making tool.

As we learnt above, the criterion of rightness that is contained in the theoretical component of a moral theory states normative factors or, more specifically, right-making features $e_1, \ldots, e_n$. On the Direct Strategy, recall, moral agents facing

"objectivism about right and wrong." It says, roughly, that the rightness or wrongness of an act depends only on objective factors. Since CU affirms Objectivism about Goodness and determines the rightness and wrongness of an act solely based on the good, it is also objectivist in Howard-Snyder's sense.

a moral choice are supposed to examine all their options, to investigate which courses of action are right by checking for properties $e_1, \ldots, e_n$, and then to choose accordingly. On the Indirect Strategy, moral agents are supposed to use heuristics instead.

Based on the distinction between the Direct Strategy and the Indirect Strategy, we can now introduce a further distinction between two views that consequentialists may take regarding the practical component of their moral theories. They can opt for Directness, which is the following view.

Directness
In each and every moral choice situation the agent should choose the Direct Strategy, i.e. use her criterion of rightness as a decision-making tool.

Or they can adopt the following idea.

Indirectness
There are choice situations in which the agent should choose the Indirect Strategy, i.e. use a heuristic as a decision-making tool.[37]

CU is commonly depicted as a version of Direct Consequentialism which means that it adopts the claim Directness along with its criterion of rightness.[38] In other words, CU is viewed as maintaining that the moral agent ought to calculate the consequences of the acts available to her, evaluate them in terms of general happiness, and choose one which is at least as good as any alternative act.

This, then, concludes the first step of our analysis. We have factorized CU into logically distinct components. In the next two steps of our procedure, we will fathom the spectrum of alternative components that we need to take into account when setting up our case against consequentialism. In step (ii), we will investigate which variants of the consequentialist family we may safely ignore. In step (iii), then, we will examine which alternative variants of consequentialism we have to take into consideration. Before we turn to these tasks, however, we will try, in the next two sections, to develop a better grasp of CU. We will do that by connecting it with notions that are commonly associated with the doctrine, either to characterize or to motivate it.

4.1.3 Characteristics

There is an obvious objection to our factorization of CU. It may be said that a number of important ideas that are commonly associated with the doctrine are

[37]The distinction between Directness and Indirectness is foreshadowed, e.g., by Bentham (1838, 16), Mill (1863, 26), Sidgwick (1907, 413), and Moore (1903/1959, Ch.V) and has received important discussions in recent decades by Bales (1971), Pettit and Brennan (1986), and Sumner (1987).

[38]In the present context, it should be re-emphasized that CU is not the view of classic utilitarian thinkers. None of them seems to have endorsed Directness. For literature references, see footnote 37.

missing. CU is often regarded as *impartial*, *agent-neutral*, and *aggregative* (cf. Sinnott-Armstrong 2011). So why do these features not show up as components of the classic utilitarian doctrine? The answer is simple. These characteristics are not additional features of CU. Rather, they are the logical consequence of the CU-components that we identified above. To see this, let us consider each of them in turn.

Let us start with the idea of moral impartiality.[39] First up, let us clarify its meaning. Intuitively, a moral theory is impartial to the extent that it neither favours nor disfavours anyone in particular. This specification leaves open whether impartiality is an axiological or a deontic property. It leaves open, that is, whether it is intended as a property of CU's conception of the right or as a feature of CU's account of the good. It seems that both interpretations are possible.[40] We can formulate both a principle of axiological impartiality and a principle of deontic impartiality.[41]

Axiological Impartiality
An act is good to the extent that it is impartially good. It is impartially good to the extent that it promotes the well-being of *all* morally relevant individuals *weighted equally*.

Deontic Impartiality
The moral status of an act depends only on the extent to which it is impartially good (as specified by the principle of Axiological Impartiality).

It is obvious that a proponent of CU has to subscribe to both of these claims. Axiological Impartiality is a direct consequence of CUG and, in particular, of Universalism and Equal Treatment (cf. Mukerji 2013c, 300). Deontic Impartiality is implied by the combination of CUG and the consequentialist Core Idea, viz. that rightness depends only on goodness. Therefore, to say that CU is an impartial doctrine does not add anything to it. It merely explicates a characteristic of CU that it possesses in virtue of its components.[42]

[39]Note that impartiality need not be understood as an exclusively moral notion (cf. Jollimore 2008). We, however, shall only discuss it to the extent that it can be seen as a moral idea.

[40]Talk of the "impartial good" is quite common (cf., e.g., Hooker 2000, 191, Williams 1981, 15). This suggests that moral impartiality may be understood as an axiological notion. And there is also the notion of "impartial rightness" (Blum 1988, 479), which is clearly a deontic concept.

[41]Note that these interpretations of the idea of impartiality are not the only possible ones, but merely those which seem most congenial to consequentialism. Arguably, the notion of impartiality is not limited to the family of consequentialist doctrines. Deontological theories, too, have been understood as offering an impartial view of morality (cf. Jollimore 2008).

[42]It may be noted, furthermore, that both the axiological and the deontic variant of classic utilitarian impartiality are independent of the idea of Summation. There are alternatives to Summation which, when combined with a consequentialist moral theory that contains Universalism and Equal Treatment, also yield an impartial conception. In place of Summation, a moral theorist may adopt, e.g., Maximin that we will consider in Sect. 4.2.3.2. It is the view that the goodness of a state of affairs is proportional to the well-being enjoyed by the worst-off individual. Such a theorist, too, calculates the good based on all individuals' well-being weighted equally and, hence, has a claim to impartiality. This is also emphasized by Rawls (1971/1999, 165).

The second property that is often ascribed to CU is agent-neutrality (cf., e.g., Sinnott-Armstrong 2011). It is once again not clear whether we are dealing with an axiological or a deontic property here. Many authors use verbiages like the "agent-neutral good" (Ridge 2011) or talk about states that are "good from an agent-neutral point of view" (Lippert-Rasmussen 2005, 159), which suggests that we are dealing with an axiological property. Others use agent-neutrality as a property that attaches to a moral theory's conception of the right (cf. Darwall 2003, 131).[43] They regard it as a deontic notion. We may state the two ideas as follows.[44]

Axiological Agent-Neutrality
The goodness of an act does not depend on the identity of the agent.

Deontic Agent-Neutrality
The rightness of an act does not depend on the identity of the agent.[45]

The two theses are logically independent such that a moral theorist can endorse either one without making a commitment regarding the other.[46] But CU is com-

[43]In fact, elsewhere Stephen Darwall uses the axiological notion of agent-neutrality as well (cf. Darwall 2004, 15).

[44]Both statements are intended as universally quantified statements that range over all choice situations, acts, and agents.

[45]This is, at least, one common interpretation of deontic agent-neutrality. Note, however, that there are others. Ridge (2011), e.g., distinguishes between a reason-based, principle-statement-based, and perspective-based conception of agent-neutrality.

[46]To prove this, we have to show that there are (1) theories which are axiologically and deontically agent-neutral; (2) neither axiologically nor deontically agent-neutral; (3) axiologically agent-neutral, but not deontically agent-neutral; (4) not axiologically agent-neutral, but deontically agent-neutral. This can be proven by ostension.

Ad(1): Classic Utilitarianism. According to CU, an act is right if and only if it maximizes the sum total of happiness of all sentient creatures. This doctrine fulfils Axiological Agent-Neutrality. The goodness of an act does not depend on the identity of the agent. It only depends on the extent to which it promotes overall happiness. It is deontically agent-neutral since it takes the good to be the sole determinant of the right.

Ad(2): Ethical Egoism. Ethical Egoism is the theory that an act is right if and only if it maximizes the well-being of the agent. This theory violates Axiological Agent-Neutrality since it takes the good to be the good of the agent. It also violates Deontic Agent-Neutrality since the moral status of an act is determined solely by its axiological status (which depends on the identity of the agent).

Ad(3): Moderate Deontology. Many moderate deontologists believe that moral agents should promote the good subject to certain constraints (cf. Kagan 1998, 72–73). Such doctrines can fulfil Axiological Agent-Neutrality and violate Deontic Agent-Neutrality, viz. in the case where they endorse an agent-neutral account of the good (e.g. the Classic Utilitarian Conception of the Good) and constraints that are relative to the agent (e.g. the requirement that the agent be loyal to *her* friends).

Ad(4): Absolutist Deontology. Versions of absolutist deontology claim that whether an act is right or wrong depends only on intrinsic properties of and/or past-related factors about the act. They regard considerations of goodness as irrelevant to considerations of rightness. A moral theorist who subscribes to such an absolutist deontological doctrine can, therefore, accept a theory of the good which violates Axiological Agent-Neutrality (e.g. the view that an act is good to the

mitted to both because both are implied by its components. Axiological Agent-Neutrality is a consequence of CUG. To be more precise, it is a consequence of the fact that its components Welfarism, Summation, Universalism, Equal Treatment, and Welfare Hedonism contain no reference to the identity of the agent. CU's Deontic Agent-Neutrality follows from two facts: firstly, that its account of the good is agent-neutral and, secondly, that the issue whether an act is right depends only on the goodness that it produces. As it turns out then, agent-neutrality, too, is not an additional feature of CU, but a logical consequence of its components.

Let us turn to the third item on our list, viz. the *aggregative* character of CU. Again we should clarify what it means for a moral theory to be aggregative. As Norcross (2009, 84) points out, we have to distinguish, once again, between an axiological and a deontic sense of aggregation.[47]

Axiological Aggregation
Overall welfare, W, is an aggregative functional f of individual well-being.

Deontic Aggregation
Whether an act is right depends at least partly on overall welfare, W, which is an aggregative functional of individual well-being.

It is, in fact, easy to make intuitive sense of these propositions. Axiological aggregation is simply the idea that "harms and benefits [of different individuals] can be traded off against each other in determining the overall goodness (or badness) of a state of affairs" while Deontic Aggregation "involves the claim that harms and benefits can be traded off against each other in determining which choices are required, permissible, or forbidden." (Norcross 2009, 84) Note that the two theses are logically linked. The latter implies the former, but not *vice versa*. That is, no moral theory that fulfils Deontic Aggregation can violate Axiological Aggregation. But there could be theories which fulfil Axiological Aggregation while violating Deontic Aggregation.[48] That means there can be theories which are axiologically aggregative without being deontically aggregative. Certain forms of absolutist deontology are examples of this kind of theory. Absolutist deontologists may hold an aggregative view about overall welfare, W. On their view, however, rightness is not at all a matter of promoting W and, hence, they can accept an overall moral view which violates Deontic Aggregation. However, since, on CU, the rightness of an act depends exclusively on the overall welfare that it produces, the doctrine is not deontically aggregative unless it is also axiologically aggregative. To show, then,

extent that it benefits the agent). She can at the same time fulfil Deontic Agent-Neutrality, e.g. if adheres to the principle that an act is wrong if and only if it is a lie.

[47] For simplicity's sake, we ignore the possibility of non-welfarist accounts of the good in our explication of aggregation.

[48] As Alastair Norcross points out, however, "[g]iven the structure of consequentialist theories, a commitment to axiological aggregation entails a commitment to deontic aggregation." (Norcross 2009, 84)

that CU fulfils both Deontic Aggregation and Axiological Aggregation, we only need to show that it fulfils the latter.

The trouble with Axiological Aggregation is that it is ambiguous. It is not clear what it means to say that overall welfare is determined by an *aggregative* functional of individual well-being. There are at least three different senses in which this claim may be interpreted. On the first interpretation, a moral theory is axiologically aggregative if and only if its W contains the CU-component Summation.

Axiological Aggregation (Version 1)
Overall welfare is measured by the *sum* of the appropriately weighted levels of well-being of all morally relevant individuals.

This, it seems to me, is the most common usage amongst contemporary moral philosophers (cf., e.g., McNaughton and Rawling 2009, 345; Hooker 2009b, 126; Sumner 1996, 186).

In economics, the notion of aggregation is used as well, though in a much broader sense. Following Arrow (1951/1963), social choice theorists have analysed how individual betterness orderings over social states can be "aggregated" into a social betterness ordering using Social Welfare Functions (SWFs).[49] A SWF f is a function which takes betterness orderings, R_i, of individuals $i = 1, \ldots, n$ over a set X of social states $x, y, z, \ldots$ as its arguments and maps them onto a social betterness ordering R (cf. Arrow 1951/1963, 23).[50] A moral theory can be interpreted as axiologically aggregative if its W contains an operation that can be formally represented as a SWF.[51]

Axiological Aggregation (Version 2)
The goodness of an act is computed *via* a function that can be formally represented as an Arrovian SWF.

[49] A betterness ordering R_i of individual i is a transitive, binary relation between social states. Note that economists usually use the idea of a preference ordering as opposed to that of a betterness ordering.

[50] This talk of betterness relations is, in fact, equivalent to our talk of individual well-being functions $u_1, \ldots, u_n$. On this point, see footnote 14.

[51] This interpretation of aggregation is problematic because it is very inclusive. It comprises almost all serious moral theories. It only excludes those theories which either do not possess an account of the good or do not determine what is good based on what benefits individuals. A moral code which only contains rules, such as "Don't lie" or "Keep your promises," can do without a conception of the good and is an example of the former kind of theory. A divine command theory, which determines what is morally good based on God's will, is an example of the latter kind of theory. Nevertheless, some philosophers, e.g. Vallentyne and Kagan (1997, 5), accept various operations as forms of aggregation which are aggregative only in this weak sense. On the corresponding interpretation of Deontic Aggregation, a moral theory is aggregative in the deontic sense if its account of the good contains an operation which can be represented as a SWF and if, on its account of the right, the right depends solely on the good. As Hirose (2007, 275) points out, this interpretation of Deontic Aggregation, in fact, comprises all consequentialist moral doctrines.

Note that every welfarist (or extra-welfarist) consequentialist theory fulfils this kind of Axiological Aggregation. It, hence, appears to be too broad. A narrower and more intuitive interpretation of Axiological Aggregation is this.

Axiological Aggregation (Version 3)
The goodness of an act is computed *via* a function that can be formally represented as an Arrovian SWF and allows trade-offs.[52]

Here is what this means. Consider two social states, x and x'. x' differs from x only in one respect. There is one individual $i \in N$, who is worse-off under x' than under x. A moral theory is axiologically aggregative if its account of the good allows us to transform x' into another state x'' which is as good as x by making individuals $j, k, l, \ldots \in N_{-i}$ better off (for every value of i).

CU is aggregative in all three senses.

- It is aggregative in the first sense because it contains Summation as a component.
- It is aggregative in the second sense because it contains some operation f for combining the separate well-being of individuals into one value, viz. Summation.
- And it is aggregative in the third sense because this operation allows trade-offs between the weal and woe of different individuals. In any given state, we can take away any number of units of happiness from any individual, add an equivalent amount to the happiness of another individual, and CU will judge that the resulting state is as good as the state from which we started.

Let us sum up, then. CU is an impartial, agent-neutral, and aggregative doctrine. These characteristics are a consequence of its logical building blocks.

4.1.4 Motivation

Before we move on to step (ii) of our inquiry, let us consider, very briefly, why anyone would find CU attractive as a moral theory. Of course, in spelling out wherein its appeal lies, we are not aiming for a final verdict. We are simply trying to see the motivation for the individual claims as well as the doctrine as a whole.

Obviously, each of the claims has something to be said in its favour. Maximization has great appeal, as John Rawls explains, because it appears to "embody the idea of rationality." "It is natural," he says, "to think that rationality is maximizing something and that in morals it must be maximizing the good." (Rawls 1971/1999, 21) Other authors have argued along the same lines (e.g. Scheffler 1985; Portmore 2007, 2011; Roberts 2002; Scarre 1996, 18). Now, why is it natural to think, as Rawls says, that Maximization is *rational*? To understand this, it is instructive to take a brief look at the theory of rational choice, as it is studied by economists. Economists believe that it is rational for an agent to maximize her *utility* expected.

[52]This understanding of Axiological Aggregation is proposed, e.g., by Hirose (2004, 66 and 2011, 65) and Vallentyne (1987, 26).

It is important, however, to note that they mean something quite specific by the term "utility." They do *not* mean by it, e.g., "sensory pleasure" or something like that. To them, utility is not an entity in the world but, rather, a derived notion whose content is fixed against the background of certain rationality axioms about the structure of preferences over objects of choice x, y, z .... One axiom purports, e.g., that a rational agent's preferences are *transitive*, such that, if she prefers x to y and y to z, she also prefers x to z. The other axioms, too, are purely formal in nature and seem at least *prima facie* acceptable. They are simply "requirements of coherence" that regulate the appropriate treatment of objects according to preferences and probabilities. I do not wish to go into any detail about them.[53] My point is simply this. If an agent's preferences are coherent in the sense that they conform to these axioms, they can be represented, as von Neumann and Morgenstern (1944/1955) have proven, by a real-valued expected utility function which ascribes higher numbers to more preferred objects, lower numbers to less preferred objects, and the same number to equally preferred objects. Maximizing expected utility is interpreted, then, as choosing the highest-ranked object that is available. To maximize expected utility simply means to choose according to rational preferences that obey the requirements of coherence. So once the coherence axioms are granted, "expected utility maximization takes care of itself," as Dreier (2004, 139) says. The question "What maximizes my utility?" simply is the question "What ought I, rationally, to do?" In the context of the theory of choice it is, hence, hard to see how anything other than Maximization could be rationally warranted. Now, moral theory seems to be similar to rational choice theory. At least, it is not initially clear why it should not be (though we will come back to this point in Sect. 4.2.2.1). So the conclusion about utility maximization could conceivably be extended, by way of analogy, to moral theory, such that in moral theory, too, the only rational response to the good is to maximize it. This, at any rate, is one way to make sense of Rawls's claim that Maximization seems rational.

The second component of CU that we identified above was Welfarism. Formally, it is the notion that intrinsic goodness, G, is equal to overall welfare, W, which, in turn, is taken to be a functional, f, of individual well-being functions $u_1, \ldots, u_n$ whose relative importance is measured by weights $w_1, \ldots, w_n$. This idea, too, can be given a very persuasive motivation. In Sect. 4.1.1, recall, we distinguished a number of interpretations of welfarism. As we defined Welfarism, it subsumes all of these ideas. Hence, everything that can be said in favour of any of the more specific interpretations of welfarism can also be said to support Welfarism. One point that is often made in this context is that welfarism is compatible with the Pareto Principle,

[53] See, however, von Neumann and Morgenstern (1944/1955). Accessible discussions of the concept of utility according to von Neumann and Morgenstern can be found in Nida-Rümelin (2002, 136–142) and Resnik (1987/2000, 88–96).

while non-welfarism is not (cf. Kaplow and Shavell 2001). To explain, there are two versions of the Pareto Principle that are relevant here.[54]

Strong Pareto Principle
If an object of evaluation, *x*, is at least as good as another, *x'*, for all individuals and better for at least one individual, it is better.[55]

Pareto Indifference
If an object of evaluation *x* is as good as another *x'* for all individuals, then *x* and *x'* are equally good.[56]

These two principles, I take it, are rather plausible. The Strong Pareto Principle claims that *x* should be seen as better than *x'* if it is better for some and worse for none. As Ng (2004, 27) says, to deny this seems to take "a rather peculiar ethic." Pareto Indifference seems equally plausible. As Martin Rechenauer says, if "the individuals completely agree in their utility functions with respect to these alternatives [i.e. *x* and *x'*; NM], then of course you should think of these alternatives as being equally good." (Rechenauer 2003, 9) Now, these two principles jointly imply the Person-Affecting Restriction that we discussed above.[57] That means, if you accept Strong Pareto and Pareto Indifference, you had better also accept the Person-Affecting Restriction. The Person-Affecting Restriction, in turn, commits you to Welfarism, as we pointed out above. Considerations based on the Pareto Principle, hence, give us a strong reason to accept Welfarism.

Four further CU-components that we identified above had to do with the precise content of overall welfare, *W*. The first item on that list was Summation. It is intuitively plausible as well. Like Maximization, it gets its plausibility from an analogy with rationality, as, e.g., Rawls (1971/1999, 21) and Scheffler (1982/1994, 11) have noted. Suppose I face a non-moral decision problem between two act

[54]As Broome (1991) points out, the Pareto Principle (in all its versions) is actually narrower than our interpretation of it. The Pareto Principle, Broome says, is, in fact, tied to a perferentist interpretation of well-being. In our formulation, it is not. It is, hence, closer to what Broome calls the "principle of personal good," which leaves open how well-being is measured. Note, however, that our usage is not all too uncommon. Ng (2004, 27), e.g., uses it more broadly. He says that it "can be defined with respect to preference or welfare."

[55]The weak version of the Pareto Principle was briefly mentioned in Sect. 3.1.2.

[56]In conjunction, the Strong Pareto Principle and Pareto Indifference are called the Full Pareto Principle (cf. Suzumura 2001, 96).

[57]This theorem can be proven as follows: Let *xPx'* represent the idea that *x* is strictly better than *x'*, *xRx'* that *x* is at least as good as *x'*, and *xIx'* that *x* and *x'* are equally good. Furthermore, let $xP_i\ x'$, $xR_i\ x'$, and $xI_i\ x'$ denote that *x* is better than *x'*, at least as good as *x'*, and as good as *x'*, respectively, for the individual $i \in N$. In its comparative form, the Person-Affecting Restriction can be formulated as saying that, if *xPx'* holds, then we have $\exists i{:}xP_i\ x'$ (cf. Arrhenius 2009, 289). To show that this idea is implied by Strong Pareto and Pareto Indifference, assume that *xPx'*. By Strong Pareto, this implies that $\neg\exists i{:}x'P_i\ x$ because if it was the case that $\exists i{:}x'P_i\ x$, we would have either *x'Px* or *¬xPx'* (both of which contradicts our assumption). $\neg\exists i{:}x'P_i\ x$, in turn, implies that $\forall i{:}xR_i\ x'$. *xPx'* implies, furthermore, that *¬xIx'* which, in turn, implies, by Pareto Indifference, that $\neg\forall i{:}xI_ix'$. These two findings, taken together, imply that, if *xPx'*, then we have $\forall i{:}xR_ix'$ and $\neg\forall i{:}xI_ix'$. This conjunction yields that $\exists i{:}\ xP_ix'$, which proves our theorem.

options, a_1 and a_2, and I want to find out which is better for me. How do I do it? It seems rational for me, though this time in a *substantive* sense, to take stock of the benefits and losses that each option promises and balance them off against one another. The better choice option for me is, then, the one which has the more favourable balance of benefits over losses. Since the case of rational choice seems to be analogous to the case of moral choice, it appears natural to suggest that I should decide in the same way when I face a moral choice problem. I should first work out whether a_1 or a_2 offers a more preferable balance of benefits over losses and decide on that basis (cf. Mukerji 2013c, 300).

The fact that aggregation *via* Summation usually has rather intuitive consequences in a wide range of cases lends it further plausibility. Take the case we came across on page 32: A flood is threatening the lives of people on both sides of an island. The captain of a freight ship has to choose between saving the people on the north or the south side because he cannot save them both. It seems that, in such a case, the captain morally ought to do what will save more people and prevent greater harm. He ought to go to the south side if there are more people on the south side. And he ought to go to the north side if there are more people on the north side. A consequentialist moral conception which commits to the principle of Summation evaluates the goodness of the captain's acting by summing up its effects on the well-being of the individuals whose life is at stake. It will most likely judge that the captain ought to save the many rather than the few and is, hence, in line with common sense – at least when it comes to this particular example.[58]

As we discussed in the previous section, Universalism and Equal Treatment conjoined offer a *prima facie* attractive interpretation of the idea of impartiality. Since the moral point of view has often been identified with the impartial point of view, it seems as though impartiality is, in fact, definitional of morality (cf. Jollimore 2008). This would, of course, mean that every reasonable moral theory should endorse Universalism and Equal Treat or, at the very least, claims that are very similar to them. Furthermore, impartiality seems to be a good guiding ideal for moral behaviour in many areas. Think, e.g., of public officials. We would find it morally outrageous if public officials were *not* impartial. It is, after all, "the essence of public service as such that public servants should serve the public at large." (Goodin 1995, 9) Or think of parents who give preference to some of their children over others. These examples may be taken to suggest that the implications of impartiality are usually well in line with our moral intuitions.

Welfare Hedonism offers what appears to be a rather appealing specification of the welfarist idea and has had many proponents throughout the ages and up until recent times (e.g. Feldman 2004). We arrive at it by a simple train of thought.

[58] I say that this will *most likely* (as opposed to *necessarily*) be the case because not every moral conception which subscribes to Summation needs to subscribe to Equal Treatment, too. A moral conception which endorses Summation and is based on an inegalitarian conception of moral worth might, e.g., embrace the view that the king's welfare is infinitely more valuable than everybody else's. If the king is on the side of the island with the smaller number of people, such a moral theory would rule that the captain should head for this side, even though it contains Summation.

Our happiness does, in fact, matter to *us*. Given that we know this, we may find it plausible to assume that, therefore, other people want to be happy, too. And we can take this as evidence that happiness really is a good. This reasoning was put forward by John Stuart Mill in his famous book *Utilitarianism* (cf. Mill 1863, 51–52). Now, one may see in this a "naturalistic fallacy," as G. E. Moore did (cf. Moore 1903/1959, 18). That is, one may suspect that Mill thought "people desire happiness" just *means* the same as "happiness is good." But this can be questioned (cf. Ryan 1966; Sayre-McCord 2001, 336). If we assume, as is *prima facie* plausible, that people do, in fact, desire happiness and if we assume, further, that they are not idiots and are capable of judging what is good for them, we might end up finding Welfare Hedonism quite a plausible view.

Objectivism is, at first glance, also convincing. Obviously, it matters to us what *actually* happens (cf. Sen 2010, 213). After all, all of our actions seem to be motivated by our desire to change the world and to bring it closer to the world that we would like it to be. Why, then, should we not morally evaluate our acts in terms of what they actually achieve?

Finally, CU is in line with the attractive notion that "the whole point of ethical judgments is to guide practice." (Singer 1979/1993, 2) This is partly due to Directness, which claims that the moral agent ought to apply the classic utilitarian criterion of rightness to every single moral choice situation. Directness appears to ensure, then, that the verdicts which flow from the classic utilitarian criterion of rightness have maximal practical import.

There is a plausible motivation, then, for all of the fundamental moral claims that lie at the heart of CU. In addition, the doctrine as a whole has a number of desirable properties which might motivate moral theorists to endorse it. First of all, it is very simple and economical in terms of basic notions (cf., e.g., Mukerji 2013c, 301; Williams 1973, 137). Happiness or utility is the only key concept. In contrast, other moral theories involve complex and abstract moral notions, such as *desert*, *need*, or *justice*, which are hard to pin down. Secondly, CU is apparently free of moral dilemmas.[59] The doctrine, it seems, will never judge that in a choice between two options, a_1 and a_2, both of them are wrong. Either a_1 is better than a_2, a_2 is better than a_1 or they are equally good. In the first case, it is right to do a_1. In the second case, it is right to do a_2. And in the third case, both a_1 and a_2 are right. Thirdly, CU appears to exhibit a high degree of systematicity. It construes the right as maximizing the good and thus establishes a logical connection between the two central concepts of ethics. Furthermore, the components of the classic utilitarian notion of the good are internally connected. Universalism links up with Welfare Hedonism. Equal Treatment and Universalism are united through their connection with impartiality. Lastly, all moral verdicts in CU are united by the fact that they

[59] On this point, see Norcross (1995) and Slote (1985b). It can, however, be shown that under certain deontic-logical assumption it is possible to reformulate CU's criterion of rightness, such that it, in fact, allows for moral dilemmas (cf. Mukerji 2013c, 306).

follow from one simple standard of right and wrong. To sum up, then, the reasons for finding CU *prima facie* attractive are legion.[60]

4.2 Variants of Consequentialism

Let us take stock of where we are. We examined a paradigmatic consequentialist doctrine, viz. CU. We factorized it into a number of logically distinct moral claims. Furthermore, we looked at a number of ideas that can be used to characterize and motivate CU. This concludes step (i) of our methodic procedure, FRA_2. We know now what the paradigmatic consequentialist doctrine looks like. We know, therefore, what the structure of a consequentialist theory is and which fundamental moral claims consequentialists paradigmatically subscribe to. What we have to do now is to examine which alternative forms of consequentialism we have to take into account when we put together our case against consequentialism.

In what follows, we will go through steps (ii) and (iii) of FRA_2. In step (ii), we will investigate which determinable components can be put aside for the purpose of our investigation. In step (iii), then, we will take stock of the spectrum of alternative forms of consequentialism which depart from the paradigmatic member of the family.

4.2.1 Unmotivated Variants

In this section, we shall examine the distinctions between Subjective and Objective Consequentialism, Direct and Indirect Consequentialism, and variants of consequentialism which employ alternative theories of well-being. Our aim will be to show that we can justifiably neglect these distinctions in our trolleyological investigation.

4.2.1.1 Subjective Consequentialism

In our factorization of CU, we identified a component that we called Objectivism. It claims that the axiological status of an act depends only on the goodness of the consequences that it produces as a matter of *objective* fact. Combine this with the Core Idea behind consequentialism, viz. that rightness depends only on the goodness of consequences, and you arrive at a view that may be called Objective Consequentialism.[61]

[60] For a systematic critique of almost all of the above points, see Mukerji (2013c, 301–307).

[61] Slote (1992/1995, 239) proposes the label "actualism."

Objective Consequentialism
Whether an act is permissible depends only on the goodness of its *objective* consequences.[62]

Motivation for this view is not hard to find. The implications of Objective Consequentialism seem to coincide with our intuitive judgements. We often judge a person's actions by the difference that they make in the world. When they have bad consequences, we usually want to say that the person did something wrong. And if they have good consequences, we normally feel inclined to judge that she did the right thing. At any rate, so it seems at first glance. To illustrate this point, consider the following case.

Poisonous Medicine
Smith gives Jones a medicine that kills him.[63]

How would a form of Objective Consequentialism judge Smith's acting in *Poisonous Medicine*? His act has bad consequences. Poor Jones, after all, ends up dead! Surely, then, Objective Consequentialism would condemn Smith's act as morally wrong. Intuitively, this is the right judgement to make.

Or is it? Maybe we should not be too hasty in our conclusion. There could be relevant facts about the case that we do not know. What, e.g., would we say, if the case turned out to be as follows?

Poisonous Medicine*
Smith gives Jones a medicine that kills him. However, Smith did not know that the medicine was poisonous. He expected that it would cure Jones.

Given the information that is provided in *Poisonous Medicine**, is it still appropriate to say that Smith did something wrong, as an objective consequentialist presumably would? This is not certain. At any rate, we should get clear on what we mean when we call an action "right" or "wrong." Philosophers commonly differentiate between two senses in which they use these expressions.[64] There is the *fact-relative* sense, on the one hand, and the *belief-relative* sense, on the other. Obviously, it was wrong for Smith to give Jones the lethal substance in the fact-relative sense. But since he did not know that it was lethal, his act should probably not be called "wrong" in the belief-relative sense. After all, Smith did what he thought was best for Jones. Now, the question arises which of the two notions we are trying to give an account of when we engage in moral theorizing. Many philosophers believe that we should aim to give an account of the belief-relative sense of right and wrong. One of their reasons for thinking this is, as John Stuart Mill has famously pointed out, that "we do not call anything wrong unless we mean

[62] Note that the label "objective consequentialism" is used by authors following Railton (1984) to refer to what we shall call below "Indirect Consequentialism."An instructive discussion that disentangles these concepts is given by Forschler (2009). See, also, footnote 73.

[63] We borrow the following sequence of examples from Parfit (2011, Ch. 21).

[64] See, e.g., Gibbard (1990, 42), Nida-Rümelin (1993, 81-84), and Parfit (2011, 150).

to imply that a person ought to be punished in some way or other for doing it."[65] (Mill 1863, 72) In *Poisonous Medicine** it seems somewhat inappropriate to blame Smith for what he did, let alone punish him. So, perhaps, we should not call what he did "wrong." Those who are convinced by this line of reasoning may feel inclined to accept some version of Subjective Consequentialism.[66] It combines the Core Idea of consequentialism with the view that the axiological status of an act depends only on its subjective consequences.

Subjective Consequentialism (SC)
Whether an act is morally permissible depends only on the goodness of its *subjective* consequences.

It is unclear, of course, what the term "subjective consequences" means here. One interpretation which seems to be especially useful in the context of *Poisonous Medicine** is to take this expression as referring to the outcome that the agent *expects*.

Subjective Consequentialism – Version 1 (SC_1)
Whether an act is morally permissible depends only on the goodness of its expected consequences.[67]

Smith did not expect that the medicine would kill Jones. He expected, rather, that it would cure him. That is, he anticipated that his act would have a good consequence. Should we not judge what Smith did on *that* basis? SC_1, at any rate, would have us do that. It classes Smith's act in *Poisonous Medicine** as right, and plausibly so.

But maybe we have still not considered all the relevant facts about the case. Suppose that there is, in fact, one further factor about Smith's act that we have overlooked so far.

Poisonous Medicine**
Smith gives Jones a medicine which, unbeknownst to him, is poisonous and kills Jones. Smith expected that the medicine would cure Jones. However, given the evidence available to him, he should, in fact, have been able to foresee that it would kill the guy.

In this case, it seems as though Smith should not get off the hook so easily. Granted, he did not know that the medicine would kill Jones. But the fact that he could have foreseen this consequence seems to be morally relevant in this case.

[65] For a discussion of the connection between blameworthiness and wrongdoing see, e.g., Adams (2002, esp. 238), Mason (2002), Parfit (1986, 31–35), Skorupski (2000, 142), and Tännsjö (1995).

[66] Note that many different views have been referred to as "subjective consequentialism." Railton (1984, 152), e.g., uses this label to denote "the view that whenever one faces a choice of actions, one should attempt to determine which act of those available would most promote the good, and should then try to act accordingly." This idea is much closer to what we called Directness than it is to the view we refer to as "subjective consequentialism."

[67] There are various names for this idea in the literature. Driver (2001/2003, xiv) and Slote (1992/1995, 239), e.g., call it "expectabilism." Feldman (2006, 69) calls it "expected utility consequentialism," as it has an obvious analogue in Bayesian rational choice theory, viz. the view that the rationality of an act depends entirely on its expected utility for the agent.

Parfit (2011, 151) recommends, therefore, to distinguish a third sense of "right" and "wrong," viz. the *evidence-relative* sense in which Smith's act was certainly wrong. Is this perhaps the interpretation of the word "wrong" that our moral theories should target? Consequentialists who believe that it is will tend to favour a different interpretation of Subjective Consequentialism, viz. the view that the rightness of an act depends only on the goodness of those consequences which, given the available evidence, the agent could have foreseen.[68]

Subjective Consequentialism – Version 2 (SC_2)
Whether an act is morally permissible depends only on the goodness of the consequences that the agent could have foreseen, given the evidence available to her.

SC_2, rather plausibly, gives moral agents a responsibility for the consequences of their actions to the extent that they could foresee them.[69] It also gives them a responsibility for their beliefs, which also seems to be rather credible (cf. Nida-Rümelin 2011b, 33–47). Many consequentialists should, therefore, find SC_2 a rather compelling variant of consequentialism.

Interestingly, there are also forms of consequentialism that we may call Objective-Subjective Consequentialism (OSC) because they are partly objective and partly subjective.[70] The following two variants focus on a subset of the objective consequences of the act, viz. those which the agent expected or should have foreseen, respectively.

Objective-Subjective Consequentialism – Version 1 (OSC_1)
Whether an act is morally permissible depends only on the objective consequences that the agent expected.

Objective-Subjective Consequentialism – Version 2 (OSC_2)
Whether an act is morally permissible depends only on its objective consequences that the agent should have foreseen, given the evidence available to her.

Both of these versions, however, are rather implausible, as the following cases show.

Medicine
Smith gives Jones a medicine. He expects that it will kill Jones. But, in fact, it cures him.

In this case, I take it, we feel that there was something morally wrong with Smith's act. He expects the medicine to kill Jones and gives it to him nevertheless. The medicine does not, in fact, kill Jones. But this is a matter of pure luck. Hence, it should not exculpate Smith. But, according to OSC_1, it does because OSC_1 bases

[68] Some theorists would, perhaps, hesitate to call this version of consequentialism "subjective." As Walter Sinnott-Armstrong (2011) points out, "reasonably foreseeable consequences are (..) not subjective insofar as they do not depend on anything inside the actual subject's mind." However, they are subjective insofar as a "particular subject would foresee [them] if he or she were better informed or more rational." Nothing in our argument turns on this terminological issue, however.

[69] It is, hence, not subject to the objection that it violates the principle *Ought Implies Can*. This objection has been pressed against objective consequentialism. See, e.g., Howard-Snyder (1997).

[70] I am grateful to Erasmus Mayr for making me aware of the possibility of hybrid forms.

the evaluation of an act only on objective consequences. In fact, it seems as though a proponent of OSC_1 would be committed to the view that Smith's act in *Medicine* is morally indifferent as it appears to lack moral status. On OSC_1, we evaluate what an agent does entirely based on the objective consequences she expects. In this case, however, there are no consequences that the agent had expected. Hence, there is nothing about the act which, according to OSC_1, could be evaluated. Another interesting aspect of OSC_1 is that an agent can avoid acting wrongly simply by not forming any expectations about the consequences of what she does. OSC_1, it seems, supports thoughtlessness.

Medicine*
Smith gives Jones a medicine. Given the evidence available to him, he should expect it to kill Jones. But, surprisingly, it cures Jones.

In this case, too, we might feel that Smith acts wrongly. He should have known that the substance was likely to kill Jones but gave it to him nonetheless. The fact that no harm was done is, again, entirely a matter of luck on Smith's part. Nevertheless, OSC_2 finds no fault with his acting in *Medicine** since everything turned out okay.[71]

Now that we have an overview of the various alternatives to Objective Consequentialism, let us proceed with our investigation. As we said above, we are not interested to find out whether the most plausible form of consequentialism is objective or subjective, though this may be an interesting issue in the consequentialist in-house debate.[72] What counts for us is merely the question whether the distinction between objective, subjective, and, for that matter, objective-subjective versions of consequentialism is at all motivated in the context of our inquiry. This is plainly not the case if the implications of all of these variants coincide in trolley cases. Let us ask, therefore, whether this is in fact so.

Recall the characteristics of trolley cases. In particular, remember Characteristics 3 and 5. In a trolley case, we assume that the description of the case contains information about the acts that are available to the agent as well as their consequences. Moreover, we stipulate that the agent knows all the relevant facts about them. Hence, we can assume that the agent should foresee the objective consequences of her acts and that she does, as a matter of fact, expect them. In other words, we can assume that objective consequences, foreseeable consequences, and expected consequences are identical. Because OC, SC1, and SC2 judge the moral status of acts only based on objective consequences, expected consequences, and foreseeable consequences, respectively, their implications in trolley cases will, therefore, be identical. Furthermore, since all objective consequences are foreseeable to the agent, foreseeable objective consequences are identical to objective consequences. And since all objective consequences are in fact expected by the agent, expected objective consequences are also equal to objective consequences. This means that the implications of OSC1 and OSC2 coincide with those of OC, SC1, and SC2. The

[71] In fact, OSC_2 would judge that this act, too, lacks moral status.

[72] For views on this question see, e.g., Carlson (1999b), Howard-Snyder (1997, 1999 and 2005), and Qizilbash (1999).

distinction between these variants of consequentialism will, hence, be irrelevant to our trolleyological investigation. So we can safely put it aside.

4.2.1.2 Indirect Consequentialism

To start this section, let us turn to two theses that we christened Directness and Indirectness in Sect. 4.1.2. For convenience, let us restate them here.

Directness
In each and every moral choice situation the agent should choose the Direct Strategy, i.e. use her criterion of rightness as a decision-making tool.

Indirectness
There are choice situations in which the agent should choose the Indirect Strategy, i.e. use a heuristic as a decision-making tool.

We can partition consequentialist doctrines into versions of Direct Consequentialism and Indirect Consequentialism, depending on whether they accept the former or the latter view.[73] The aim of the following pages is to show that it is possible, for the purpose of our inquiry, to set aside the distinction between these two forms of consequentialism. To be sure, we shall not argue that the differentiation is always unimportant. As Griffin (1994, 179) emphasizes, it "is one of the most important developments in utility theory of recent decades."[74] Rather, we shall seek to support the claim that we can ignore it *in the context of our discussion*.

The strategy of our argument will be essentially the same as it was in the previous section. The starting point of our reasoning is the idea that "[i]ndirect forms of consequentialism are worth discussing separately only if they have different implications from direct consequentialism." (Brink 2006, 386) To show, then, that we can ignore the distinction between them, we need to establish that, when we

[73]Note that various terms are used to refer to what we call "Indirect Consequentialism." Pettit and Brennan (1986) call it "restrictive consequentialism." Railton (1984) calls it "objective consequentialism."

[74]In recent decades, many participants of the debate about consequentialism have focused on the distinction between direct and indirect forms. R.M. Hare's two-levelled utilitarian theory, e.g., hinges on the distinction between the direct application of his criterion of rightness at what he calls the "critical level" and the use of alternative principles for choice at what he refers to as the "intuitive level" (cf. Hare 1981). Flanagan (1993, 34–35), too, has entertained the possibility of indirect versions of consequentialism. Although consequentialism tells us "that the action is best which produces the best outcome," he thinks that "it need not tell us that agents should always act or be motivated to act so as to produce the best outcome." Similarly, David Brink has said that a "criterion of rightness explains what makes an action or motive right or justified; a decision procedure provides a method of deliberation. Teleological theories do provide criteria of rightness, but need not provide decision procedures." (Brink 1986, 421) Cocking and Oakley (1995, 87) have also put forward the view that "[i]ndirect consequentialists seem right to stress that a consequentialist moral agent need not aim at maximizing the good." (For further references as well as historical examples of indirect consequentialism, see also footnote 37.) By now, most consequentialist moral theorists seem to endorse Indirectness (cf. Hooker 2003, 142).

apply variants of Indirect Consequentialism and otherwise identical variants of Direct Consequentialism to trolley case, we get the same practical implications. To simplify our task, we shall confine ourselves to *plausible* versions of Indirect Consequentialism.

Here is how we shall proceed. In a first step, we shall analyse when the use of the Indirect Strategy is, in fact, plausible. On that basis, we will then formulate a plausibility requirement for Indirect Consequentialism. It states conditions under which variants of the doctrine can plausibly recommend the Direct Strategy. Then, we will show that these are never fulfilled in a trolley case, such that plausible versions of Indirect Consequentialism will always recommend the Direct Strategy. That, in turn, means that, in a trolley case, they will make the same practical recommendations as otherwise identical forms of Direct Consequentialism. On the above contention, this suggests that we can, for the purpose of our inquiry, ignore the distinction between Direct Consequentialism and Indirect Consequentialism.

Let us start, then, by asking how we can distinguish plausible versions of Indirect Consequentialism from implausible ones. Roughly, the answer to this question is that indirect consequentialist theories seem to become unreasonable if they "help themselves too liberally to the resources of indirectness." (Williams 1973, 81) To specify and establish this conclusion, we need to understand, first, when it is appropriate to apply the Direct Strategy and when it is appropriate to use the Indirect Strategy. So let us contrast the two.

On the Direct Strategy, the moral agent should apply her criterion of rightness as a method for making moral choices. That is, she should always consider all her options for acting, check them for the relevant right-making features $e_1, \ldots, e_n$, as stated by her criterion of rightness, and choose her act accordingly. On the Indirect Strategy, by contrast, the agent should follow heuristics. To illustrate the difference between the two approaches, consider CU's criterion of rightness. It says that an act is right if and only if it maximizes the happiness sum. This moral standard can be used directly or indirectly. The agent uses it directly when she fathoms all her choice options, evaluates them regarding happiness promotion, and chooses the best one. She uses it indirectly when she employs a heuristic instead. There are various moral heuristics which we could mention in this connection (cf., e.g., Appiah 2008, 56-62; Goldman 2003, 10–61). Since our argument is principled in nature, we need not go through all of them. Two examples should, nevertheless, prove useful for our subsequent discussion.

Above, we distinguished between Type-1 and Type-2 heuristics. So let us consider an example of each type. Type-1 heuristics use only part of the available information. An example of such a Type-1 heuristic is the "satisficing heuristic."[75] (Gigerenzer and Todd 1999, 7; Simon 1955, 104–110) It works like this. The agent starts by defining an "aspiration level." That is, she determines a level of happiness that she aims to achieve. Then, she goes through her choice options, evaluates them in terms of happiness promotion as she goes along, and stops at the first one that is

[75] See, also, Mulgan (2001b, 43). He talks about a "satisficing strategy."

at or above the defined threshold.[76] There are two differences between this strategy and the direct application of CU's criterion of rightness. Firstly, the agent does not count up all her options in order to evaluate them later. Rather, she evaluates them as she enumerates them. Secondly, she does not try to find the maximally good act. Instead, she merely aims to find one which is good enough.

Unlike Type-1 heuristics, Type-2 heuristics do not use the morally relevant properties $e_1, \ldots, e_n$. Instead, they use heuristic features of acts $h_1, \ldots, h_m$ that are highly correlated with $e_1, \ldots, e_n$. On CU, there is only one morally relevant property, viz. the extent to which an act promotes happiness. A Type-2 heuristic for CU ignores this property and focuses on specific heuristic qualities. We find a ready example in the rules of common-sense morality. Utilitarian philosophers have commonly interpreted norms "such as 'Don't harm others', 'Don't take or harm the possessions of others', 'Keep your promises', 'Tell the truth', etc." (Hooker 2003, 142) as moral heuristics (cf., e.g., Smart 1956). A classic utilitarian agent may choose to follow these norms since promise-keeping, truth-telling, and so on usually promote happiness.[77]

Having juxtaposed the Direct Strategy and the Indirect Strategy, we can restate an observation that we made in Sect. 1.2.2. There seems to be an obvious motivation for the former while the motivation for the latter is less clear. Here is why. The Direct Strategy appears to be the best way to ensure that moral agents do the right thing. The decision-making process that it advocates is, after all, based on factors that are immediately morally relevant. Furthermore, it uses *all* of the pertinent information. It appears reasonable to ask, then, why moral agents should use the Indirect Strategy. Why should they ignore parts of the information that matters or choose based on a heuristic that pays no attention to the relevant right-making features at all? To be sure, heuristics may usually recommend the right act. But there is certainly no guarantee that they will do that all the time. As Gerd Gigerenzer writes, "[o]ne and the same heuristic can produce actions we might applaud *and* actions we condemn, depending on where and when a person relies on it." (Gigerenzer 2008, 4) Therefore, we can record that the Direct Strategy appears to be the *default position* and that any departure from it requires a justification.

Let us ask, then, what reasons there may be. Why should an agent facing a moral choice choose the Indirect Strategy rather than the Direct Strategy? This, it seems, is itself a choice problem to which consequentialist criteria of rightness can be applied.[78] So we can rephrase our question as follows: Under which conditions would a consequentialist criterion of rightness designate the use of the Indirect Strategy as the right choice?[79] There seem to be only two possibilities:

[76] Our description of the satisficing heuristic follows Pettit (1984, 166).

[77] See, also, Bykvist (2009, 154) and Kagan (1998, 67).

[78] For simplicity, our subsequent reasoning presupposes a *maximizing* consequentialist criterion of rightness.

[79] This way of phrasing the question is also proposed by Sumner (1987, 180).

- Firstly, the Direct Strategy is not an option. In that case, the Indirect Strategy automatically becomes the best strategy available.
- Secondly, it is reasonable to expect that applying the Indirect Strategy will not have worse consequences than acting on the Direct Strategy.

With these two possibilities in mind, we can formulate a plausibility requirement for Indirect Consequentialism.[80]

Plausibility Requirement for Indirect Consequentialism
A version of Indirect Consequentialism is plausible on the condition that it recommends the Indirect Strategy only in situations where (a) this is the only option or (b) it is reasonable to expect that its consequences will not be worse than the consequences of the Direct Strategy.

This plausibility requirement is rather abstract. It is not clear when it approves or disapproves of the Indirect Strategy. So let us work that out.

Let us turn, first, to clause (a). Under which conditions is the application of the Direct Strategy not an option? To answer this question, we merely need to look at the instructions that the Direct Strategy gives us. As we said above, it suggests that the agent should consider her options for acting and gather all information that is needed to project their consequences. Of course, she has to do all of this in due time. This description makes it evident that the Direct Strategy requires the agent to clear specific epistemic hurdles. Obviously, these hurdles will, on occasion, be too high. That will be the case, particularly when the following problems arise.[81]

- *The agent is epistemically restricted.* She does not know all her choice options or does not possess all the relevant information to project their consequences.
- *The agent is cognitively restricted.* She knows all of her alternatives and possesses all the relevant information to compute their consequences, but she is cognitively incapable of doing so.
- *The agent faces a time restriction.* She knows all of her options, possesses all information to project their consequences, and is cognitively able to do so, but she has insufficient time.

Plainly, when these problems obtain, the only thing the agent can do is to follow the Indirect Strategy. In such circumstances, she has to use a heuristic decision rule. However, she should not pick any old rule. It has to be a "fast and frugal heuristic ," as Gerd Gigerenzer emphasizes. Such a heuristic focuses specifically on properties

[80] Brad Hooker argues against such a plausibility requirement for Indirect Consequentialism (cf. Hooker 2003, 99–102). His defence, however, rests on the assumption that consequentialists need not accept, as we did above, that acts are the primary evaluative focal point of moral theories. Hence, what he says seems to me to beg the question against our reasoning in favour of the Plausibility Requirement.

[81] All of these problems have previously been mentioned and discussed in the literature. See, e.g., Appiah (2008, 51-62), Bales (1971), Bykvist (2009, 154–156), Hooker (2003, 142–143), Gigerenzer (2007, 202), Goodin (1995, 7), Goldman (2003, 11 and 90), Kagan (1989, 33), Nida-Rümelin (2011a, 30–31), Nozick (1993, 14), Scarre (1996, 13-14; 172-181), and Smart (1956, 346–347).

of acts that are easy to ascertain and fast to process. It can, hence, make do with limited information, limited cognitive processing capacities, and limited time, as the case may be. As such, it will allow the agent to make a choice, even when it is unclear which choice is right.[82]

Let us turn, then, to clause (b). Suppose now that the Direct Strategy is epistemically feasible. The agent, hence, has a choice between it and the Indirect Strategy. Under which conditions could it be reasonable to expect that the consequences of applying the Indirect Strategy are no worse than the consequences of the Direct Strategy? The answer appears to be that this can never be the case. Here is why. Suppose the agent holds a maximizing consequentialist criterion of rightness and applies the Direct Strategy. She fathoms all her options, investigates their consequences, and chooses one that is no worse than all other options. In that case, she ends up doing an act that is maximally good. In contrast, if she uses the Indirect Strategy, she makes her choice based on some heuristic which does not recommend acts based on whether or not they are maximally good. Therefore, there always seems to be a fair chance that the agent will end up doing an act that is worse than the maximally good act(s). Hence, it appears that the application of the Indirect Strategy is never motivated when the Direct Strategy is available.

It can be argued, however, that this reasoning is myopic in at least three ways. For one thing, it ignores the fact that real moral agents suffer from all sorts of "ailments of practical reason" (Pettit and Brennan 1986, 448) that can lead them to make mistakes in their moral calculations. Furthermore, it focuses only on the *outcome of choice* while neglecting the *choice process* (cf. Pettit and Brennan 1986, 444). Also, the reasoning above can be criticized for focusing only on the outcome of an isolated choice, while neglecting the importance of *choice structures* (cf. Nida-Rümelin 1993, §§36, 45). Therefore, the reasoning in favour of the Direct Strategy may be charged to ignore three important factors:

- firstly, possible mistakes in moral calculation;
- secondly, potential burdens of the direct choice process (as well as benefits of the indirect choice process);
- thirdly, the relative desirability of the choice structures that are selected by direct and indirect choice procedures.

Let us consider each of these points in turn.

First up, let us look at potential sources of error that derive from the contingent mental makeup of moral agents as human beings. Of course, we need to confine ourselves to a few examples of a very rich literature. One aspect of our moral mind, however, surely deserves to be mentioned. It is the influence of our self-interest on the way we reason morally. As we learnt in Sect. 2.2, our immediate moral intuitions can be skewed by self-interest which, as John Rawls notes, makes them dubious. He cautions us to be wary of intuitive moral judgements that we make "when we stand to gain one way or the other." (Rawls 1971/1999, 42) Given that some of our moral

[82] For a book-long treatment of fast and frugal heuristics see, e.g., Gigerenzer and Todd (1999).

sentiments can apparently be "switched on and off in keeping with self-interest," (Wright 1996, 13) this recommendation seems, in fact, reasonable. Now, we may say something similar about the kind of moral calculation that the Direct Strategy advocates. At any rate, many moral theorists have argued in that vein. Brad Hooker, e.g., says that "most of us are biased in such a way that we tend to underestimate the harm to others of acts that would benefit us." He concludes that the choice procedure favoured by the Direct Strategy will, therefore, "frequently lead us to make mistakes." (Hooker 2003, 143) We can infer from this that moral agents may often do better, as far as the morality of their act is concerned, if they follow a set of simple heuristics (e.g., the rules of common-sense morality), as the Indirect Strategy suggests.[83]

A further point to be noted in this context is that human beings tend to mispredict the welfare effects of their actions systematically. Henry Sidgwick already recognized this in his discussion of empirical hedonism. He writes that "when we are absorbed in any particular pleasant activity, the pleasures attending dissimilar activities are apt to be contemned."[84] (Sidgwick 1907, 145) In other words, when I am playing tennis, and I am fully immersed in my activity, I will tend to underestimate the pleasures that playing a musical piece on my guitar would bring me. Of course, the same holds *vice versa*. To put the point more generally, any prediction that I make about the quality of an anticipated experience is likely to be influenced by my current perspective. It may be influenced by the experiences that I am having at the moment or vivid memories that may come to mind.[85] As Sidgwick notes, this may distort my predictions about the effects that individual events will have on my well-being. In moral contexts, matters seem to be even worse. Given that I am a bad judge of my own well-being, I will do poorly, it seems, when it comes to predicting the welfare effects of my actions on others. The problem is compounded, after all, by the fact that I have to overcome an *interpersonal empathy gap*. I have to take into consideration the fact that other people have different desires, needs, and preferences. I am, hence, very likely to fall prey to "blind spots." (Bazerman and Tenbrunsel 2011) That is, I will tend to ignore features of a choice situation that are unimportant to me, but quite important to others. The bottom line of all this is that my moral performance may be better if I suspend moral calculation and stick to a set of well-tested rules, as the Indirect Strategy suggests.

Though we could add further empirical-psychological considerations, let us move on to the second point, viz. the potential benefits of the indirect choice process and the potential burdens of the direct choice process. On the Direct Strategy, recall, the agent is supposed to consider all her options, calculate the goodness of their

[83] Ng (1981) and Smart (1956) make this point, too.

[84] The phenomenon that Sidgwick describes has come to be recognized as the "intrapersonal empathy gap" (Loewenstein 2005).

[85] A case in point is the well-known phenomenon that people who won the lottery do not tend to be significantly happier on a long term basis than people who have recently become paraplegics (cf. Brickman et al. 1978).

consequences, and choose accordingly. It is obvious that this choice process can be both *time-consuming* and *costly*. And it is clear that when it is, the computational ease associated with the Indirect Strategy may make it a better choice than the Direct Strategy. As Alan Goldman emphasizes, "[t]he gains from continuously looking for better possibilities for action or for better results must," after all, "be weighed against the costs of doing so."[86] (Goldman 2003, 30) If, e.g., I face a choice between two options for acting, a_1 and a_2, and I happen to know that the costs of finding out which is better are greater than the difference in value between them, I am well advised to forswear the Direct Strategy. Instead, I should choose a simpler decision-making process, as the Indirect Strategy suggests. I may, e.g., pick a_1 or a_2 at random.

Here is a further consideration that may motivate the Indirect Strategy. It also has to do with the costs of moral calculation. Suppose I have two mutually exclusive options for acting. I can either do an act a_1 or choose not to do it. Since I am a direct consequentialist, I believe that I should go about my choice by calculating and evaluating the consequences of my options. But then it occurs to me that doing this is itself an act, a_2, which I may or may not choose to perform. As we learned in the previous paragraph, the benefits of calculating are sometimes not worth its costs. So I am not sure whether I should, in fact, perform a_2. Since I am a direct consequentialist, I believe that I should go about my choice by calculating and evaluating the consequences of my options. But wait a minute! Doing *that* is yet another act, a_3, that I can choose to do or not to do, and so on and so forth. Hence, when calculation is costly and time-consuming, the Direct Strategy seems to lead to *practical paralysis*. It will have me calculate forever. This, too, speaks in favour of the Indirect Strategy.[87]

A further problem with the Direct Strategy is that it destroys benefits that are, as Pettit and Brennan (1986) put it, "calculatively vulnerable." A benefit is calculatively vulnerable if it can be attained only on the condition that the agent suspend the kind of calculation that is suggested by the Direct Strategy. As we saw above, computational ease provides a case in point. Obvious benefits attach to it. It saves costs and frees up time that can be spent engaging in morally valuable endeavours. To attain these benefits, however, we must use the Indirect Strategy, that is, apply heuristics. Moreover, we must do so without regard for calculative considerations. This is easy to see. If we adopted the policy to use a heuristic only on the condition that its benefits outweigh its costs (i.e. in the form of suboptimal choices), we would have to engage in time-consuming and costly calculation. That is, we would defeat the purpose of using the heuristic in the first place (cf. Pettit and Brennan 1986, 445).[88]

[86] See also Anderson and Milson (1989) and Stigler (1961).

[87] The regress problem is mentioned, e.g., by Bales (1971, 258), Caws (1995, 324), Hardin (1988, 4), Kagan (1998, 66), Percival (2002, 138), and Pettit (1984, 170).

[88] For this general line of argument, see Railton (1984).

Examples of calculatively vulnerable benefits are legion. Consider, e.g., spontaneity. A certain degree of spontaneity can obviously be beneficial. It can help us to break with our daily routines and ensure that we discover new things that enrich our lives. However, we cannot realize the benefits that attach to spontaneity if we use an inflexible calculative decision procedure, as suggested by the Direct Strategy. The use of such a decision process is, after all, precisely the opposite of being spontaneous. Similar problems arise, e.g., regarding honour, courage, and integrity, as Pettit and Brennan (1986, 448) point out.[89] To the extent that the benefits associated with these personal attributes outweigh their costs, it seems reasonable for consequentialist moral agents to adopt the Indirect Strategy. But an even more significant calculatively vulnerable benefit – at least for proponents of CU – is *happiness*. It, too, is a calculatively vulnerable value, as the great utilitarian philosopher John Stuart Mill explains in his autobiography. (Mill, of course, does not use the term "calculatively vulnerable," which is of later provenance.) He reports that, though he "never, indeed, wavered in the conviction that happiness is the test of all rules of conduct, and the end of life", "this end was only to be attained by not making it the *direct* end." And he goes on to claim that when you "[a]sk yourself whether you are happy, (..) you cease to be so."[90] (Mill 2004, 82; emphasis added, NM) Every consequentialist who accepts the goal of happiness and accepts Mill's diagnosis that happiness is a calculatively vulnerable good is bound, then, to look favourably on the Indirect Strategy.

Let us move on, then, to the third point, viz. the issue of choice structures. To this end, let us assume, contrary to what we said above, that moral agents always had enough information to project consequences and were perfect calculators. Let us suppose, further, that the time and costs that are involved in their calculations were not an issue and that regress problems were, hence, not an issue either. Moreover, let us stipulate that there were no calculatively vulnerable goods. Should moral agents always prefer the Direct Strategy over the Indirect Strategy, then? The answer is no. Some philosophers, at least, argue in that vein. They make two distinct points.

The first point is this. Even if moral agents successfully figured out which act was best in any given situation, this would not ensure that the structure of their acts over time was optimal. To illustrate this point in the context of rationality theory, Julian Nida-Rümelin gives the following example.[91]

[89] For a comprehensive treatment of values that defy direct pursuit, see Elster (1983, 43–108).

[90] Henry Sidgwick calls this the "fundamental paradox of Hedonism." (Sidgwick 1907, 48) See, also, Gert (1998/2005, 261) who calls it the "utilitarian paradox."

[91] Similar illustrations are plentiful in the literature. Nida-Rümelin himself offers a further example of a city planning committee that is drafting a bus network (cf. Nida-Rümelin 1993, 128). If the planning process is such that each bus route is planned individually, it may happen, Nida-Rümelin claims, that the resulting network turns out to be suboptimal even if each planning decision was optimal. Pettit and Brennan (1986, 452) give a more down-to-earth example that involves dental health. They ask whether direct calculation of the best possible act would ever lead one to brush one's teeth and conclude that "calculation after every meal would always fail to elicit a walk to the bathroom. "Yet another illustration is given by James Buchanan. It is about "individual choice

> [A] long-standing smoker asks himself whether he ought to give up smoking. The smoker's subjective valuations are such that the time integral of his value function increases monotonously the sooner he gives up smoking. (. . .) If he still does not give up smoking, his case seems to be a clear instance of 'akrasia' or weakness of the will : weighing up all different evaluative aspects, he knows what would be the best thing to do, but he still cannot get himself to do it. This interpretation, however, is not conclusive. It might just as well be the case that the smoker acts *consequentially rational*, if he does not give up smoking – despite the fact that the time integral over his valuation function increases monotonously the sooner he gives it up. (Nida-Rümelin 1997a, 140; emphasis in the original)

The reason for the impasse at which the consequentially rational smoker finds himself lies, as Nida-Rümelin explains, in the fact that he never actually faces the option to stop smoking altogether. (If he did, he would, of course, choose that option.) Instead, he faces many individual choices over an extended period. At each point, the smoker can decide to smoke yet another cigarette or turn it down. Nida-Rümelin offers an argument to the effect that it might be for the best, given the smoker's long-term preferences, if he chose to smoke at each point in the series of choices, such that he ends up with a *choice structure* which, *ex hypothesi*, he disprefers.[92] The precise argument shall not concern us here. Let us simply suppose that Nida-Rümelin's reasoning goes through. In that case, it would follow that the smoker would be better off sticking to a simple rule – *Don't smoke*! – than if he calculated the utility of each cigarette. It is reasonable to suppose that analogous cases occur in moral choice situations (cf. Goldman 1978). We should expect, that is, that there are sequences of *moral* choices where direct, consequential deliberation on a choice-by-choice basis would produce a series of decisions which, as a whole, leads to suboptimal consequences. Whenever this is so, it would provide further motivation for the Indirect Strategy.[93]

Let us briefly discuss a further interpretation of the smoker's problem. Unlike Nida-Rümelin's reading of the case, it trades on *time preferences*. We start by distinguishing three points in time: t_0 (=now), t_1, and t_2. Let us suppose that at t_0 the smoker is pondering whether he should smoke another cigarette at t_1. We may suppose that there are two considerations which factor into his decision, viz. the pleasure of smoking at t_1 and the damage in health that he will suffer at t_2 as a result of smoking at t_1. (For simplicity's sake, we suppose that every cigarette

behavior in eating" which, if done "on a meal-by-meal basis, often leads to obesity, a result that is judged to be undesirable." In line with Nida-Rümelin's above assertion, Buchanan stresses that the "individual arrives at this result (..) through a time sequence in which each and every eating decision seems privately rational. No overt gluttony need be involved, and no error need be present. At the moment of each specific choice of food consumption, the expected benefits exceed the expected costs." (Buchanan 1975, 189) I am thankful to Julian Müller who made me aware of Buchanan ideas.

[92] Nida-Rümelin, therefore, claims that an agent who applies a consequentialist criterion of rightness directly in each choice is "structurally irrational."

[93] Note, however, that Nida-Rümelin does not consider his argument as a potential motivation for Indirect Consequentialism, but as an argument against consequentialism, *period*. In order to be charitable to indirect consequentialists, we assume, however, that it can be interpreted in that way.

causes a perceptible health damage.) Let us stipulate that, according to the smoker's valuation, the pleasure of smoking at t_1 has a value of d, while the benefit of not having to suffer the associated health damage at t_2 has a value of e, where $e > d$. Given this assumption, it seems to make no sense for the smoker to smoke at t_1 because, by doing that, he obtains the smaller of the two rewards, viz. d. But this observation does not take time preferences into account. People tend to care less about the prospect of a reward the later it eventuates. Amongst economists, this phenomenon is known as "discounting." The present value, PV, of a reward is a function, f, of its value, v, its time of eventuation, t_i, and the current point in time t.

The Present Value of a Reward

$$PV = f(v,\ t_i,\ t)$$

The discount function f casts new light on the smoker's problem. It gets clear that when he ponders at t_0 whether he should smoke at t_1 he does not simply compare the respective rewards e and d. Rather, he compares the present values of e at t_2 with the present value of d at t_1, as seen from his present vantage point t_0. Let us stipulate, then, that $f(e, t_2, t_0) = b$ and that $f(d, t_1, t_0) = a$. If $a \geq b$, it appears to make sense, then, for the smoker to smoke at t_1.[94] Let us assume, however, that this is not the case. Let us assume, that is, that the smoker's long-term assessment at t_0 is such that he prefers the greater, later reward (health benefit) to the smaller, earlier reward (pleasure from smoking); formally, $b > a$. On this assumption, it appears rather unreasonable for the smoker to smoke at t_1.[95] It seems that he should resolve not to do it. Note, however, that whether or not he will feel inclined to keep to his resolve at t_1 depends not on his present valuations at t_0. It depends, rather, on the value of $f(e, t_2, t_1) = c$. Whether, come t_1, the smoker will still prefer long term health to instantaneous pleasure depends on the shape of his discount function f. If it is time-consistent, "later preferences 'confirm' earlier preferences," (Frederick et al. 2002, 358) such that $c > d$. But if f is time-inconsistent, the opposite might be true. That is, it may be that $d > c$. As Ainslie (1975) points out, people's actual time preferences can often be modelled using "hyperbolic discount functions" which are, indeed, time-inconsistent.[96] On such a function, the valuation of a greatly delayed future reward increases only slightly with t but rises sharply in small delay periods. This can have the effect that the graphs of $f(e, t_2, t)$ and $f(d, t_1, t)$ intersect shortly before t_1, as depicted in the figure below. Our smoker may have precisely this problem. Perhaps he recognizes that health at t_2 is more valuable than pleasure at t_1, formally $e > d$. Perhaps his long-term assessment at t_0 is such that he attaches a

[94] This, of course, holds only to the extent that the discounting future rewards can be seen as rational. This idea can be disputed. For a critique of discounting future rewards see, e.g., Rawls (1971/1999, 259–262) and Parfit (1986, Appendix F). See also the very differentiated discussion offered by Broome (1999, 44–67).

[95] For an argument to that effect, see Nozick (1993, 16–17).

[96] See, also, Ainslie (1986/2000 and 1999).

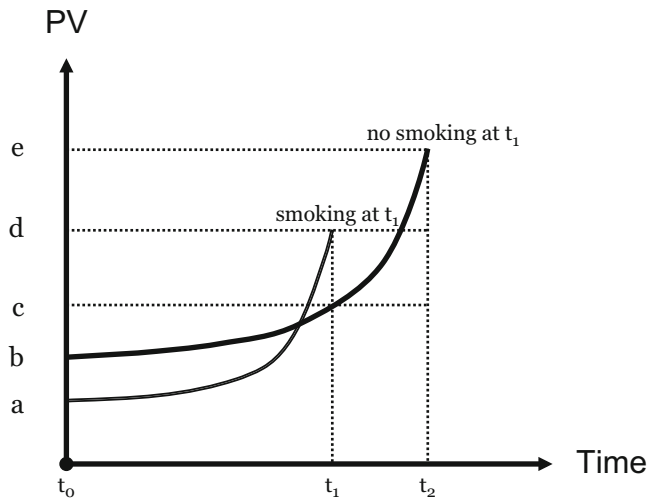

Illustration 4.1 Hyperbolic discounting

greater present value to health at t_2 than to pleasure in t_1, formally $b > a$. Shortly before t_1, however, the discount functions $f(e, t_2, t)$ and $f(d, t_1, t)$ intersect, such that the smaller earlier reward strikes him, at t_1, as more attractive than the greater, later reward, formally $d > c$ (Illustration 4.1). If this is so, the smoker obviously has a severe problem on his hands. We only looked at a short sequence of events that lead him to smoke just one more cigarette. It is clear, however, that this sequence is likely to keep on repeating itself. Just having smoked a cigarette, the smoker will come to realize that he should stop smoking because the present value of the health benefits from abstinence outweighs the present value of the pleasures from smoking in every single instance. He will, hence, resolve to stop. However, on the very next occasion, he will, nevertheless, yield to the temptation. He will have another smoke, and the process will start over. Obviously, a simple rule, as is recommended by the Indirect Strategy, can help the smoker to overcome this problem. If, instead of optimizing on each occasion, he simply follows the rule not to smoke, he will achieve a better result.[97] Since nothing depends on the details of the smoker's case, we can state, more generally, that any agent who discounts future benefits hyperbolically and optimizes her valuations at each point in time, will not be very likely to accomplish long-term goals and should, therefore, follow the Indirect Strategy. This, of course, extends to the moral case, where a moral agent sympathetically maximizes other people's present valuations since these may also be the result of hyperbolic discounting. We can conclude, then, that the second interpretation of the smoker's case also provides a potential motivation for the Indirect Strategy.

[97] This, of course, does not address the psychological problem of how the smoker can bring himself not to smoke. For suggestions on this issue, see Nozick (1993, 17–18).

Above, we said that two points can be made to motivate the Indirect Strategy even if time and costs of calculating are not an issue and there are no calculatively vulnerable goods. We just considered the first aspect, viz. potential problems with the intertemporal structure of optimal choices. The second point that can be made in support of the Indirect Strategy, even when the conditions for applying the Direct Strategy are favourable, has to do with interpersonal choice structures. Here is a sketch of the reasoning: The consequences of what we do depend not only on our acts. They also depend in large part on what other people do.[98] A standard illustration of this fact is our behaviour in traffic (cf., e.g., Buchanan and Brennan 2000, 10–15). Let us suppose that each of us can choose to drive on the right or left side of the road. There are two optimal scenarios (both from an individual and a moral standpoint). We should either all drive on the right or all drive on the left. Now, if all of us used the Direct Strategy to figure out what to do, we would almost certainly come up with different conclusions. If we calculated the best course of action all by ourselves, some of us would end up driving on the right. And some of us would end up driving on the left. The outcome would be mayhem. Constantly, there would be collisions, injuries, and deaths. In contrast, if we all adopted and committed to a simple rule, e.g., to drive on the right, the expected outcome would be much better. Presumably, there would still be accidents. However, they would certainly be less frequent. This simple illustration shows that there can be situations in which moral agents would be well advised not to engage in direct calculation and simply follow rules, as the Indirect Strategy recommends.[99]

I believe that we have said enough about the circumstances in which the Indirect Strategy may potentially be reasonable. Now, we need to establish that they do not obtain in trolley cases, as we characterized them in Sect. 2.4.1. To show this, let us briefly summarize them.

The first three conditions that might motivate the Indirect Strategy had to do with clause (a) of the Plausibility Requirement for Indirect Consequentialism. The agent, we said, may not know all her choice options or may not possess all the relevant information to project their consequences. She may be cognitively incapable of doing the necessary calculations. Or she may lack the time to do so. These problems can never arise for the agent in a trolley case. This is ruled out by Characteristic 5. The agent is assumed to know what her options are and what their consequences will be. Hence, she does not lack any information. She does not need to calculate. Moreover, since there is no need for calculation, she can choose instantaneously. So time is not an issue either.

Let us turn to the conditions that derive from clause (b). We said that one reason the Indirect Strategy may not have worse consequences than the Direct Strategy is that real moral agents will make a lot of mistakes when they calculate along the lines of their criterion of rightness. And we said that obeying simple rules, e.g. the rules of

[98] An ethical theory which focuses on spelling out the moral implications of this fact is Order Ethics (see Luetge and Mukerji 2016).

[99] The interpersonal function of rules is also emphasized by Nozick (1993, 9–12).

common-sense morality, may, therefore, turn out to have morally better effects. This possibility, too, is ruled out by the design of trolley cases. Since the agent knows the consequences of her actions in advance, she does not need to make any calculations and, hence, cannot make any mistakes in calculating them.

We said that the second reason (b) may be fulfilled is that the costs of applying the Direct Strategy may not be justified by the gain in accuracy as compared with the Indirect Strategy. Again, that is not true in trolley cases. The costs and time that are involved in calculating are not an issue since the agent does not have to do any calculating. This also rules out a vicious *regressus ad infinitum*. The necessity of this regress, recall, is shown under the assumption that the agent has to calculate.

The next potential justification for going indirect is the existence of calculatively vulnerable goods, e.g. happiness. Such goods play no role in trolley cases. To see this, recall Characteristic 4. It is the assumption that the agent's *act* uniquely determines the outcome. The agent turns the trolley to the right: one person dies. The agent does nothing: five people die. Whether she *aimed* at killing the one or the five is irrelevant. What determines the outcome is what the agent *does*.

Finally, let us turn to choice structures. We considered two different types of cases. The first concerned choices over time. We looked at the example of the smoker who wants to stop smoking. It is clear that the kind of complication that he experiences plays no role in a trolley case. This particular kind of problem can only arise for a person who faces many individual decision problems over an extended period. The agent in a trolley case, however, faces a single decision problem at a particular point in time. Hence, even if interpersonal choice structures can motivate the Indirect Strategy, this motivation is unavailable in a trolley case.

The second type of case we considered concerned interpersonal choice structures. We said that the Indirect Strategy may be justifiable in cases where many agents have to coordinate their behaviour to collectively bring about the desired outcome because individual deliberation may achieve worse results than collective rule following in interpersonal choice settings. This motivation, too, is out of the question in trolley cases, where there is only a single agent whose choice alone determines the outcome of the case.[100]

Let us recap and conclude, then. We considered some plausible motivations for the Indirect Strategy and found that all of them hinged on conditions that do not obtain in trolley cases. The Indirect Strategy, it seems, can, therefore, not plausibly be motivated in them. This, in turn, implies that all plausible versions of Indirect Consequentialism will prescribe the Direct Strategy in trolley cases. In other words,

[100]It may be objected that in certain trolley cases (or trolley-like cases) there are multiple agents. Consider Bernard Williams's famous case about Jim and the Indians (cf. Williams 1973, 98–99). This case, it seems, has all characteristics of a trolley case that we identified above. Jim faces the choice between shooting and not shooting an Indian. If he does not shoot him, Pedro will kill him and another 19 Indians, who would otherwise have been spared. In this case, one might say, there are two agents, Jim and Pedro. But this appearance is deceiving. In fact, there is only a single agent, viz. Jim. What he does *determines* the outcome of the case. Pedro is assumed to have made his choice, such that his action can be taken as fixed.

they will make the same practical recommendations as otherwise identical forms of Direct Consequentialism. As we remarked at the beginning of this section, this allows us to ignore the distinction between Direct and Indirect Consequentialism in the trolleyological investigation that we shall embark upon in Chap. 5.

4.2.1.3 Alternative Theories of Well-Being

It is beyond question that the plausibility of a consequentialist moral theory depends, to a large extent, on its theory of the good. As Russell Hardin notes, however, critics of consequentialism[101] "often make their task easy by assuming an implausible value theory and directing criticism at the implications of their value theory." (Hardin 1988, xvi) In other words, their criticisms, thinks Hardin, beg the question against consequentialist theories with more plausible axiologies.

We do not want our case against consequentialism to be vulnerable to this sort of objection. We want to acknowledge, that is, that consequentialists have a broad range of value theories and, in particular, theories of well-being at their disposal. Moreover, we want to be alive to the fact that they can use their broad axiological repertoire to defend themselves against criticisms. At the same time, though, we want to keep our discussion within manageable bounds. If possible, we want to avoid having to go through the whole arsenal of axiological theories that consequentialists can draw on. Trolley cases can help us to achieve both goals at once because they allow us to ignore competing theories of well-being.

To show this, we can use a strategy that is similar to the strategy we employed above. In the previous two sections, recall, we argued that it is justifiable to ignore the distinction between Subjective and Objective Consequentialism as well as the differentiation between Direct and Indirect Consequentialism. To support this, we said that moral doctrines on opposite sides of the respective divides coincide in their implications if they are otherwise identical. Here, we cannot use the same argument. This is because consequentialist theories which differ regarding their conception of well-being *can* have different implications in trolley cases. We can show, however, that it is possible to model trolley cases so that this difference does not matter. We can set them up in a way that ensures that their implications coincide.

How do we show this? The first thing we need to establish is that trolley cases possess an important characteristic that we have ignored up until now. They are incomplete in the sense that they typically do not provide all the information that is needed to apply a moral theory. To see this, we need some illustrative material. So let us return to one of the trolley cases that we considered earlier, viz. *Edward's Case* from page 43. In *Edward's Case*, recall, Edward is the driver of a trolley whose brakes have just failed. Five people are on the tracks right ahead of him.

[101] Hardin's remark is, in fact, about critics of utilitarianism. But since he has a very broad notion of utilitarianism that approximates our understanding of consequentialism, it is fair, I think, to take what he says as a comment about critics of consequentialism in general.

The only way for him to avoid running over them, thereby killing them, is to steer the trolley onto a spur that leads off to the right. If Edward does this, his trolley will, however, run over and kill one person. What should he do? To see that this description of the case is incomplete, let us ask what, according to CU, would be the right thing for Edward to do? From a classic utilitarian standpoint, it seems, the answer is clear. Edward should turn the trolley. He should kill the one rather than the five. However, let us be a bit pernickety here. Whether an act is right, according to CU, depends on whether it maximizes the sum of happiness in the world. Which act fulfils this condition in *Edward's Case*? Strictly speaking, we cannot tell. All we know is the number of deaths that each option would cause. Of course, it is not too far to seek that one death has a less negative effect on the happiness sum than five deaths. After all, it seems likely that the combined hedonic value of five lives will be greater than the hedonic value of a single life. But this need not be so. Perhaps the one individual on the spur would go on to live a life so pleasurable that it would outweigh the combined pleasures of the other five persons' lives. This, admittedly, is not the most natural interpretation of the case. However, it is a *possible* one. We cannot tell, therefore, what CU implies regarding *Edward's Case*. Its description does not meet the informational requirements of CU. This is a feature of trolley cases more generally.[102]

The incompleteness of trolley cases has an important consequence. Depending on how we fill the gaps in the description, we get various possible interpretations. On some of them, the implications of consequentialist theories that differ only regarding their conception of well-being – call them "w-different" – come apart. On others, they do not. Let us first consider an interpretation of *Edward's Case*, where the implications of two w-different theories coincide. To this end, let us introduce a new moral theory.

Preference Utilitarianism (PU)
An act is right if and only if it maximizes the overall fulfilment of preferences of all individuals who are capable of forming preferences.[103]

PU is identical, then, to CU except in its conception of individual well-being.[104] Unlike CU, it does not adopt Welfare Hedonism. Rather, it commits to the following view.

[102]This point is, I think, made by Shelly Kagan when he says that "typically when we think about cases, we are only thinking about *kinds* of cases." (Kagan 2001, 61-62; emphasis in the original)

[103] Hare (1981), Harsanyi (1955 and 1977a), and Singer (1979/1993) have, e.g., proposed such theories.

[104]In fact, PU also has a different version of Universalism. But this is a consequence of the fact that the Universalism of a consequentialist doctrine is systematically connected to its conception of well-being.

Welfare Preferentism
The well-being of an individual is measured by the extent to which her preferences are satisfied.[105]

Let us apply CU and PU to the following Interpretation of *Edward's Case*.

Edward's Case – Interpretation 1
Edward can turn the trolley to the left, killing person A, or carry on ahead, killing persons *B*, *C*, *D*, *E*, and *F. According to both Welfare Hedonism and Welfare Preferentism*, A *has a life ahead of herself that is as valuable as each of the lives of* B, C, D, E, *and* F.

According to the italicized bit that we added to the case, the amount of sensory happiness that is lost in the case where *A* dies is the same as in the case where *B*, *C*, *D*, *E*, or *F* dies. The same holds for the amount of preference satisfaction. Obviously, then, CU implies that Edward should kill *A*. Granted, if he does that, the sensory happiness that that individual would have enjoyed in her life is lost. But *B*, *C*, *D*, *E,* and *F* will enjoy five times as much. PU implies the same. We can, after all, say the same when it comes to preference satisfactions.

Now, let us consider an alternative interpretation of *Edward's Case* where the implications of CU and PU diverge.

Edward's Case – Interpretation 2
Edward can turn the trolley to the left, killing person *A*, or carry on ahead, killing persons *B*, *C*, *D*, *E* and *F. On Welfare Hedonism,* A's *life is six times as valuable as B's, C's,* D's, E's, *and* F's *life*. On Welfare Preferentism, however, A's *life is as valuable as each of the lives of* B, C, D, E, *and* F.

On this interpretation of the case, CU judges that Edward should kill the five because this will maximize the balance of sensory happiness. PU, however, implies that Edward should kill *A* because this will maximize overall preference satisfaction. As it turns out, then, the implications of w-different consequentialist theories can come apart in trolley cases. It is very easy to see, however, how we can avoid this. Obviously, the implications of CU and PU can diverge in the second interpretation of the case because, on the welfare-hedonistic and welfare-preferentist outlook, the lives of the individuals have different relative values. To forestall this possibility, we only need to assume that the values of the lives of the people involved in the case have the same value on all plausible conceptions of well-being. If we make this assumption in our trolleyological investigation below, all w-different consequentialist theories will imply the same deontic verdicts, and we can safely ignore the distinction between them.

[105]There are two broad interpretations of this thesis (cf. Griffin 1986, 10–15). Actualist Welfare Preferentism is the view that the measure of well-being is the extent to which an individual's actual preferences are fulfilled. Ideal Welfare Preferentism is the idea that the measure of well-being is the extent to which an individual's ideal or well-informed preferences are fulfilled.

4.2.2 Non-Maximizing Variants

Before we proceed, let us, once again, review the progress that we have made in this chapter. We started by taking on step (i) of our methodic procedure, FRA_2. To this end, we looked at a paradigmatic consequentialist moral theory, viz. CU. We factorized it into eight logically independent components, viz. Maximization, Welfarism, Summation, Equal Treatment, Universalism, Welfare Hedonism, Objectivism, and Directness. This laid bare the logical structure that unites all moral theories of the consequentialist family. We can characterize this family as a system of eight determinable components. Each member of the family adopts exactly one determinate component in place of each of these eight determinable components, viz. either the CU-components I just mentioned or alternatives to them. To show that we can reasonably reject the consequentialist family, we need to establish that all possible combinations of determinate components consequentialists might endorse are flawed. Since this is a daunting enterprise, we considered possible methodological shortcuts. That is, we looked for ways to shorten the discussion that still allow us to draw the desired conclusion. To this end, we said that we should try to show, in step (ii) of our inquiry, that we can put aside at least some of the determinable components identified in step (i). We did that in the three previous sections, where we showed that the trolleyological nature of our inquiry allows us to ignore the distinction between Objective and Subjective Consequentialism, Direct and Indirect Consequentialism, as well as different accounts of individual well-being.

In the remaining sections of this chapter, we shall take on step (iii) of our investigation. That is, we shall survey the alternatives to each of the CU-components. Logically speaking, there are innumerable such alternatives. As we said above, however, it seems reasonable to suppose that only a few of these alternatives deserve our attention, viz. those which seem motivated. Which ones are they? As Tännsjö (2002) thinks, alternative versions of consequentialism can largely be regarded as reactions to the problems that have been associated with CU. We can see them as attempts to remedy these faults. If this is true, we reasoned, it should be possible to detect the most interesting variants of consequentialism by considering the most pressing concerns that may incite a moral theorist to abandon CU and to adopt a different view. Alternatives to the CU-components, it seems, should suggest themselves in the process. This said, let us begin our investigation of the alternatives to the paradigmatic consequentialist components, starting with Maximization.

Maximization is, as we know by now, the claim that an act is right if and only if it maximizes the good. Non-standard forms of consequentialism can diverge from this claim and endorse a contrary view. The main competitor of Maximizing Consequentialism is Satisficing Consequentialism. It denies the tight logical connection between an act's property to maximize the good and its rightness. We may formulate the view as follows:

Satisficing
An act is right if and only if it is good enough.[106]

Slote (1984) first proposed Satisficing as an alternative to Maximization and subsequently defended it (see Slote 1985a and 1989). In the meantime, Slote seems to have abandoned his idea and appears to have converted to virtue ethics (cf., e.g., Slote 1997/2007). Satisficing (and with it the idea of a Satisficing Consequentialism) has survived, however. It has been taken up and defended (at least partly) by a number of philosophers, e.g. Dreier (2004), Hurka (1990, 2004), Turri (2005), and Vallentyne (2006). Other authors have endorsed forms of Satisficing Consequentialism, but have referred to them using different labels.[107]

4.2.2.1 Motivation

Much has to be said to clarify the idea of Satisficing. Before we do that, however, let us consider some problems about Maximization that might motivate Satisficing as an alternative component. We can make least three points in this connection.[108]

The first aspect to be considered is that a maximizing doctrine, such as CU, appears to be incompatible with the autonomy of the moral agent. As Williams (1981) remarks, our lives are built around certain "ground projects" which give us reason to carry on living. Since CU allows us to do only that act which produces the best possible state of affairs, it limits the range of permissible options to only a single act (unless, of course, two or more acts happen to tie regarding goodness). This appears to constrain our moral freedom to an unreasonable extent because it prevents us from pursuing our ground projects. This, admittedly, is not a matter of necessity. Certain people may choose to live a life which most contributes to the overall good. Others may choose plans which, coincidently, turn out to have the same effect. Normally, though, a consequentialist doctrine, such as CU, which subscribes to Maximization, seems to constrain moral agents in the autonomous pursuit of their projects and will make it impossible for them to live a life of their choosing. Interestingly, CU constrains not only self-interested pursuits but also selfless undertakings. To illustrate, let us assume that Jones wants to donate a lung to his dear friend Smith. Supposing perfect comparability of well-being, let us assume that the facts about Jones's and Smith's life qualities are such that Jones's act will

[106] As Mulgan (2001b, 43) points out, there are two ways of using the notion of satisficing. It can be used as an element of a theory's theoretical component that replaces Maximization. Alternatively, it can be used, as we have seen in Sect. 4.2.1.2, as a decision procedure (while retaining the maximizing criterion of rightness). Satisficing, as it is expressed here, is a view that immediately applies to the moral status of acts. It is part of the consequentialist criterion of rightness and not an alternative decision procedure.

[107] Robert Elliot, e.g., has discussed a version of consequentialism he refers to as "improving consequentialism." (Elliot 1997, 46 and 2003, 184)

[108] All of these points are mentioned by Slote (1984). For a further discussion, see Mukerji (2013c, 302–303).

lower his own lifetime well-being to a greater extent than it will increase Smith's. In that case, donating the lung would, in fact, lower the overall amount of well-being. On CU, doing it is, hence, morally forbidden. As Michael Slote has pointed out, this is surely out of touch with common-sense. As he explains, moral agents generally possess an "agent-sacrificing option" in a situation where they can sacrifice their own good for the sake of others (cf. Slote 1985a, 11–12). This would clearly be so in the case we just considered.

The second important criticism of CU that is often attributed to its maximizing nature is the fact that it apparently leaves too little room for supererogatory acts. To explain this point adequately, we have to introduce a number of minimal requirements for supererogation that seem, intuitively, to make sense.[109]

Minimal Requirements for Supererogation
Let S be a choice situation in which the agent has a set of n options for acting $A = \{a_1, a_2, \ldots, a_n\}$. Furthermore, let A^P be the subset of A that contains all and only the permissible options in A and let A^E be the subset of A^P that is best for the agent. Moreover, let the term $W_{\text{-}agent}$ represent the overall welfare minus the well-being of the agent.
An act, a_i, is, then, not supererogatory unless

(i) it is an element of A^P (Permissibility Requirement),
(ii) it is not an element of A^E (Self-Sacrifice Requirement),
(iii) there is an act, a_j, that is also an element of A^P and affects $W_{\text{-}agent}$ less favourably overall than does a_i (Altruism Requirement).[110]

I believe that these conditions make sense. The Permissibility Requirement is certainly necessary because it would seem unreasonable to regard a morally forbidden act as supererogatory. Moral permissibility appears, hence, necessary.

The Self-Sacrifice Requirement seems to be necessary, too. An act should not be regarded as supererogatory if it is the one act out of all permissible acts that most promotes the well-being of the agent. Intuitively, such acts do not deserve praise, which is an essential characteristic of the supererogatory. There is a second reason for adopting the Self-Sacrifice Requirement. When it is combined with the Permissibility Requirement, it implies that supererogatory acts are necessarily morally optional, which is intuitively correct. For it cannot, logically, be the case that an act is an element of A^P, not an element of A^E, and yet not optional.

Lastly, then, even if an act, a_i, is permissible *and* involves a self-sacrifice on the part of the agent, it should not be seen as supererogatory, if there is no alternative, a_j, that produces less $W_{\text{-}agent}$. Intuitively, if there was no permissible alternative to a_i that produces less $W_{\text{-}agent}$, then a_i would produce the least amount of welfare for others that is consistent with duty. It would not be commendable.

[109]Note that, in formulating these requirements, we suppose that supererogation is an "all-or-nothing" affair. That is, we assume that an act is either supererogatory or not. As Martin Rechenauer has pointed out to me in personal conversation, there may be contexts in which it may be desirable to use a comparative idea of supererogation. For our purposes, however, I believe that this simpler idea suffices.

[110]I am indebted to Douglas Portmore and Martin Rechenauer with whom I discussed my ideas on supererogation.

Now, is CU compatible with the idea that there are acts which fulfil all three of these requirements? It seems that it is. Consider the following example.[111] Jones has a candy bar. He can do two things with it. His first option is to give it to Smith. His second option is to eat it himself. From the standpoint of overall well-being, let us assume, the two options are equally good. That is, if Jones were to eat the candy bar, this would increase his well-being by the same amount that it would increase Smith's well-being if Smith were to eat the candy bar. Now, can a proponent of CU regard Jones's first option as supererogatory? To show that this is, at least, a possibility, we need to check whether the act of giving the candy bar to Jones meets all three requirements for supererogation. Firstly, giving the candy bar to Jones is permissible from the standpoint of CU. It is not the case, after all, that there is another option which would promote overall well-being to a greater extent. Both options tie regarding their goodness and are both permissible. The Permissibility Requirement is, hence, fulfilled. Secondly, eating the candy bar is best for Jones. Giving the candy bar to Smith involves a sacrifice. So the sacrifice condition is also satisfied. Thirdly, there is a permissible option for Jones that is better from an altruistic standpoint, that is, better for Smith. Giving the candy bar to Smith is permissible. It is better for Smith. So the Altruism Requirement is met as well. As it turns out, then, there seems to be no reason to suspect that CU cannot accommodate supererogatory acts.

What, then, is the problem with CU? The problem, it seems, is not that CU cannot accommodate *any* supererogatory acts. What seems troublesome is the fact that it allows too *few* of them. It, hence, denies acts that appear to be clear instances of supererogation their rightful status.

To see this, consider the following scenario. Suppose Smith's and Suzy's kidneys are failing. Jones has the same tissue type and would be a suitable donor for both of them. Both Smith and Suzy live a much happier life than Jones. Jones has four options for acting. He can give neither Smith nor Suzy a kidney. In that case, he will live, and Smith and Suzy will die. He can donate a kidney to Smith. In that case, only Suzy will die. He can donate a kidney to Suzy. If Jones does that, only Smith will die. Moreover, he can donate both his kidneys, one to Smith and one to Suzy. That way he would die himself, but Suzy and Smith would both live. Intuitively, all of these options are permissible, and the latter three are supererogatory. However, on CU, no act is supererogatory. There is one best act, viz. giving up both kidneys. Furthermore, since, on CU, only the best Jones can do is permissible, there is only one permissible act. This rules out that he can fulfil the Self-Sacrifice Requirement in this situation. According to CU, there is, hence, no room for supererogation in this situation. There is just bare duty! As it turns out, then, CU does not class an act as supererogatory, even in a case where it seems entirely clear that it should be so classed.

A third point to be made against CU is that it might require unreasonable moral sacrifices from the moral agent. Not only does she apparently not get to choose

[111] Paul McNamara suggested this example to me in personal conversation.

what she does and is undeserving of moral praise. She might sometimes be morally *required* to do things, in the service of morality, that are drastically at odds with her own self-interest and even with the plain wish to survive. (The previous example makes this clear as well. The only permissible option is for Jones to give up both kidneys.) Again, this is not a matter of necessity. It might be that the agent does best from the standpoint of a maximizing conception of morality if she just promotes her own self-interest.[112] But it is likely that CU's maximizing aspect will frequently lead to unreasonable moral demands.

These points against CU, I take it, may motivate Satisficing Utilitarianism (SU) which accepts the same components as CU, but substitutes Maximization by a version of Satisficing.

Satisficing Utilitarianism (SU)
An act is right if and only if it sufficiently promotes the sum total of happiness of all sentient creatures.

SU is likely to give the agent more moral freedom. It does not require her to do the best act. She is morally free to do any act that is good enough. If the level of goodness that counts as good enough is set sufficiently low, it gives her a range of permissible options that she may choose. Within limits, this makes it possible for the agent to choose among her options from amoral motives that may come, e.g., from her life projects. Obviously, SU also makes more room for the supererogatory. As we saw above, only optional acts can be called supererogatory. CU's problem is that it seems to allow too few of them. SU allows more and increases, hence, the number of acts that may qualify as supererogatory. As regards the issue of demandingness, it cannot be ruled out that satisficing forms of consequentialism are also quite taxing on the agent. After all, all of the options that are good enough may be associated with tremendous self-sacrifice (cf. Vallentyne 2006, 27). It seems, however, that SU is, if anything, less demanding than CU since it permits the agent to do the morally best act, as CU does, and allows further options in addition to that.

This much, at any rate, can be said to motivate Satisficing. However, Satisficing may still seem implausible because it appears to be straight-out irrational. In Sect. 4.1.4, we said that Maximization appears to be rational because of an analogy to the theory of rational choice. In rational choice theory, the principle of expected utility maximization looms large. Since it is *prima facie* not clear why rational choice theory and moral theory should not be analogous, the idea that the agent should maximize the good appears to be the only rational option as far as moral theory is concerned. To defend Satisficing against this concern, one can do one of two things. Firstly, one can argue that it is not necessarily irrational if an agent fails to maximize utility in a choice-theoretical setting. Secondly, one can challenge the analogy between moral theory and choice theory.

Defenders of Satisficing have commonly chosen the first strategy. Michael Slote, in particular, has argued that the utility-maximizing conception of rationality is,

[112] Utilitarian defenders of the market economy have commonly argued that way.

in fact, misguided. To this end, he has drawn on examples of purported rational satisficing that have received some attention in the economics literature (e.g. Simon 1955 and 1959; Cyert and March 1963/1992). This strategy is implausible, however. As we saw above, the very idea of expected utility maximization is wedded to axioms of rationality. It is not clear which axiom Slote would propose to reject.

There may be a more promising way to salvage Satisficing.[113] One may attack the supposed analogy between moral theory and rational choice theory. As we saw above, the latter explains the idea of utility in reference to the theory's axioms about rational preference structures. It represents the utility of an object of choice by a real number that signifies the place of that object in the agent's preference ordering. The utility maximizing object is simply the one that is rational for her to choose, given her preferences. Outside this theoretical context, however, the notion of utility is meaningless. The question "What maximizes my utility?" *is* the question "What ought I, rationally, to do?" Now, the idea of the good, which is utility's analogue in moral theory, may be different. Perhaps, the question "Which act would be best to perform?" can be answered independently of the question "What ought I, morally, to do?"[114] John Rawls, at any rate, may be interpreted as gesturing towards that idea when he says that a consequentialist (or teleological) theory

> accounts for our considered judgments as to which things are good (our judgments of value) as a *separate class* of judgments intuitively distinguishable by common sense, and then proposes the hypothesis that the right is maximizing the good *as already specified*. (Rawls 1971/1999, 22; emphases added, NM)[115]

At this stage, it is not important to decide whether Satisficing is ultimately a defensible moral claim. The only thing that was important in this section was to ascertain that it is positively motivated and not hopelessly implausible. We can, therefore, put it on the list of alternative determinate components that consequentialists might embrace. We are, however, not done yet. For what we have said so far is far too unspecific. As it will turn out below, there are various forms of Satisficing. Some of them can be ruled out immediately because they appear to make no sense. Others are at least not immediately nonsensical. These, then, are the ones we shall consider when we finally connect the dots of our case against consequentialism in Chap. 5.

4.2.2.2 Options

Satisficing views are commonly classed as forms of Comparative Satisficing Consequentialism and forms of Non-Comparative Satisficing Consequentialism, depending on how they interpret the term "good enough."

[113] We follow a suggestion by Dreier (2004).

[114] For a criticism of this idea, see Foot (1985).

[115] I owe the quote to Dreier (2004, 145).

Comparative Satisficing
An act is right if and only if it is good enough, as measured by a *relative* standard.

Non-Comparative Satisficing
An act is right if and only if it is good enough, as measured by an *absolute* standard.

Comparative Satisficing says that whether an act counts as good enough depends on how it affects overall goodness, as compared to the best act that is available to the agent (cf. Slote 1984, 156).[116] Non-Comparative Satisficing interprets the term "good enough" in a non-comparative way. It fixes the level of goodness that an act has to bring about to count as right in absolute terms. Based on this distinction, there seem, then, to be two main alternatives to Maximization: Comparative Satisficing Consequentialism and Non-Comparative Satisficing Consequentialism. But are these really the only options open to consequentialists? We can find out only by analysing what a denial of Maximization entails. As we stated it above, it is the claim that an act is right if and only if it maximizes the good. Its logical complement is Non-Maximization.

Non-Maximization
It is not the case that an act is right if and only if it maximizes the good.

Non-Maximization negates a biconditional. It is, therefore, logically equivalent to the disjunction of the negation of the two conditionals that are implied by the biconditional. One disjunct says that there are wrong acts which maximize the good (maximizing acts, henceforth). The other purports that there are right acts which are not maximizing. A moral theorist who endorses Non-Maximization has, therefore, three options. She can subscribe to the first claim and deny the second and *vice verse*. Moreover, she can endorse both. As we shall see, all three options are logically compatible with what we called the Core Idea of consequentialism, viz. the notion that the rightness of an act depends only on the goodness that it produces.[117] That is, all three options are logical possibilities for consequentialists. This, however, does not mean that they all make intuitive sense. Let us consider each of these possibilities in turn.

Non-Maximization – Version 1 (Non-Maximization$_1$)
Some maximizing acts are wrong, but all right acts are maximizing.

This claim, it appears, contradicts the Core Idea of consequentialism flat out. According to the Core Idea of consequentialism, the rightness of an act depends only on its goodness. Therefore, two acts which are indistinguishable from the standpoint of goodness must be assigned the same moral status. Either they must all be right, or they must all be wrong. Since all maximizing acts are indistinguishable from the standpoint of goodness, they must all be either right or wrong. Supposing that right

[116] It would be possible to use a reference point other than the *best* act. In the following, we shall, however, disregard this possibility.

[117] These options do not only allow consequentialist doctrines. Each option is also compatible with non-consequentialist doctrines which invoke further normative factors in determining rightness. Concrete examples of such theories are given in footnotes 118, 124, and 126.

acts exist, the second part of Non-Maximization$_1$ entails, however, that at least some maximizing acts are right. Therefore, *all* maximizing acts are right. But the first half of Non-Maximization$_1$ says, in contradiction to this, that some maximizing acts are wrong. This argument, however, is unsound since the assumption that maximizing acts are indistinguishable from the standpoint of goodness is false. The fact that an act is, relatively speaking, the best option for acting available does not say anything, in and of itself, about its *absolute* goodness. At times, our best option for acting may be quite good. At others, it may be very bad. The latter is the case in tragic choices, where nothing we can do seems to be truly satisfactory. There should be moral doctrines, then, which are compatible both with the Core Idea of consequentialism and with Non-Maximization$_1$.[118] One candidate is a theory which endorses the following component.[119]

Maxificing
An act is right if and only if it (i) maximizes the good and (ii) is *good enough* by some absolute standard.

This doctrine determines the moral status of an act only based on its goodness. That is, it is compatible with the Core Idea of consequentialism. Furthermore, it judges that only maximizing acts are right because, according to clause (i), the maximization of the good is a necessary condition for rightness. And it allows that individual maximizing acts are wrong, viz. those which fail by the criterion stated in (ii), that is, in case they do produce enough goodness in absolute terms. So it is a variant of Non-Maximization$_1$ and lies, therefore, in one of three logical spaces that are allowed by Non-Maximization. This, of course, does not mean that it is well motivated. Let us examine it a bit more closely to determine whether it might make sense.

Firstly, it leaps to the eye is that the component we are looking at is a strange hybrid theory that contains elements of Maximization and Non-Comparative Satisificing. Insofar it seems appropriate to call it Maxificing.[120] It is a combination of Maximization and Satisficing whose possibility, I believe, has hitherto escaped the (conscious) attention of moral philosophers.[121]

[118] As pointed out in footnote 117, there are non-consequentialist doctrines in the domain of Non-Maximization$_1$, too. An example is the doctrine that an act is right if and only if it (i) maximizes the good and (ii) does not violate certain moral rules (e.g. the rule not to lie), where (i) and (ii) are lexically ordered. Such a moral theory is very close to being consequentialist. But it is nevertheless incompatible with the Core Idea of consequentialism, which says that the rightness of an act is determined *solely* based on its goodness.

[119] An alternative to this component is the doctrine that an act is right if and only if it (i) maximizes the good and (ii) is *below* a certain level of goodness (or exactly equal to this level). This doctrine, too, would satisfy Non-Maximization$_1$. But it is hopeless on the face of it. We can, hence, put it aside.

[120] Note that the term "maxificing" has previously been used, e.g. by Narveson (2004) and Dorsey (2005). The latter, in turn, attributes it to John Roemer, who has apparently used it first an unpublished paper manuscript. These authors, however, use "maxificing" in a completely different sense from the one I am intending. As Dorsey (2005, 578) explains, "'Maxificing' is a conglomeration of 'maximizing' those persons who 'satisfice'."

[121] As we shall see below, however, versions of Maxificing Consequentialism have been proposed, e.g., by Thomas Hurka. He seemed to have been unaware, though, that he was proposing a doctrine of this sort.

Another interesting feature of the doctrine is, secondly, that it opens up the possibility to make sense of moral dilemmas which is independent of the usual explanation that involves the idea of value incommensurability (cf., e.g., Richardson 1997, 115–117). Maxificing versions of consequentialism judge that all options for acting are wrong in situations where even the best act is not good enough by the standard specified in (ii).

Other than that, however, it is hard to see why one should find the above creed interesting, let alone endorse it as an alternative to Maximization. Like the latter, it regards the maximization of the good as a necessary condition for rightness. It inherits, hence, problems which, as we saw in the previous section, are commonly associated with maximizing versions of consequentialism, such as CU. These problems derive from the fact that the maximization of the good is a necessary condition for rightness. Maxificing also contains that condition and seems, hence, unmotivated. It does not solve the problems of Maximization and, perhaps, even adds new ones as it contains an additional clause (ii). We shall, therefore, leave it aside for the purpose of our investigation and move on to the second version of Non-Maximization.

Non-Maximization – Version 2 (Non-Maximization$_2$)
Some maximizing acts are wrong, and some right acts are not maximizing.

The second option open to consequentialists who reject Maximization is to say that an act's maximizing the good is neither a necessary nor a sufficient condition for its rightness. In other words, some maximizing acts are wrong, and some right ones are not maximizing. This, too, is compatible with the Core Idea of consequentialism. We can make this clear by way of an example. Consider Malevolent Consequentialism, which is the doctrine that an act is right if and only if it *minimizes* the good.[122] It judges the moral status of acts only by the goodness of their consequences and is, hence, in line with the Core Idea of consequentialism. And since it judges that all maximizing acts are wrong, and all right acts are not maximizing, it implies precisely what Non-Maximization$_2$ says, viz. that some maximizing acts are not right, and some right acts are not maximizing.

Malevolent Consequentialism is, of course, absurd on the face of it. There are, however, consequentialist doctrines in the domain of Non-Maximization$_2$ which are not, on the face of it, abstruse. One of these is Non-Comparative Satisficing Consequentialism. It maintains that an act is right if and only if it is good enough by some absolute standard.[123] It allows for the possibility that some maximizing acts are wrong. This will be the case in situations in which even the best the agent can do does not achieve the level of goodness required for rightness. And it entails that certain right acts are not maximizing, viz. in situations where certain

[122] Scarre (1996, 10) refers to this idea as "philosophical sadism."

[123] According to Slote (1984), an example of such a view is Karl Popper's "Negative Utilitarianism", hinted at in his *Open Society* (1947). Popper says "that there is, from the ethical point of view, no symmetry between suffering and happiness, or between pain and pleasure." (Popper 1947, 241) On this view, we do not have the duty to do the most good, but only a duty to avoid unnecessary suffering. For this reason, it seems, Slote interprets Poppers theory as a satisficing view.

non-maximizing acts are still good enough as judged by the absolute standard of goodness that is required.[124]

As Slote already pointed out in his seminal paper "Satisficing Consequentialism" (1984), the non-comparative variant of Satisficing Consequentialism is not very plausible, however. The problem is that, on Non-Comparative Satisficing Consequentialism, an act is judged as right only based on the goodness that it brings about and independently of what else would have been possible for the agent to do. Slote illustrates this, using Jeremy Bentham's doctrine as an example. He interprets Bentham's remark that his utilitarian moral principle "approves or disapproves of every action whatsoever, according to the tendency which it appears to have to augment or diminish the happiness of the party whose interest is in question" (Bentham 1838, 1) as a non-comparative satisficing view. According to it, thinks Slote, an act is right if it "alters [i.e. improves, NM] the balance of happiness over unhappiness even to the slightest extent." (Slote 1984, 154) The problem with Bentham's view is, then, that "if an alternative is available which would produce much more good, we should perhaps not normally feel that a slight addition to happiness was (morally) good enough." (Slote 1984, 154) An example in Slote's text makes this clear. If our criterion of rightness says that an act is right as long as it does some good on balance, then a medic on a battlefield acts permissibly if he merely hands out bandages although he could be treating the heavily wounded. For he does, after all, *some* good.

Proponents of Non-Comparative Satisficing Consequentialism can, of course, avoid this particular problem by raising the threshold. This will lead to another problem, however. There might be certain situations where the agent cannot help doing something which has bad consequences. In these situations, Non-Comparative Satisficing Consequentialism must "treat whatever the agent does as wrong." (Slote 1984, 155) In other words, it must judge that the agent faces a moral dilemma. But this is obviously abstruse. First of all, the existence of moral dilemmas is controversial. However, even those who believe that they exist do not go so far as to say that every choice situation in which all options are bad ones are moral dilemmas. In fact, some are most clearly not. Take a military commander who has a choice between executing plan *A* and plan *B*. Plan *A* will result in the death of 1000 troops. Plan *B* will have the consequence that the same 1000 troops die and, in addition to that, another 9000. In this situation, it is pretty clear what is right for our commander to do. He should execute plan *A*. It, too, will have terrible consequences. Nevertheless, it is the right thing to do – and quite evidently so. Non-Comparative Satisficing, however, will make the absurd judgement that plan *A* is wrong.

Since Non-Comparative Satisficing Consequentialism seems, hence, not to be an attractive option, Slote favours Comparative Satisficing Consequentialism. In

[124] A non-consequentialist version of Non-Maximization$_2$ is, e.g., the theory that an act is right if and only if it is not a lie. On such a view, every maximizing lie is wrong. And every non-maximizing act that is not a lie is right. There are, then, some maximizing acts that are wrong and some right acts are not maximizing.

contrast to the former, it elaborates the notion of "good enough" as "some sort of percentage or other mathematical function of the best results attainable by the agent." (Slote 1984, 156) Plainly, though, such a version of Satisficing Consequentialism would be incompatible with the second form of Non-Maximization that we are presently investigating. For it would deny that some maximizing acts are wrong. It is, however, possible, as Hurka (2004, 71) points out, to combine Comparative and Non-Comparative Satisficing Consequentialism into hybrid theories that we may refer to as Hybrid Satisficing Consequentialism (HSC).

Hybrid Satisficing Consequentialism – Version 1 (HSC_1)
An act is right if and only if it (i) is good enough, as compared to its alternatives and (ii) is good enough by some absolute standard.

This view, too, is compatible with Non-Maximization$_2$. It judges that certain maximizing acts are not right, viz. in situations where even the best act is not good enough by the absolute standard specified in (ii). Furthermore, it allows for right acts that are non-maximizing, viz. in situations in which some non-maximizing acts are good enough compared to the best act, as (i) demands, and relative to the absolute standard, as stated in (ii).

At least at first glance, this form of HSC_1 is more plausible than pure Non-Comparative Satisficing Consequentialism because it is more demanding when it matters. It would not permit the medic's act in Slote's example because the latter fails by the condition stated in (i). What the medic does is not good enough as compared to what he could have done. This makes the hybrid variant more attractive than Non-Comparative Satisficing Consequentialism. However, since the former contains an element of the latter, it inherits the other problem of its non-comparative cousin which we considered above. In the case of the military commander, it will judge that he faces a moral dilemma. We can prevent this, of course, by setting the bar for goodness in (ii) lower. This, it seems, will make the doctrine more acceptable. But it will also make it practically indistinguishable from Comparative Satisficing. That, in turn, will make it hard to understand why we should keep it in play as a separate option.

Maybe we missed something in our discussion of Non-Maximization$_1$ and Non-Maximization$_2$ that would rehabilitate certain possibilities in these domains. On the assumption, however, that we did not, we can rule out these options for the purpose of our investigation. No consequentialist doctrine which falls within their scope appears to be plausible. This leaves consequentialists with a third variant.

Non-Maximization – Version 3
No maximizing acts are wrong, but some right acts are not maximizing.

Let us consider, then, which doctrines fall into the domain of this component. It is compatible with some crazy views.[125] But we shall ignore them and look for serious contestants. The most obvious one of them is Comparative Satisficing,

[125] One of these views is, e.g., the idea that an act is right if it has either the best possible consequences or the worst possible consequences. This insane idea would fit, say, the mindset

which we introduced above.[126] It judges an act right if and only if it is good enough compared to the best possible option available to the agent. This doctrine entails Non-Maximization$_3$. On all plausible versions of this view, all maximizing acts are right.[127] And it allows for non-maximizing acts to be right.[128] At first glance, there is no need to rule this doctrine out, as it can plausibly address the worries associated with the other forms of Satisficing Consequentialism which we considered above.

Let us reconsider the worries that we had about Non-Comparative Satisficing. It seemed to be too permissive in one way and too harsh in its judgements in another. As for the first problem, we looked towards Bentham's non-comparative view and considered Slote's case of a medic who hands out bandages, while he could, instead, be treating the wounded. We observed that this acting might come out as right on a non-comparative view. As we saw, this problem can, of course, be addressed using the resources of Non-Comparative Satisficing Consequentialism. We can avoid it simply by putting the bar higher and raising the level of goodness that an act has to produce to count as right. This quick fix leads to another problem, however. In certain situations, the agent can, at best, do little good. Non-Comparative Satisficing Consequentialism will then judge that nothing the agent can do will be right. And this is equally counter-intuitive. Sometimes doing a little good is clearly right if the alternative is doing something that produces an atrocious result. It all comes down to this: The non-comparative variant of Satisficing will always have the one or the other problem. There is no way to set the bar for goodness just right because the question where it should be set depends on the situation. Hence, it seems, that any form of Satisficing Consequentialism can only be plausible if it contains some reference to the options available to the agent and their relative goodness. (Michael Slote already pointed this out in his first paper on the matter.) Now, this is precisely what Comparative Satisficing does. In the first case, it judges that what the medic does is wrong because it egregiously falls short of what he might have accomplished. In the second case, it judges the act of the commander who chooses plan *A* right. It is good enough as it is, indeed, the best he can do. So Comparative Satisficing can satisfactorily solve two cases which trouble its non-comparative counterpart.

---

of the Nazis, who wished that Germany should either rule the world (which they thought would be best) or perish. But it is not a serious contestant as a moral theory.

[126] A non-consequentialist doctrine that conforms to Non-Maximization$_3$ is the moral theory put forward by Samuel Scheffler. It combines, as Scheffler says, "the deeply plausible sounding feature that one may always do what would lead to the best available outcome overall" with an "agent-centred prerogative which has the effect of denying that one is always required to produce the best overall states of affairs." (Scheffler 1982/1994, 4–5)

[127] An implausible version of Comparative Satisficing that is incompatible with Non-Maximization$_3$ would hold that an act is right if and only if it produces, say, 101 % of the good that the best act produces. On such a view, all acts would be wrong.

[128] This holds, of course, only for forms of Comparative Satisficing Consequentialism which set the amount of goodness that counts as good enough below the level of goodness associated with the best act. This is a natural assumption, though. For doctrines which violate this condition are either equivalent to Maximizing Consequentialism or obviously absurd, as they judge all acts to be wrong.

Of course, there are various possibilities to flesh out the details of Comparative Satisficing and not all of them may be plausible. To come up with a plausible formulation of it, we will have to get the specifics exactly right. Obviously, the relative level of goodness at which an act counts as right has to be set at just the right level. If it is set too low, the doctrine will become too permissive. If it is set too high, it might face the same problems that we identified in our discussion of Maximization. It might impose unreasonable demands, restrict the autonomy of the agent too severely, and leave too little room for supererogation. Let us, however, put up these issues until we finally consider the case against consequentialism in the next chapter.

Before we move on, let us investigate further whether Non-Maximization$_3$ allows any other possibilities that deserve our attention. To this end, let us reconsider the version of HSC$_1$ that we considered and rejected above. It connects the elements of Comparative and Non-Comparative Satisficing Consequentialism *via* conjunction. Of course, this is not the only way to combine the two. Another version of the hybrid theory connects the two parts *via* disjunction.

Hybrid Satisficing Consequentialism – Version 2 (HSC$_2$)
An act is right if and only if it (i) is good enough as compared to its alternatives or (ii) is good enough by some absolute standard.

On this doctrine, no maximizing acts are wrong because they fulfil (i), which is a sufficient condition for rightness. Moreover, there can be right acts which are not maximizing because they fulfil (ii), which is also a sufficient condition for rightness.

HSC$_2$ seems, however, rather unattractive as well. If the absolute level of goodness required in (ii) is too low, it inherits the problems of the Benthamite stripe of Non-Comparative Satisficing Consequentialism. It will morally permit acts which do not seem to be morally permissible, such as the medic's handing out bandages instead of treating the heavily wounded. As we saw above, we can avoid this by setting the bar for goodness higher in (ii). The higher we set it, the more plausible the doctrine gets. However, as we do this, the more it will look like Comparative Satisficing Consequentialism. And this is an option we are already considering. Let us, therefore, dismiss HSC$_2$.

Thomas Hurka proposes another possibility that we cannot dismiss out of hand. He offers what he thinks is a form of Non-Comparative Satisficing Consequentialism, which he calls "absolute-level satisficing." (Hurka 2004, 71) Quite clearly, though, it is yet another form of Maxificing Consequentialism.[129] It involves a

[129]In one passage, Hurka says that in situations where no matter what the agent does her act will fall short of the absolute standard of goodness the "implications [of absolute-level satisficing, NM] coincide with those of maximizing." (Hurka 2004, 71) This means that, on Hurka's view, agents act rightly in such situations if they do the best they can do. For, according to Maximization, the act with the best consequences is always right. According to Non-Comparative Satisficing Consequentialism, however, even maximizing acts are wrong in such situations since they do not achieve the required absolute standard of goodness. Hence, it seems more plausible to regard Hurka's "absolute-level satisficing" as a hybrid between Maximization and Non-Comparative Satisficing rather than as a pure form of the latter.

differentiation between two types of cases, *A*-cases and *B*-cases. An *A*-case is one in which at least the best act alternative available achieves a given absolute standard of goodness. A *B*-case, in contrast, is one where no option for acting achieves that standard.

Maxificing - Hurka's Version
An act is right if and only if the following holds: (i) If the agent faces an *A*-case, she produces enough good as defined by an absolute standard. (ii) If she faces a *B*-case, she maximizes the good.

This creed, too, implies Non-Maximization$_3$. It says that all maximizing acts are right. This is the case in both *A*-cases and *B*-cases. And it allows for certain non-maximizing acts to be right, too, viz. in *A*-cases.

Hurka motivates his view rather plausibly. Take again the example of the medic on the battlefield. We have already established that it would be wrong for him just to hand out bandages. He should contribute more significantly to the good. Slote, however, claims that to fulfil his moral duties, the medic need not maximize. He need not go out and see who is injured the most and treat that person. It is good enough if he attends "to the first (sufficiently) badly wounded person he sees without considering whether there may be someone in even worse shape nearby." (Slote 1984, 153) And Slote goes on to claim that this seems, in fact, to be an agreeable judgement from the standpoint of common-sense morality. Hurka, however, rightly criticizes this reasoning by remarking that "[t]he reason we think the medic may treat the first sufficiently badly wounded person he sees (...) is strategic." What the medic does seems right because, if he went out and checked the whole battlefield for the most badly wounded soldier, he would just waste time and would certainly do less good than he could. (As we discussed in Sect. 4.2.1.2, treating the first sufficiently badly wounded person may, hence, be a sensible heuristic for maximizing the good). Hurka goes on to explain that for "a true test of comparative satisficing we must imagine that the medic comes to the battlefield, sees right off that one of two soldiers is more seriously wounded, yet attends to the other." (Hurka 1990, 110) This test shows, it seems, that Slote's Comparative Satisficing is not free from problems and that Hurka's Maxificing might eventually turn out to be more plausible. The latter may just treat the situation as a *B*-case and may, therefore, require that the agent maximize. The problem with Slote's version can be mitigated, of course, by raising requirements. A case could be made for the view that only acts whose consequences are almost maximally good should count as good enough. This, however, would make Comparative Satisficing look very much like Maximization. It would, hence, bring with it the problems of Maximization and would, therefore, make Slote's Comparative Satisficing appear unmotivated as an alternative component. At this point, I think, we need not adjudicate whether there is a solution to that problem. No harm is done, I believe, if we put both Slote's and Hurka's views on our list of alternatives to Maximization.

4.2.3 Alternative Welfarist Conceptions of the Good

In this section, we shall consider alternatives to the claims Summation, Universalism, and Equal Treatment. The reason we take them on in one go is that the combination of these claims shapes the character of a consequentialist moral doctrine in a *holistic* way. When we consider a particular consequentialist criterion of rightness, it is often not possible to tell which claims its account of the good contains precisely. It is multiply realizable by various combinations of fundamental components.

To see this, consider, e.g., Ethical Egoism. We can construe the doctrine as a close relative of CU whose criterion of rightness says that an act is morally right if and only if it maximizes the good, where the good is taken to be the agent's own good (cf. Shaver 2010).[130] How can this criterion be characterized? It is possible, first of all, to view it as a consequentialist doctrine which abandons Universalism and accepts a version of the view that we referred to as Partialism on page 100. We may conclude, then, that Ethical Egoism is a non-universalist consequentialist moral doctrine.

There are further possibilities, however. We can reconstruct the theory, starting from CU, by modifying another component of its axiology, viz. Summation. Summation gives us an operation which we can use to compute the good that attaches to act alternatives. Many such operations are possible. The family to which Summation belongs is, as we already know, *aggregative*. Doctrines in that family combine the well-being of different individuals into one value, trading off the good of one against the good of others. In contrast, other functions are *non-aggregative*. One example of a non-aggregative function is one which simply picks up the well-being of the agent and ignores everything else. A consequentialist who subscribes to all claims contained in CU, but swaps Summation for this non-aggregative function, is also an ethical egoist.[131] On this reconstruction, however, Ethical Egoism is not

[130]My example may seem to be ill-chosen. Hooker (2009a, 153), e.g., denies that it is a form of consequentialism. Sometimes it is even questioned whether Ethical Egoism is genuinely a moral doctrine. John Rawls, e.g., says that "that although egoism is logically consistent and in this sense not irrational, it is incompatible with what we intuitively regard as the moral point of view." And he says, furthermore, that it is significant "not as an alternative conception of right but as a challenge to any such conception." (Rawls 1971/1999, 117) On this point, see, also, Nida-Rümelin (2011b, 37). However, some authors, e.g. Chong (1992), have argued, contrary to Rawls, that Ethical Egoism can be seen as a moral doctrine after all. Others merely take it for granted that Ethical Egoism is "clearly a consequentialist doctrine." (Spielthenner 2005, 218) See, also, Dreier (1993, 22–23) and Portmore (1998, 2). What I say does not depend on this issue at all. All that counts here is that Ethical Egoism can be seen as a moral doctrine in the *formal* sense in which we defined this notion above.

[131]This reconstruction of Ethical Egoism is somewhat problematic, however, since it makes the doctrine appear somewhat incoherent. It holds "Universalistic Hedonism," which is the view that all sentient beings matter morally. This, however, is not reflected in the function that we then use to calculate the good.

a non-*universalist* version of consequentialism, but a non-*aggregative* one, since it does not allow any trade-offs between the welfare of all morally relevant individuals.

A third way of reconstructing the doctrine is to view it as a moral theory which accepts the claims contained in CU, but rejects CU's Equal Treatment in favour of the view we called "Unequal Treatment." Equal Treatment, recall, is the view that the good of all is to be weighted equally. Unequal Treatment is the idea that we should attach different weights to the well-being of the morally relevant individuals.[132] A particular version of this view is the notion that only the welfare of the agent should carry any weight, while everybody else's should be assigned a weight of zero. Combine this with the rest of CU and you get, once more, the doctrine that an act is right if and only if it maximizes the well-being of the agent.

The example of Ethical Egoism illustrates, then, a general point about consequentialist doctrines. Their criteria of rightness may be characterized in various ways because their theories of the good may be characterized in various ways. Hence, we should not discuss the individual claims that belong to their conceptions of the good in isolation. We should, rather, consider in a *holistic* way the various conceptions of the good that are open to consequentialists. To this end, we shall first look at some apparent problems about CU that may motivate consequentialists to look for alternatives to the classic utilitarian conception of the good. After that, we shall fathom the range of options that are open to them.

4.2.3.1 Motivation

Many philosophers have pointed out that CU is "supremely unconcerned with (. . .) interpersonal distribution."[133] (Sen 1973/1997, 16) This, one may say, is because it only looks at the *sum* of happiness that individuals share between them and not at how this sum is shared out. This aspect of the doctrine can lead to counter-intuitive implications in a number of cases. Let us look at some of the most important difficulties for CU which arise in this context. First up, let us look at a case that is introduced by Tim Scanlon.

Jones's Case
Suppose that Jones has suffered an accident in the transmitter room of a television station. Electrical equipment has fallen on his arm, and we cannot rescue him without turning off the transmitter for fifteen minutes. A World Cup match is in progress, watched by many people, and it will not be over for an hour. Jones's injury will not get any worse if we wait, but his

[132] It seems to me that Robert Shaver interprets Ethical Egoism in this third way. He contrasts it with Kantianism, Utilitarianism, and Common-Sense Morality, all of which require, as he says, "that an agent give *weight* to the interests of others." (Shaver 2010, emphasis added, NM)

[133] See, e.g., Brink (1989, 270–273), Feldman (1995), Feldman (1997, 151–174), McDermott (1982), Mendola (2006, 2–3), Rawls (1971/1999, 20-21), Roemer (1998, 130), Scanlon (1973, 1047), Scheffler (1982/1994, 11), Schroth (2008), Sidgwick (1907, 416–417), and Sinnott-Armstrong (2011).

> hand has been mashed and he is receiving extremely painful electrical shocks. Should we rescue him now or wait until the match is over? (Scanlon 1998, 235)

How does CU answer Scanlon's question? Presumably, since so many people are watching and enjoying the game, the sum of their individual pleasures outweighs the intense pain that is felt by Jones. Therefore, CU would probably demand that we do not rescue Jones, which, intuitively, is the wrong conclusion to draw. The mere fact that this would produce more good in the aggregate is not a sufficient reason, it seems, to let Jones suffer this much. If we did that, we would not take into account the fact that a *single* person has to bear the entire costs of our choice. We would not, as John Rawls has famously put it, "take seriously the distinction between persons." (Rawls 1971/1999, 24)[134]

Another problem for CU can arise when we compare and evaluate the moral quality of states of affairs which differ in population sizes. In such cases, CU may give rise to a "repugnant conclusion", as Parfit (1986, 381) famously suggested. The problem, in brief, is this. As we just noted, CU looks only at the sum total of happiness and not at how this total is distributed. Therefore, it does not distinguish a case in which one person has, say, 10 units of happiness from a case in which two persons have 5 units. From the perspective of CU, 100 persons might as well have 0.1 units of happiness (cf. Mukerji 2013c, 305). Morally speaking, it is all the same. The total in each case is, after all, identical. This has an unwelcome consequence. It is that CU may imply that the preferable state of affairs is one in which a 100 billion people "have lives that are barely worth living." (Parfit 1986, 388) The sum of happiness may, after all, be maximal in that state.[135]

---

[134] Rawls's originally intended his remark about the normative distinction between persons as a point against Smart (1973). Smart writes that "if it is rational for me to choose the pain of a visit to the dentist in order to prevent the pain of toothache, why is it not rational of me to choose a pain for Jones, similar to that of my visit to the dentist, if that is the only way in which I can prevent a pain, equal to that of my toothache, for Robinson?" (Smart 1973, 37) In Rawls' view, Smart violates the normative separateness of persons because he is willing to have Jones suffer the smaller pain to spare Robinson the greater one. Many moral philosophers subsequently made similar assertions (e.g. Williams 1985, 88; Nagel 1978, 142). Aside from Rawls, Nozick (1974, 33) may be seen as the *locus classicus* in this respect. By now, the claim that utilitarianism violates normative separateness seems a moral-philosophical commonplace. It should be pointed out, though, that this assertion has come under much disrepute and criticism. As Norcross (2009, 76) says, it is "often made, but rarely explained in any detail, much less argued for." Furthermore, it might have unwelcome consequences. It is not clear, e.g., whether the principle of normative separateness leads into a radical and counter-intuitive "number scepticism," as proposed by Taurek (1977). If that was indeed true, those who favour the principle could not explain why it is sometimes morally required to save the greatest possible number from harm. Scanlon (1998, 229-241) and Kamm (2007, 48-77), e.g., have argued that normative separateness does not imply this. For an instructive summary discussion, see Hirose (2007).

[135] It should be noted, though, that not all theorists find this conclusion unacceptable. Alastair Norcross, e.g., says that "I don't find Parfit's Repugnant Conclusion to be so repugnant" (Norcross 1998, 155). John Broome also finds the repugnant conclusion acceptable though he, too, thinks that it contradicts our intuition. However, he refuses to regard this as a serious problem and concludes "that [the] intuition (. . .) must be wrong." (Broome 2004, 212)

A further, related problem about CU is that it ignores the *fairness* of a distribution of well-being (cf., e.g., Rawls 1958, Sen 1973/1997, 16). Consider a situation in which I have 10 units of happiness and you have 10 units as well (and neither of us is particularly deserving or undeserving of their happiness). On CU, this situation is just as good as a situation in which, *ceteris paribus*, I have 20 units and you have nothing (cf. Mukerji 2013c, 305). Hence, on CU, an agent who faces a choice between bringing about the one distribution or the other acts rightly no matter what she does. This is because, according to CU, both options maximize the good. Intuitively, though, the distribution in which I am very happy, while you are very unhappy, is inequitable. It appears to be worse than the distribution in which we both have 10 units of happiness.[136]

The next problem echoes a point we made earlier. As we have learnt, CU is extremely demanding. According to CU, a moral agent may, e.g., be morally required to give away both her kidneys if, by doing that, she can save two people. We have already discussed one way for the consequentialist to solve this problem. She can reject Maximization and adopt a sufficiently lenient version of Satisficing. Many consequentialists, however, are reluctant to give up Maximization and may look for alternative ways to deal with this problem (e.g. Scheffler 1985; Portmore 2007 and 2011, Roberts 2002). There may be hope for them. As McElwee (2011) points out, the crass demands of CU have not only been connected to its maximizing nature, but also to its impartiality. We have already talked about this aspect of CU on page 105. As we pointed out, the impartiality of CU is a feature that can we can attribute to the ideas of Universalism and Equal Treatment. Maybe it is possible, then, to tinker with these components in order to find alternative routes for circumventing the demandingness problem? If there is, this may give consequentialists a further motivation to embrace those alternatives.

Another problem for CU is the fact that it is apparently incompatible with the idea of *special obligations*.[137] Intuitively,

> one ought to give preferential consideration to the interests of some persons as against others, including not only oneself but also other persons with whom one has special relationships, such as, for example, the members of one's own family or friendship circle or local community or nation or various other restricted social groups. (Gewirth 1988, 283)

This obligation is *special* insofar as it is rooted in a special *relationship* – one that we do not share with people at large. It is easy to see the intuitive force of this idea. Imagine, e.g., the following scenario. A friend of yours is in danger and needs your

[136] Some theorists might accept that, morally speaking, both distributions are equally good. They might still hold, however, that it is *ceteris paribus* impermissible for a moral agent to bring about the inequitable distribution because the goodness of an act's consequences is not the only thing that determines its rightness. Consequentialists, however, can only claim that it is impermissible to bring about the inequitable distribution if they argue that the equitable distribution is *better* in comparison. This is because they cannot invoke factors over and above consequences, as this would violate the Core Idea of consequentialism.

[137] An instructive discussion of the idea of special obligations is found in Jeske (2008).

help.[138] She is in a burning building, say, trapped together with other people. In this situation, should you attach the same weight to her well-being as to everyone else's? It seems that you should not. If you could save only one person in that building, you ought, most certainly, to save your friend. To be sure, in doing that you would be partial. You would abandon the principle of Equal Treatment. However, it seems that this would be sanctioned – nay, *required* – by morality. CU, however, disagrees. It demands that you ask whose life has the greatest value from an impartial point of view where everyone counts the same. To be sure, the classic utilitarian calculation, too, might yield the conclusion that you ought to save your friend. But this would be a matter of coincidence. It is quite likely that CU will demand that you save someone else, as it is oblivious to the idea of special obligations.

It appears, then, that special obligations can ground a morally mandated preferential treatment of some individuals as against others. However, special obligations are not the only such factor. Many authors have mentioned desert or merit, too. John Stuart Mill, e.g., says that

> it is universally considered just that each person should obtain that (whether good or evil) which he *deserves*; and unjust that he should obtain a good, or be made to undergo an evil, which he does not deserve. (Mill 1863, 66; emphasis in the original, NM)[139]

A further factor that appears to be relevant in this context is *historical injustice*. If, in the past, a person who deserved a given level of well-being, but got less than that, it seems as though, from the moral point of view, benefits to her are comparatively more valuable than benefits to other people. On CU, though, we cannot take into account desert or historical injustice. We have to consider everyone's well-being impartially and sum it up. There is no room for anything else in CU. This point, too, may motivate consequentialists to endorse an alternative to CU's conception of the good, viz. one that does not treat everybody equally.

Before we consider some options that allow consequentialists to avoid these complications, let me address a further problem for CU that moral philosophers commonly ascribe to its aggregative nature. CU seems to make very high informational demands. What does this mean? Above we imagined two states of affairs. In the first, we both have 10 units of happiness. In the second, I have 20 units and you have 0 units. In saying this, I supposed that my happiness and your happiness can be measured and compared on a common scale. I supposed, that is, that they are *commensurable*. This, of course, is a problematic assumption. However, due to its conception of the good, CU relies on it. After all, how else could we make sense of a happiness *sum*? To calculate a happiness sum, it is obviously necessary to compare the happiness of any two persons with one another (cf., e.g., Hirose 2011, 72). As

[138] I have used this example in Mukerji (2013b) and Mukerji (2013c). It is a freely adapted version of the classic "famous fire cause" (discussed, e.g., by Barry 1995, 222–233 and Williams 1981). The original case is due to Godwin (1793, 82–83).

[139] Many authors on both sides of the debate about consequentialism agree on this. For an example of a paradigmatic deontologist, see Ross (1930/2002, 137–138). For a typical consequentialist, see Sidgwick (1907, 279).

some believe, this is impossible since happiness is a subjective mental state which is epistemically accessible only to the individual who has it.[140] Economists have been aware of this problem for a long time. Stanley Jevons, e.g., already asserted back in 1871 that he sees no way "to compare the amount of feeling in one mind with that in another" since the "susceptibility of one mind may, for what we know, be a thousand times greater than that of another." And he concluded from this that "[e]very mind is thus inscrutable to every other mind, and no common denominator of feeling is possible." (Jevons 1871, 21)[141] Many economists should subsequently come to share some version of Jevons's view (cf., e.g., Robbins 1932/2007 and 1938; Samuelson 1947).[142] In philosophy, too, the view that interpersonal comparisons of utility are problematic has gathered adherents, particularly amongst philosophers with inclinations towards logical positivism. As Brad Hooker points out, the charge that CU relies on unrealistic assumptions about the possibility of interpersonal comparisons has become one of the most common objections to the doctrine (cf. Hooker 1990, 68).[143] Nevertheless, we will disregard this problem of interpersonal comparison and make the (charitable) assumption that consequentialists can solve it.

4.2.3.2 Options

What are the consequentialist's options as regards the above complications? Let us start by reconsidering *Jones's Case*, where Jones's hand is being mashed up by the electrical equipment in the transmitter room of the television station. To save him, we have to turn off the broadcast of the soccer game and many viewers will miss out on the enjoyment of watching the game. In the aggregate, this enjoyment supposedly outweighs Jones's intense pain. It follows, then, from CU that we ought not to save Jones, which, intuitively, is the morally wrong choice. It seems right to save Jones.[144]

How could a consequentialist explain this judgement? She has to argue, it seems, that the alternative of saving Jones is, in fact, *better* than the alternative of not saving

[140] Discussions of the problem of interpersonal utility comparisons can be found, e.g., in Bergström (1982), Broome (1999, 29-32, 37-43, 207-208), Griffin (1986, 113–120), Hare (1981, 117–129), Harsanyi (1955, 317–320), Harsanyi (1977a, 638–642), Little (1950, 53–68), Ng (2004, 16–18), Scarre (1996, 16-18, 136-137), and Sen (1973/1997, 12–17; 1997b, 2010, 277–284).

[141] Interestingly, the founding farther of utilitarianism, Jeremy Bentham, was already aware of the problem of interpersonal comparison. In fact, in a rarely quoted, undated manuscript passage he went so far as to say that instead of comparing the utilities of different people "you might as well pretend to add 20 apples to 20 pears." (Dinwiddy and Twining 2004, 49)

[142] In the case of Robbins, the influence of logical positivism seemed to play a role (cf. Walsh 1996, 179–181; Putnam 2002, 53).

[143] Hardin (1988, xv) makes this point as well.

[144] Some moral theorists have challenged this idea. Alastair Norcross, for one, argues that a great number of small benefits to many (e.g. avoiding a headache) can outweigh great harms (e.g. death) to a few. For his argument, see Norcross (1998).

Jones. An argument to this effect may take various routes. The first option is to look for alternatives to the component Summation. Maximin is one such alternative.[145]

Maximin
For all acts, a_i and a_j, that are available to an agent, the following holds: a_i is better than a_j if and only if the well-being of the worst-off individual that is associated with a_i is greater than that which is associated with a_j and equally good if and only if it is equal.[146]

A consequentialist who departs from Summation and adopts Maximin can give a satisfactory explanation for our judgement in Jones's case. Maximin evaluates states of affairs based on the situation of the worst-off individual(s) in these states. In the present case, there are two options to consider. Either we save Jones, or we choose not to. If we do not save Jones, he is presumably the worst-off individual.[147] Hence, his well-being level is what determines the goodness of the state of affairs that results from this act. If we do not save Jones, the viewers of the soccer game are, let us assume, worse off than he is. They are disappointed because they cannot watch the game. To rank our options regarding their goodness according to Maximin, we have to compare, then, the intense pain that Jones suffers if we do not save him with the disappointment of the many viewers that would result from saving him. Above I mentioned that such comparisons throughout persons are problematic. Unless we take a very sceptical stance here, the outcome of the comparison should, however, be rather obvious. Surely, it is not nice to miss out on the game. But, intuitively, it beats Jones's agony.

According to Maximin, saving Jones has the better result, then. Now, this makes it possible for a consequentialist to explain our intuitive judgement about the case. We think it is morally mandated to save Jones. And this, conveniently, is what follows from a consequentialist doctrine which replaces Summation with Maximin.

Prima facie, Maximin is then an attractive alternative to Summation for consequentialists. I propose, nevertheless, to dismiss it as an option since it has highly problematic implications in other cases. Consider, e.g., a scenario in which Jones and Smith both had an accident in the transmitter room. Electric equipment has fallen on their arms. It gives them painful electric shocks and will continue to do so for the next 15 min unless they are rescued. This time, no game is being broadcast. So we can save Smith and Jones without disappointing any viewers. But, tragically, in this case, we cannot save them both from their agony. We can either save the one and let the other suffer. Alternatively, we can do the reverse. Or we could do

[145] We interpret Maximin as an axiological thesis. Note, however, that some authors take it to be a deontic claim. See, e.g., Shrader-Frechette (1991, 102).

[146] Maximin is commonly associated with John Rawls who first proposed it in a moral-philosophical context (cf. Rawls 1971/1999). (Rawls, however, does not view it as a moral principle, but as a principle of choice under uncertainty which is rationally justified under very specific conditions. He assumes that the contracting parties in his contractualist thought experiment will use it to choose a conception of justice.) For a consequentialist moral conception which is based on Maximin, see Mendola (2005b and 2006).

[147] If this assumption is false, the Maximin solution does not work anymore. But we shall not deal with this case at this point.

nothing at all and let them both suffer. The latter option, of course, sounds like a bad joke. It is clear that we ought to save at least one person in this case. However, on Maximin, saving one is, in fact, as good as saving none. The situation of the worst-off individual(s) is the same in all three states and, hence, Maximin does not prefer one over the others. It is easy, however, to fix this problem using a principle which is very much akin to Maximin. It is Leximin.[148] Leximin evaluates states of affairs *vis-a-vis* each other just like Maximin does, except that, whenever Maximin is indifferent between two states, Leximin determines their rank-order based on the *second*-worst-off individuals in these states. If their well-being is the same, then Leximin ranks the states based on the *third*-worst-off individual and so on.

Leximin
For all acts, a_i and a_j, that are available to an agent, the following holds: a_i is better than a_j if and only if the well-being of the worst-off individual that is associated with a_i is greater than that which is associated with a_j. If a_i and a_j tie in this comparison, their ranking is determined by the well-being of the second-worst-off individual that is associated with them, respectively, and so on. If a_i and a_j are equal with respect to the well-being of the worst-off, second-worst-off, ... and best-off, then a_i and a_j are equally good.

Leximin apparently gives the same answer in *Jones's Case* as Maximin and leads, hence, to an intuitively satisfactory result. But it also gives a satisfactory answer in the latter case of Jones and Smith. Here, recall, Maximin failed. The three options for acting – viz. to save Jones, to save Smith, and to Save none – tie in regards to the well-being level of the least well-off individuals. However, only the act of rescuing Jones and the act of saving Smith tie when we compare them regarding the well-being of the second-worst-off individual. The option to help neither Jones nor Smith is dominated and, hence, ruled out as a morally justifiable option.[149]

Vallentyne (2006) puts forward a further option which can solve the complication that CU runs into in Jones's case. He suggests that consequentialists may choose a more or less "coarse-grained" theory of the good. With the aid of such a theory, a consequentialist can explain why both the act of saving Jones and that of helping Smith lead to a morally superior outcome.[150]

The coarse-grained variation of CU, Coarse-Grained Utilitarianism (CGU), may look something like this. In contrast to CU, CGU does not just add up the precise amounts of well-being that an individual enjoys and judges overall goodness based on the sum. Rather, before it does the summing, it puts the various welfare values into rough *brackets* each of which carries a value itself. After that, it does the sums.

[148] Amartya Sen suggested Leximin in response to Rawls's proposal of Maximin. See, e.g., Sen (1970a, 138).

[149] Note, however, that this is certain only on a maximizing consequentialist moral theory. The option to save none may still be justifiable on a satisficing doctrine.

[150] It shall be noted, though, that my point here is different from Vallentyne's. His proposal is not meant to solve the specific problem that CU faces in Jones's case. Vallentyne sees it, rather, as an attempt to establish that even a maximizing consequentialist moral theory can be compatible with moral freedom. If such a conception has a coarse-grained theory of value, he says, "[t]ypically, there will be many actions that maximize value."(Vallentyne 2006, 26)

It is easy to see how this procedure can solve the consequentialist's problem in *Jones's Case*. Suppose there are 100 million people who would like to watch the game. Their well-being is at 10 hedonic units, as is Jones's. If the viewers do get to watch the game, their welfare increases to 11 units and Jones's drops to 1. Suppose, further, that the coarse-grained theory maps every individual's well-being onto a three-unit bracket: [0,1,2], [3,4,5], [6,7,8], [9,10,11], and so on. These brackets are, in turn, assigned a value which, let us assume, is their respective median value: 1, 4, 7, 10, and so on, respectively. Now, if we choose the option to save Jones, the viewers stay in the same bracket, and so does Jones. If, however, we do not save Jones, his well-being drops into bracket [0,1,2] which has a value of 1, while the viewers stay in the same bracket. Hence, on CGU, the overall goodness of saving Jones is nine units better than the option not to save him. CGU can explain, then, why it is morally required to save Jones.[151]

Let us move on, then, to the second problem, viz. the "repugnant conclusion." As we saw in the previous section, CU may prefer a state in which many people's lives are barely worth living to a state in which few people live very good lives. CU may, therefore, imply that we should work towards the former rather than the latter state, which may be seen as counter-intuitive. Leximin can fix this problem, too. According to Leximin, we determine the relative goodness of two states by looking at the worst-off individual(s) in those states. In the given case, then, the relevant comparison is between the well-being of the great number of fairly miserable individuals in the one state and the small number of fairly well-off individuals in the other state. In this comparison, the state with the happier individuals carries the day, and the repugnant conclusion is averted.

An appropriately specified version of CGU can solve this problem, too. We may stipulate that the value of well-being of the comparatively miserable people in the one state lies in a bracket whose value is zero. We may suppose, furthermore, that the value of well-being of the comparatively comfortable individuals in the other state lies in a bracket whose value is positive. If we do that, the state that contains the small number of fairly well-off individuals comes out as better which also solves the problem.

There is also a third way to address the repugnant conclusion. Consequentialists can adopt, in place of Summation, the following principle.[152]

Average
For all acts, a_i and a_j, that are available to an agent the following holds: a_i is at least as good as a_j if and only if the average of the appropriately weighted well-being of all morally relevant individuals that is associated with a_i is at least as great as the average of

[151] Again, on a satisficing variant of CGU, this is not necessarily so. The option not to save Jones may lead to a morally good enough outcome and may also be morally permissible. Also, the solution only works on this particular construal of the case.

[152] As noted by Jamieson (1984, 210), this principle has the same moral implications as Summation if the states of affairs associated with the act alternatives in a given choice situation do not differ in population size. For this reason, Average cannot solve the problem in *Jones's Case*. It implies the same counter-intuitive judgement as Summation.

the appropriately weighted well-being of all morally relevant individuals that is associated with a_j.[153]

Obviously, the average of happiness is larger amongst the small number of people than amongst the big number. Hence, if a consequentialist drops Summation and adopts Average as a criterion for ordering alternatives regarding goodness, she can also avert the repugnant conclusion.[154]

Let us turn, then, to the third problem which is the issue of fairness. As we said above, CU cares only about the sum total of happiness and not about its distribution. On CU, a distribution where you and I have 10 units of happiness is as good as an alternative allocation in which I have 20 hedonic units, and you have 0. The latter seems unfair, however, given that none of us is more deserving of happiness than the other.

Apparently, the problem about CU is that it attaches no intrinsic value to equality (cf. Holtug and Lippert-Rasmussen 2007, 2).[155] A consequentialist may want to adopt, then, a ranking method which is sensitive to it. The most prominent proposal in this respect is what might be called Pure Egalitarianism.[156]

Pure Egalitarianism
An act, a_i, is better than another, a_j, if and only if a_i leads to an equal distribution of well-being, while a_j does not. Otherwise, it is equally good.

Pure Egalitarianism, one may say, simply partitions the alternatives that are available to an agent into two classes. Class 1 comprises all options that promote equal distributions of happiness. Class 2 contains all choices that lead to unequal

[153] A principle of average utility is proposed, e.g., by Harsanyi (1977b, 28).

[154] There is, however, a worry about Average that has often been pointed out. Average well-being increases, obviously, if we kill off those individuals whose well-being is below the current average well-being. As far as I know, this was first noted by Henson (1971). The same point has subsequently been made by other authors, too (e.g. Jamieson 1984 and Sinnott-Armstrong 2011). In what follows, we shall ignore this problem. I should, however, remark that it may not only affect Average. Rather, it may be a more general point that also affects, e.g., Maximin and Leximin. If I kill the worst-off individual in any given state of affairs, the resulting state will, *ceteris paribus*, be better, according to both Maximin and Leximin, than the prior state since the worst-off individual in the resulting state is then better off than in the prior state. This conclusion depends, of course, on metaphysical suppositions that may be questioned. But these seem to be identical to the assumptions that inform Henson's point regarding Average.

[155] It has been pointed out, however, by classic utilitarians that CU tends to favour distributions of material goods that are rather equal. The main reason for this is the "law of diminishing returns" which holds true for most goods and people. It purports, technically speaking, that the relationship between the goods held by an individual and her welfare is a function that increases at a diminishing rate. That is, each unit of a given good increases welfare by a lesser amount than the unit before. On this point see, e.g., Smart (1973, 36), Hare (1981, 164–165), and Broome (1991, 175–177).

[156] I take the term "Pure Egalitarianism" from Parfit (2003, 117). It is not to be confused with what has been called "Radical Egalitarianism." (Nielsen 1978 and 1981) The latter is not an axiological ranking method but a view of justice, to wit, that "[j]ustice in society as a whole ought to be understood as a complete equality of the overall level of benefits and burdens of each member of that society." (Nielsen 1981, 121)

distributions. Alternatives are then ranked *vis-a-vis* each other based on their membership in the one or the other class. All elements of class 1 are equally good as are all elements of class 2. Also, all elements of class 1 are ranked above all elements of class 2.

This ranking method can solve the problematic exemplary case that I raised above. The distribution in which you and I both have 10 hedonic units is in class 1, the one in which I have 20, and you have 0 is in class 2. Hence, on the pure egalitarian principle, the equal distribution is judged to be better, which is intuitively correct.

It is easy to see, however, that Pure Egalitarianism is too crude. One of its problems is that it does not distinguish between the various equal arrangements (cf. Holtug 2003, 161).[157] The distribution (10,10) is taken to be as good as the distribution (11,11). The latter, however, is *obviously* more desirable. A simple way to fix this problem is to build in a tie-breaker and to say that amongst distributions in class 1 those with the larger total are better.[158] Then, (11,11) comes out as better than (10,10). This criterion is still problematic, though, since every unequal distribution gets ranked below every equal distribution. As John Rawls taught us, however, certain unequal distributions may be morally more desirable than even the best equal distribution.[159] It seems very reasonable to suppose that an uneven distribution is better than an even distribution if every individual – and, in particular, the least well-off – is better off under the former than under the latter. This, roughly, is the axiological idea behind John Rawls's "difference principle," whose lexical extension is structurally identical to Leximin (cf. Rawls 1971/1999, 72).[160] Leximin might, therefore, be an attractive option for the consequentialist. With it, she could meet not only the problem presented by Jones's case and by the repugnant conclusion. She could address the fairness challenge as well.

Some theorists, however, insist that Leximin, too, is problematic. It can be criticized from two completely different angles.[161] Egalitarians may criticize it

[157] Another problem is that it does not distinguish the relative goodness of various unequal distributions. This issue was taken up by economists and philosophers who tried to put egalitarian thought on a more solid foundation. Ground-breaking work was done by Atkinson (1970) and by Sen and Foster (1973). An instructive summary discussion is given by Cowell (2000).

[158] Egalitarians may justify this, e.g., in reference to what Bertil Tungodden has called "Weak Utilitarianism" which is the principle that "[i]f one alternative has more total utility than another, it is better in one respect." (Tungodden 2003, 17)

[159] For a rather informal elaboration of this point see, e.g., Mukerji (2009, 30–33). The relevant passage by Rawls can be found in Rawls (1971/1999, 65-70).

[160] Note, however, that Leximin cannot be equated with the difference principle. There are various differences. For one thing, the difference principle is a deontic principle, whereas Leximin is an axiological principle. Also, the difference principle is intended as a design principle for social institutions, while Leximin, as we have introduced it here, is a general method for ranking act alternatives in terms of their goodness.

[161] This is, of course, not to say that these are the only kinds of critique. Consider, e.g., Nozick's famous critique of the difference principle. He focuses on the fact that this principle is, as he says, an "end-state principle" (cf. Nozick 1974, 160–164).

because it seems not, in fact, to be an egalitarian ranking principle at all. Indeed, it is not concerned with overall equality *per se*, but only with the situation of the worst-off.[162] It may prefer a comparatively inegalitarian distribution to a relatively egalitarian distribution if the former is preferable from the perspective of the worst-off. Also, it approves of any increase in inequality "when this would benefit the best-off group without harming the worst off." (McKerlie 1994, 31)

The second critique of Leximin is that in giving absolute priority to the worst-off, it treats the second-worst-off unfairly. Benefits to them, however great, can never morally outweigh losses, however little, to the worst-off (cf. Roemer 1998, 137). Those who find the egalitarian critique more pressing may opt for Moderate Egalitarianism as a ranking principle. In determining the goodness of a distribution, it considers both its sum and its equity.

Moderate Egalitarianism
An act, a_i, is better than another, a_j, if and only if a_i leads to a distribution of well-being that is, on the whole, more preferable in terms of its aggregate welfare *and* equity.

Unlike a proponent of Leximin, a moderate egalitarian can prefer a distribution that is comparatively worse for the worst-off *vis-a-vis* a different available distribution if it offers a better mix of equality and aggregate welfare.

Those who find the second critique more urgent may opt for Prioritarianism, which is a version of Unequal Treatment.[163]

Prioritarianism
The well-being of all morally relevant individuals is to be weighted inversely proportionally to the amount they enjoy.

Prioritarianism can be combined, e.g., with Summation. This combination yields the view that a distribution of benefits is better than another if the weighted sum of benefits is greater. Since benefits to the less well-off are given increased weight, this view will tend to prefer acts that promote more equitable distributions of welfare. Unlike a pure egalitarian view and Maximin, it will not, however, completely ignore the second-worst-off individuals. Benefits to them are also taken to be quite important. Hence, one may argue, they are not treated unfairly.[164]

A further component which may help the consequentialist to address the problem of fairness is Multiplication.

[162] Once more, this raises the issue of inequality measurement that I have alluded to in footnote 157. As above, I shall leave this issue aside.

[163] The first statement of Prioritanism was given by Parfit (2003).

[164] Note that our use of the term "Prioritarianism" is rather idiosyncratic. Other authors use it in a more inclusive way, such that it encompasses Leximin. Richard Arneson, e.g., writes that "[a]t one extreme, prioritarianism gives absolute priority to the worse off, so that securing any gain of any size however tiny for a worse-off person is morally to be preferred to securing a gain however huge for any number of persons who are better off. Prioritarianism here becomes leximin." (Arneson 2000, 58) I am grateful to Martin Rechenauer for clarification regarding this matter.

Multiplication
Overall well-being is the product of the appropriately weighted levels of well-being of all morally relevant individuals.

It, too, has a tendency towards equality. Consider again the problem of distributing 20 units of a good between you and me. What is the best distribution? Obviously, it is the equal distribution 10/10 since the product of your and my share is maximized when you and I both get ten units.[165]

Let us turn, then, to the issue of demandingness which was the fourth problem we considered in the previous section. As we already discussed above, some theorists have suggested that consequentialists should address the demandingness problem by way of a departure from Maximization (e.g. Slote 1984). However, there may, conceivably, be another way. To soften the extreme moral demands that confront the classic utilitarian agent, consequentialists may modify the components Universalism and/or Equal Treatment.

Let us consider Universalism first. It says that all individuals capable of well-being deserve moral consideration. Obviously, if their class were more limited, the resulting doctrine would be less demanding than CU. We can see that this is so by considering the limiting case, to wit, a consequentialist view that regards only the agent herself as morally relevant. As we pointed out above, this is a form of Ethical Egoism. The distinctive feature of this doctrine is that it does not impose on the agent any self-sacrificial moral demands at all.[166] It merely requires that she maximize her own good. Now, the consequentialist need not go so far as to become an egoist. In fact, most consequentialists will try to eschew egoism since it is considered a rather

[165]It can easily be shown that any distribution of one unit of a valuable good from a better-off person to a worse-off person increases the product of units that is shared between the two, unless the two swap places due to the redistribution. To see this, consider a good, g, which is split up into two shares, g_1 and g_2, which are given to individuals 1 and 2, respectively. Let 1 be better off by at least two shares, i.e. $g_1 - 1 > g_2$. Then, it will always be possible to increase the product g_1g_2 by taking one unit from 1 and giving it to 2. That is, $(g_1 - 1)(g_2 + 1) > g_1g_2$. Here is the proof. By assumption: $g_1 - 1 > g_2$

$$g_1g_2 + g_1 - 1 > g_1g_2 + g_2$$
$$g_1g_2 + g_1 - g_2 - 1 > g_1g_2$$
$$(g_1 - 1)(g_2 + 1) > g_1g_2 \quad q.e.d.$$

The result can easily be generalized to any number of individuals n. Consider a distribution $g_1, \ldots, g_n$ of g throughout n individuals. It is clear that any redistribution of one unit of g from a better-off individual, i, to a worse-off individual, j, will increase the product of individual goods holdings on the assumption that i and j would not swap positions due to the redistribution. Formally speaking, if $g_i - 1 > g_j$, then $g_1g_2 \ldots (g_i - 1) \ldots (g_j + 1) \ldots g_n > g_1g_2 \ldots g_n$. As we have already shown, the assumption that $g_i - 1 > g_j$ implies that $(g_i\text{-}1)(g_j + 1) > g_ig_j$. Now we simply multiply both sides of this inequality by $g_1 \ldots g_{i-1} \ldots g_{i+1} \ldots g_{j-1} \ldots g_{j+1} \ldots g_n$ to show the result.

[166]The emphasis lies on *self-sacrificial*. We are not claiming that Ethical Egoism does not impose any moral demands on the agent. In fact, some theorists have argued that, contrary to popular opinion, Ethical Egoism can be quite a demanding moral doctrine. See e.g. Person (1985).

implausible view (cf., e.g., Kagan 1998, 63). There are presumably many options that lie on a spectrum which ranges from an impartial regard for all to absolute selfishness. There should then be an attractive option somewhere on this spectrum.

However, it seems that this is not the case. There appears to be no version of non-universalist consequentialism that is even remotely plausible. Here is why. A consequentialist who gives up Universalism kicks some individuals out of the domain of morally relevant subjects. This means that what happens to them it is morally *irrelevant*. It is utterly implausible to suggest this. It makes much more sense to say that all sentient beings deserve to be given some regard though some of them may have a higher moral status than others.[167] In other words, consequentialists are, it seems, well advised to stick to Universalism and to adopt some form of weighting method in place of Equal Treatment. A philosophical view which comes to mind in this connection is C. D. Broad's conception of "self-referential altruism."[168] (Broad 1971, 279–280) It directs the agent to exhibit a universal concern for all while allowing her to attach variable weights to the good of different individuals. On a view such as this, it is possible to scale up the moral importance of the agent's good, which, of course, will reduce any moral concessions that she is required to make. To sum up, then, it seems as though the demandingness objection can be addressed both by a departure from Universalism and by a renunciation of Equal Treatment. Only the latter, however, appears to yield a plausible solution.

The sixth problem of CU that we considered above is the fact that the doctrine does not allow for special obligations. We said that most of us believe that we owe a greater concern to those with whom we share some sort of special relationship (e.g. a family or community membership, a friendship, and so on). It seems, therefore, to be a *prima facie* desirable feature of a moral theory that it permits us to care more for certain people than for others and to act accordingly. It is not difficult to see that a consequentialist moral conception much like C. D. Broad's self-referential altruism can tackle this problem, too. As we saw, self-referential altruism requires that moral agents take everyone's welfare into consideration. However, it allows them to accord different weights to different individuals. They may be allowed to give greater weight to their own concerns. And they may also attach increased importance to those near and dear to them and to other people with whom they share a moral bond. The issue of special obligations may, then, also motivate consequentialist theories that adopt some version of Unequal Treatment.

The same can be said when it comes to the issue of moral desert. Above we stated that it seems plausible to assume that people should be treated according to their moral merits and that preferential treatment should be given to those who

[167] This gradualist view is embraced, e.g., by John Rawls in his discussion of the demands of justice as regards animals. There he says that animals do not hold the same equal basic rights as do humans, but hastens to add that "they have *some protection* certainly." (Rawls 1971/1999, 442; emphasis added, NM)

[168] Gewirth (1988) refers to this idea as "ethical particularism."

had to suffer historical injustice. As it turned out, CU seems not to allow this, which, intuitively, appears to be wrong. But the solution that we just considered in connection with the problems of demandingness and special obligations should also provide a remedy to this problem of CU. Consequentialists, it seems, can throw out CU's Equal Treatment in favour of a version of Unequal Treatment that allows them to accommodate concerns for moral desert. To this end, they simply need to attach increased moral weight to the welfare of those who have, e.g., furthered morality's cause in the past and to those who have suffered historical injustice.

Moral Desert
The well-being of all morally relevant individuals is to be weighted according to their moral desert.

Endorsing this principle would, it seems, help consequentialists to bring their doctrines further in line with our intuitions.[169] Some moral theorists, however, may not be entirely convinced by this solution. Their reservations may be based on something like the following argument. Consequentialist theories, they might say, are essentially future-regarding and cannot take past-regarding factors into account. Moral Desert, however, is a past-regarding principle and is, hence, incompatible with a consequentialist theory. In a more structured form, we can state the argument as follows:

(P1) Consequentialist theories are incompatible with past-regarding factors.
(P2) A moral theory that endorses Moral Desert involves past-regarding factors.
(C) Consequentialist theories are incompatible with Moral Desert.

Both premises seem credible. P1 appears to follow from the Core Idea of consequentialism in conjunction with a weak metaphysical assumption.[170] According to the Core Idea of consequentialism, the moral status of an act is determined entirely by the goodness that it promotes. The goodness of the past seems to be fixed because the past itself seems to be fixed. Nothing we do can change it.[171] Therefore, the

[169]The problem of adjusting utility for desert has received an influential discussion by Rescher (1966). In recent times, Feldman (1995) has put forward a detailed elaboration of this approach.

[170]Note that it also sits well with our initial characterization of consequentialism that we gave on page 15. We distinguished between three types of normative factors, viz. past-regarding, intrinsic, and future-regarding factors. Intuitively, we said, consequentialism should be regarded as the class of moral theories that accepts only future-regarding factors, viz. consequences. And this is, in fact, a view that is widely shared (e.g. Graafland 2007; Oddie and Milne 1991; Carlson 1995; Nida-Rümelin 1997b; Geirsson and Holmgren 2000). Now, we have already established that the distinction between the act itself and its consequences seems very fuzzy because there are usually many adequate descriptions of what happens and each of these descriptions draws the line between the act and its consequences in a different place. It may be questioned, therefore, whether the distinction between intrinsic and future-regarding factors can, indeed, be maintained. Perhaps, then, a consequentialist theory can accommodate factors that moral philosophers have traditionally seen as intrinsic. But in any case, it seems, we can agree that consequentialist doctrines cannot take past-regarding factors into consideration.

[171]As it has been noted by Geach (1969), however, every event may be said to cause a *trivial* change about the past – a so-called "Cambridge change." When I scratch my head now, this does not only

moral status of an act can only depend on its present and future effects. Past-regarding factors cannot play a role. P2 seems to be unobjectionable as well. A moral theory which endorses the principle of Moral Desert appears to make room for past-regarding factors. It allows increased moral weight to be attached to the well-being of those who *furthered* morality's cause in the past or to individuals who *suffered* historical injustice. Both premises of the above argument appear, then, to be well supported. It seems, therefore, to follow that consequentialists cannot endorse Moral Desert.

In fact, however, consequentialists can reject this argument. They can do that because it seems to be guilty of an equivocation (cf. Mukerji 2013a). There are, in fact, two senses of the word "past-regarding" that we should keep part. These are as follows.

Past-Regardingness 1
A normative factor is past-regarding$_1$ if and only if it is a fact about the past.

Past-Regardingness 2
A normative factor is past-regarding$_2$ if and only if it is a fact about the present or the future that gets its moral significance from a fact about the past.

Two examples should help to make this distinction clear. Consider, first, the obligation to keep a promise. When I have a promissory obligation to do a_i, it is, we may say, because I promised someone, at some point in the past, that I will do a_i. On this construal, my obligation to do a_i derives directly from a fact about the past and is, hence, rooted in a past-regarding$_1$ factor. In contrast, consider my duty to promote the welfare of those who deserve it. Insofar as my obligation to do a_i is an instance of this duty, it is rooted in a fact about the present and the future, viz. the fact that it promotes or will promote the well-being of, say, Smith and Jones. However, this fact gets its moral significance from a fact about the past, viz. that Smith and Jones have done something at an earlier point in time that makes them especially deserving of well-being.

With this distinction in mind, consequentialists can answer the argument. P1, recall, says that consequentialist theories are incompatible with past-regarding factors. Consequentialists can accept this premise if the term "past-regarding" is interpreted as past-regarding$_1$. They can accept, that is, that consequentialist theories are incompatible with the idea that facts about the past matter in and of themselves. They can insist, however, that consequentialist theories can be past-regarding$_2$. They can hold, that is, that it is possible, on a consequentialist moral theory, to judge acts in terms of facts about the present and the future whose moral significance derives from facts about the past (cf. Vallentyne 1988). As for P2, consequentialists can claim that adopting the principle of Moral Desert merely makes a consequentialist theory responsive to past-regarding$_2$ factors. This gets clear when one combines this

change a fact about me in the present. It also changes a fact about Socrates. For it makes it true that Socrates lived in a world in which my head was scratched at that precise time. This, however, cannot be seen as a genuine change.

principle with the building blocks of a typical consequentialist theory. Consider, e.g., the following doctrine.

Desert-Adjusted Utilitarianism (DAU)
An act is right if and only if it maximizes the desert-adjusted sum total of happiness of all sentient creatures.

DAU contains all CU-components but drops Equal Treatment in favour of Moral Desert. It is like CU in that it judges the rightness of an act in terms of the happiness that it produces for the morally relevant individuals. Since no act can have an impact on past happiness, what determines its moral status, on DAU, is its present and future effects on the distribution of happiness. Unlike CU, however, DAU is responsive to past-regarding$_2$ factors. In evaluating the distribution of happiness to which an act leads, it takes into account facts about the past, viz. considerations of desert. It gives comparatively more weight to the happiness of deserving individuals.

Consequentialists can, hence, distinguish between two ways in which the above argument can be interpreted. The first is this:

(P1) Consequentialist theories are incompatible with past-regarding$_1$ factors.
(P2) A moral theory that endorses Moral Desert involves past-regarding$_2$ factors.
(C) Consequentialist theories are incompatible with Moral Desert.

On this reading, consequentialists can reject the argument because the conclusion does not follow.

The second interpretation is this:

(P1) Consequentialist theories are incompatible with past-regarding$_2$ factors.
(P2) A moral theory that endorses Moral Desert involves past-regarding$_2$ factors.
(C) Consequentialist theories are incompatible with Moral Desert.

On this interpretation, the conclusion follows. But consequentialists can deny the first premise. So it seems as though they can reject the argument in any case.

Some moral theorists would, of course, not accept this reply. They may still insist that consequentialism cannot be past-regarding in any way – not even past-regarding$_2$.[172] I believe, however, that it would be a mistake to rule out this possibility. For it would imply that we should exclude from the consequentialist family a doctrine which moral philosophers usually regard as a paradigm case of consequentialism, viz. the doctrine of PU that we came across on page 133. PU maintains that an act is right if and only if it maximizes overall preference satisfaction. As Jeff McMahan writes, most people "have preferences that extend beyond the boundaries of their lives – for example, preferences concerning the posthumous disposition of their property or the treatment of their dead bodies." (McMahan 2002, 497) To the extent that PU takes account of such preferences, it is, hence, past-regarding$_2$.[173] For it judges acts based on their present and future effects

[172] Such a view is defended, e.g., by Nida-Rümelin (1993). See, in particular, §14.

[173] The idea that the preferences of the dead should be counted seems not at all alien to proponents of PU. Kymlicka (2002, 17–18) ascribes such a view, e.g., to Richard Mervyn Hare. And Richard

but allows certain facts about the past – viz. the preferences of the dead – to play a part in determining their moral significance. By excluding past-regarding$_2$ doctrines from the family of consequentialist theories, we would, hence, exclude PU, which, I believe, *every* ethicist would class as consequentialist.

It is, however, possible to exclude past-regarding$_2$ consequentialist theories for the purpose of our inquiry. Again, this has to do with the fact that we use trolley cases. When we discussed the characteristics of trolley cases in Sect. 2.4.1 we said, in connection with Characteristic 3, that trolley cases have no history. This means that none of the individuals involved in a trolley case can be morally deserving since moral desert arises from past events. I believe, therefore, that we can permissibly eliminate any past-regarding$_2$ considerations from our discussion.

In summary, let us record, then, that there seem to be two types of welfarist consequentialism which deserve our attention. There are, firstly, consequentialist theories which adopt alternatives to Summation. In particular, versions that endorse Leximin, a Coarse-Grained view of the good, Average, Moderate Egalitarianism, and Multiplication should be seen as relevant to our investigation. There are, secondly, interesting versions of consequentialism that adopt Unequal Treatment and allow/require the moral agent to give preferential treatment to some individuals as against others according to various factors. We mentioned, in particular, the prioritarian view, which gives preference to those who are comparatively worse off, and self-referential altruism, which is the idea that the agent should give preference to individuals according to the nature of her relationship with them. Moral Desert, we just established, is not relevant in the context of our discussion.

4.2.4 Alternatives to Welfarism

Up to this point, we have only considered variants of consequentialism which, in keeping with Welfarism, assume that the goodness of an action depends only on the extent to which it promotes the appropriately weighted well-being of the morally relevant individuals. As we noted above, the logical complement of this view is Non-Welfarism, viz. the idea that goodness does not depend only on the extent to which it promotes overall welfare. One alternative in the domain of Non-Welfarism is Extra-Welfarism, which is the idea that there are further factors *X, Y,* ... (over and above overall welfare *W*) that determine the goodness of an act. The other is Anti-Welfarism, which is the idea that only other factors *X, Y,* ... count from the standpoint of goodness. The latter view is somewhat absurd because the well-being of individuals is, of course, important. That is why we will leave it aside and only consider Extra-Welfarism.

Brandt says that "[s]ome philosophers would not accept this restriction [viz. that the preferences of the dead should not count; NM] and therefore consider that we are creating utility if we satisfy the past desires of the dead, even those dead long ago." (Brandt 1992, 162)

Extra-Welfarism claims that there are additional factors besides overall welfare that determine the goodness of an act. To start, we might ask which possibilities there are. On page 97–98, recall, we briefly considered a very comprehensive list of items which, according to the philosopher William K. Frankena, are good in themselves. Perhaps, then, we can find additional factors on that list? It comprised items such as consciousness, health, various forms of pleasure, beatitude, contentment, knowledge, wisdom, beauty, virtues, love, friendship, freedom, and so on. Some of these things obviously get their significance from the fact that they are parts of well-being (e.g. pleasure and health). Others, however, are not commonly seen as parts of well-being (e.g. freedom). It does not seem even remotely contradictory to say that a person's life is high in well-being, but low in freedom. Perhaps, then, freedom would be one of the additional factors *X, Y,* ... that an extra-welfarist consequentialist philosopher would endorse? This thought seems initially appealing. Recall, however, our discussion of the various variants of welfarism in Sect. 4.1.1 and, in particular, the way in which we specified the meaning of Welfarism. We said that Welfarism, as we understand it, is a form of Technical Welfarism. As we discussed, the distinctive mark of Technical Welfarism is that, unlike Philosophical Welfarism, it is not subject to the requirement of "descriptive adequacy." That means that a theory can be welfarist in the technical sense even if it applies a measure of individual well-being which does not have much to do with how we intuitively think about well-being. The only requirement for this is, as we noted, that it construes goodness as a functional that takes as its arguments only the representations of well-being $u_1, \ldots, u_n$ of the morally relevant individuals. What is represented by $u_1, \ldots, u_n$ does not matter. They may, e.g., be interpreted as increasing functions of the hair that grows on the individuals' bodies. This, of course, means that a consequentialist theory which contains Welfarism as a component may, in fact, include many of the factors that we find on Frankena's list. There seem to be only two categories of things which cannot be construed as a part of the welfare of individuals. The first comprises things whose intrinsic goodness does not depend on the existence of individuals who are in a position to appreciate them. The second contains moral constraints.

G. E. Moore's Ideal Utilitarianism provides an example of the first kind. He famously opposed the purely hedonistic utilitarian theories of his precursors. They based their axiologies on the idea that only the mental states of individuals were good in themselves. Henry Sidgwick, e.g., claimed that "no one would consider it rational to aim at the production of beauty in external nature, apart from any possible contemplation of it by human beings." (Sidgwick 1907, 114) Moore begged to differ. He did consider this rational and gave one of his most infamous thought experiments to demonstrate it. He has us imagine a world that is "exceedingly beautiful" and contains everything we may admire, such as "mountains, rivers, the sea; trees, and sunsets, stars and moon." (Moore 1903/1959, 83) Then he has us imagine the ugliest world we can possibly think of, viz. one which is simply a "heap of filth, containing everything that is most disgusting to us." (Moore 1903/1959, 83–84) In addition, we are supposed to assume that there are no human observers who might enjoy the beauty of the one world or be skeeved out by the ugliness of

the other. Even so, says Moore, it is not irrational "to hold that it is better that the beautiful world should exist, than the one which is ugly." I, for one, do not share Moore's intuition. However, let us suppose he is right. Let us suppose, that is, that intrinsic goods like beauty did exist. Even if that was so, it would be irrelevant to our present investigation since we use trolley cases. In these thought experiments, non-welfarist factors like beauty do not play a role. Hence, we can put them aside for the purpose of our investigation.

What about the second class of factors, viz. constraints? Do they matter in trolley cases? They obviously may. Recall, e.g., *George's Case* that we considered on page 45. George faces a choice between letting five die and killing one. In this case, there are two factors which might conceivably matter from the moral point of view. The first is the consequences of George's act for the welfare of the individuals. The second is the fact that the two options involve different *kinds* of acts. One act can be described as *letting* the five die, while the other can be described as *killing* the one. As many moral philosophers believe, killing the one is forbidden since there is a moral constraint against killing, but no constraint against letting die. Constraints seem, therefore, to matter in trolley cases. Hence, consequentialists apparently do have a motivation to reject Welfarism in order to incorporate constraints into their moral theories. It appears to be problematic, however, for consequentialists to motivate constraints since the Core Idea of consequentialism seems to be flat out incompatible with them.

In what follows, we will first explore the consequentialist case for constraints. After that, we will make one brief point about them which suggests that there are limits to the ways in which consequentialists can use constraints to defend their moral outlook.

4.2.4.1 Motivation

Let us start by asking why consequentialists should not be able to allow for constraints ? Why should they not be able to say, e.g., that there is a constraint which forbids one to break one's promises? Initially, it seems that consequentialists can give an obvious explanation for such a prohibition. They can say that the breaking of a promise usually has bad consequences. And they can insist that this explains why promise-breaking is wrong. This reasoning should sound familiar as it reiterates a point that we made earlier in the discussion. In Sect. 4.2.1.2, we said that adherents of Indirect Consequentialism may recommend that moral agents use Type-2 heuristics to make moral choices. They may hold, that is, that moral agents should decide what to do based on certain heuristic properties $h_1, \ldots, h_m$ of acts instead of the intrinsically relevant normative factors, viz. consequences. As we said, Type-2 heuristics are, e.g., rules "such as 'Don't harm others', 'Don't take or harm the possessions of others', 'Keep your promises', 'Tell the truth', etc." (Hooker

2003, 142) They capture aspects of our behaviour that we intuitively view as being subject to moral constraints.[174]

This account, however, does not seem to capture what it means for there to be a constraint in the ordinary sense. When we say that there is a constraint against doing Φ, we do not mean that Φ-ing is wrong because it has this or that *effect*. We mean that it is wrong *in and of itself*. We believe that Φ-ing can be wrong, even in a case where it is, as far as consequences go, the best we can do in that situation. It seems as though consequentialists cannot account for such a constraint. Here is why. Consider the Core Idea behind consequentialism. It is the notion that the permissibility of an act depends only on one factor, viz. the goodness of its consequences. Given this idea, moral constraints cannot matter. Because, if they did, there would be one factor over and above goodness which determined whether an act is permissible. This is incompatible with the Core Idea.

What can consequentialists say in reply to this objection? They have to concede, of course, that their theories cannot acknowledge constraints in the *formal* sense. They cannot, on pain of contradicting the Core Idea, recognize constraints as *independent* normative factors. But they can insist that it is possible to construct consequentialist theories which are "deontically equivalent" (Portmore 2007, 40) to non-consequentialist theories that explicitly allow for constraints. To see how they might do this, consider Judith Jarvis Thomson's famous thought experiment *Transplant*.

> **Transplant**
> (. . .) imagine yourself to be a surgeon, a truly great surgeon. Among other things you do, you transplant organs, and you are such a great surgeon that the organs you transplant always take. At the moment you have five patients who need organs. Two need one lung each, two need a kidney each, and the fifth needs a heart. If they do not get those organs today, they will all die; if you find organs for them today, you can transplant the organs and they will all live. But where to find the lungs, the kidneys, and the heart? The time is almost up when a report is brought to you that a young man who has just come into your clinic for his yearly check-up has exactly the right blood-type, and is in excellent health. Lo, you have a possible donor. All you need do is cut him up and distribute his parts among the five who need them. You ask, but he says, "Sorry. I deeply sympathize, but no." Would it be morally permissible for you to operate anyway? (Thomson 1985, 1395)

Intuitively, you are not permitted to chop up the healthy patient to save the five. A ready explanation for this is that there is a constraint against harming the innocent. How can a consequentialist deal with this constraint? To be sure, she cannot say, *simpliciter*, that you are forbidden from chopping up the healthy guy because there is a moral prohibition against harming the innocent. She has to build the constraint into her theory of the good. To this end, she can hold that "the doing or refraining from doing, of certain kinds of acts are themselves intrinsically valuable states of affairs constitutive of the Good." (Alexander and Michael 2008) She can argue, e.g., that chopping up the healthy guy would have the consequence that an innocent person is

[174]This point is a guiding theme in the work of the consequentialist moral theorist Philip Pettit. See, e.g., Pettit (1987, 1988a, 1988b).

killed. And she can claim that this consequence is worse than five innocent persons *dying* because it involves a type of act that contributes so much intrinsic badness to the resulting state of affairs that this outweighs the good of saving a net four lives. This would explain why killing the one is wrong and letting the five die is right. And it would do so, it seems, in keeping with the Core Idea of consequentiaslism. For this reason, let us call this sort of consideration a *consequentialist constraint*.

In response, critics of consequentialism may point out that this move is problematic for various reasons. The first reason is that constraints seem to concern the action itself, while consequentialism focuses exclusively on its consequence. This objection, however, is based on a particular understanding of what it means for a moral theory to be consequentialist – an understanding which, recall, we had to reject in Sect. 3.1.1. The problem with it is that this interpretation of consequentialism supposes that we can draw a clear line between the act and its consequences. As we saw, however, this seems not to be so. When a person acts, there always appear to be various adequate descriptions. And each of these descriptions represents a different way of carving up what happens into the act and its consequences. Above, we illustrated this using the example of a college student who improved her grades throughout the semester. To describe what she did, she can say this: "Last semester I studied very intensely. As a consequence, my grades improved." Under this description, the student identifies the studying as her act, while the improvement of the grades is identified as her act's consequence. Alternatively, she can say this: "Through intense studying, I was able to improve my grades last semester." On this equally adequate description, the improvement of the grades is picked out as the student's act, not as its consequence. The bottom line is that the distinction between the act and its consequences has an air of arbitrariness to it. It seems, hence, unfit to serve as the basis of an argument which establishes that consequentialist theories cannot incorporate constraints.

Be that as it may, critics of consequentialism might still have an ace up their sleeves. They can say that even if the above explanation works, it does not capture what moral constraints are all about (cf. Nozick 1974, 28–30). Consider a modification of *Transplant*. Imagine yourself to be the surgeon again. This time, you do not face a choice between killing one and letting five die. Instead, you have to choose between killing one person or letting two be killed. Imagine, e.g., that there is an evil scientist who has captured two innocent people. He will kill them unless you kill your next equally innocent patient (perhaps in order to test whether you believe in the existence of moral constraints?). In this case, it still seems as though there is a prohibition against harming the innocent which morally prohibits you from killing your next patient. But the above reply will not do to explain this. Presumably, killings are bad, no matter who does them. Hence, if you can prevent two persons from being killed by killing one, doing this is the best you can do and should be morally permissible on a consequentialist view.

To deal with this problem, some consequentialist philosophers have suggested that the badness of a killing (and other bad things) may be *relative to the agent*. Amartya Sen illustrates this possibility with a story he takes from the *Mahabharata,*

which is an ancient Indian epic.[175] The part of the tale that Sen homes in on centers around two characters, the warrior Arjuna and his friend and advisor Krishna. The two argue about whether or not the rightful royal family of the Pandavas should fight against the Kauravas, who have wrongfully usurped their kingdom. Arjuna, who, due to his role as army commander, is expected to partake in the battle, turns to his friend Krishna and asks him

> whether all this is worth it. He does not doubt that theirs is the right cause, and that this is a just war, and also that his side will definitely win the battle given its relative strength – not least because of Arjuna's own remarkable skills as a warrior and a general. But, Arjurna observes, so many people will die in this battle. *He also recognizes that he himself will have to kill masses of people.* (Sen 2000, 481; emphasis added, NM)

The italicized phrase is important here. As Sen explains, Arjuna is not only concerned about the fact that many people will be killed in the battle. He takes note "of the special badness of the events as *he* must evaluate them." (Sen 2000, 485; emphasis added, NM) It is bad enough that many people will die. But this badness is compounded – from Arjuna's perspective – by the fact that "he himself would have to do some of the killing." What Sen is getting at with his remarks is that Arjuna makes a distinction between bad things that happen, *period*, and bad things that happen *as a result of his acts*. And consequentialists, Sen thinks, are well advised to follow Arjuna's example.

With the distinction between "good/bad" and "good/bad-relative-to-the-agent" in their conceptual toolbox, consequentialists can revisit the problematic case that we introduced above.[176] In the modified *Transplant* case, it seems arguable that it is, intuitively, not morally permissible for you, the surgeon, to kill your next patient even though the mad scientist will then kill two of his patients. There appears to be a constraint against doing harm to the innocent which takes precedent in this case. Consequentialists can match this constraint by arguing as follows. You, the surgeon, would have to kill one person to save two from being killed by the mad scientist. From an agent-neutral viewpoint, the killings are equally bad. It seems, hence, to be better if you were to kill your patient since this would result in one person being killed instead of two being killed. But, they can proceed, it is *your* point of view that counts here.[177] And from that perspective, one killing *done by you* may be worse than two *done by somebody else*. Hence, it is possible, on an agent-relative

[175]It should be noted that we are only looking at one aspect of Sen's rather complex view. Elsewhere, he addresses various forms of relativity and explores their logical connections. He calls them "doer-relativity," "viewer-relativity," and "evaluator-relativity." On these distinctions, see Sen (1983).

[176]The distinction between "goodness" and "goodness-for" has been an ongoing theme in recent ethical debate. A fare number of scholars have expressed their sympathies for it (e.g. Broome 1991, Sen 1982 and 1983; Portmore 2001, 2005 and 2011). For critical rejoinders see, e.g., Regan (1983), Schroeder (2007), and Brown (2011, 761–763).

[177]The idea that the position of the agent is particularly important for moral evaluation is emphasized by Sen (1993).

version of consequentialism, to embrace the view that it is wrong for you to kill one person to save two from being killed by someone else.

At this point, further criticisms may be raised, however. One may say that constraints cannot be factored into the axiologies of consequentialist moral theories because all consequentialist theories are allegedly agent-neutral (cf. Howard-Snyder 1996, 113–114; McNaughton and Rawling 1991). This may require some explaining. As we discussed in Sect. 4.1.3, two senses of agent-neutrality can be distinguished. The concept can be used in a *deontic* sense and an *axiological* sense. Recall that, if a theory is agent-neutral in the former sense, this means that it judges whether an act is permissible independently of the identity of the agent. If it is agent-neutral in the latter sense, it means that the theory contains an axiology which evaluates the act without reference to the agent who does it. The meaning of agent-neutrality that is relevant here is, I take it, the deontic one. It is claimed, then, that consequentialist theories judge whether an act is morally permissible without paying any regard to the identity of the agent and that they cannot, therefore, build constraints into their axiology.

The logical connection here is not obvious and seems to depend on further assumptions. A more detailed argument may run somewhat like this. It is supposed that all consequentialist moral theories are deontically agent-neutral. As we established earlier, a consequentialist theory is deontically agent-neutral if and only if it is axiologically agent-neutral since consequentialist theories determine the permissibility of an act only based on its axiological value. Hence, we can conclude that all consequentialist theories are axiologically agent-neutral. Now, it is sensible to suppose that the axiology of an axiologically agent-neutral theory cannot contain agent-relative factors. Constraints, however, seem to be such agent-relative factors. (For this reason, Scheffler (1985) calls them "agent-centred restrictions.") Therefore, it follows that consequentialist moral theories cannot factor in constraints.

The structure of this argument is as follows:

(P1) All consequentialist moral theories are deontically agent-neutral.
(P2) A consequentialist moral theory is deontically agent-neutral if and only if it is axiologically agent-neutral.
(P3) The axiology of an axiologically agent-neutral moral theory cannot contain agent-relative factors.
(P4) All constraints are agent-relative factors.
(C1) All consequentialist moral theories are axiologically agent-neutral. (from P1, P2)
(C2) The axiologies of consequentialist moral theories cannot contain constraints. (from C1, P3, P4)

The explicit form of the argument exposes its weaknesses. The premises P1 and P4 seem dubious. As for P1, it appears to be untrue that all consequentialist moral theories have to be deontically agent-neutral. A common counter-example is the moral doctrine of Ethical Egoism, which we considered in Sect. 4.2.3. Above, I argued that we should regard it as *both* agent-relative *and* consequentialist.[178] It is

[178] Cf. footnote 32 of Chap. 3.

agent-relative since, on Ethical Egoism, the permissibility of an act turns on the identity of the agent. My action is permissible if and only if it most promotes *my* self-interest, your act is permissible if it most promotes *your* self-interest, and so on. It is also true to the Core Idea of consequentialism since permissibility is solely a matter of goodness, viz. goodness for the agent.

But even if we grant that all consequentialist theories are agent-neutral, it still would not follow that they cannot allow for constraints since P4 is apparently false. It does not seem to be the case that all constraints are agent-relative factors. To be sure, some of them are. E.g., I may be morally constrained by my allegiance to somebody, while you are not. Then, the permissibility of my acts may depend, at least in part, on how it affects *that* person. In contrast, you may not owe any loyalty to her such that the permissibility of your acting does not, to the same extent, depend on how it affects her. But consider another constraint, such as the constraint against lying. If there is, in fact, such a constraint, it seems to be agent-neutral (cf. Skorupski 1995, 51). If lying is *prima facie* wrong, the fact that an act is a lie speaks against it, *period*. No reference to the agent is necessary. It seems, then, that the above argument is unsound and does not rule out that consequentialist theories can incorporate constraints.

Another concern may be raised, however. Moral philosophers often emphasize that consequentialist theories are exclusively future-regarding (cf., e.g., Graafland 2007, 174; Oddie and Milne 1991, 53; Carlson 1995, 10; Nida-Rümelin 1997b; Geirsson and Holmgren 2000, 110). They evaluate acts only based on their results which manifest solely in the future. Constraints, on the other hand, are past-regarding. Since exclusively future-regarding theories cannot incorporate past-regarding factors, consequentialist moral theories cannot, it seems, incorporate constraints. This reasoning, consequentialists may claim, is problematic for two reasons.

Firstly, there are constraints which are not past-regarding. To be sure, many of them relate to past points in time. E.g., the fact that I have given a promise at some time may morally constrain the way in which I act in the future. But not all constraints are of this type. Unlike the obligation to keep a promise, the duty not to lie, e.g., does not seem to depend on anything that happened in the past (unless, of course, I promised not to lie, in which case the obligation not to lie may be seen as an obligation to keep a promise). Therefore, consequentialist theories may, at least, incorporate certain constraints.

Secondly, it is not even clear that consequentialists cannot accommodate past-regarding constraints in their moral framework. In Sect. 4.2.3.2, we distinguished between two senses in which a normative factor can be past-regarding, viz. past-regarding$_1$ and past-regarding$_2$. If a factor is past-regarding$_1$, it is a fact about the past. If a factor is past-regarding$_2$, it is a fact about the present or the future that gets its moral significance from a fact about the past. This distinction is important. As we established, consequentialist theories cannot be past-regarding in the sense that they cannot take into account past-regarding$_1$ factors. But a good case, it seems, can be made for the possibility that consequentialist theories may take into account past-regarding$_2$ factors. Where does that leave us as far as promissory

obligations are concerned? A consequentialist, it seems, cannot say that a given act a_i is wrong because the agent at some point in the past promised to do $\neg a_i$. A consequentialist cannot say that because, if she did, she would point directly to a past-regarding$_1$ factor, i.e. a fact about the past. This is inconsistent with the Core Idea of consequentialism. However, it seems as though she can construe things differently. She can say, that is, that doing a_i will have the *consequence* that a promise is broken (cf. Broome 1991, 4). In that case, she points to a fact about the present/future that gets its moral significance from a fact about the past, viz. the fact that a promise was made. This seems to be legitimate.

As it turns out, then, consequentialists seem to have a good motivation to incorporate constraints into their moral theories.[179] Furthermore, it appears that consequentialists may be able to rebut the standard arguments for rejecting consequentialist constraints.

4.2.4.2 Options

As we have seen, consequentialists are apparently able to incorporate constraints into their moral theories. It may be of interest, then, to enquire which shapes a scheme of consequentialist constraints may take. We could say much about the various kinds of constraints, e.g. constraints against harming others, against lying, against promise-breaking, and so on. Furthermore, we could go into the issue of how consequentialists might weigh these *vis-a-vis* welfare considerations and how they might balance the relative importance of the various constraints *vis-a-vis* each other. As interesting as these issues might be in their own right, we will not, however, pursue them here.[180] For as it will turn out below, our argument does not hinge upon them. For us, it suffices to make one brief principled point that concerns the connection between constraints and negative duties.[181] This connection seems to

[179]It should, perhaps, be mentioned that the motivation for incorporating constraints is greater for some consequentialists than for others because some consequentialist views clash with moral constraints more often than others. To see this, consider, e.g., *The Roman Circus* case: "The Romans in the audience want the victims to die a painful death. The victims want to survive. If we have sufficiently many in the audience, then their desires should rule, and we have to say that what the Romans are doing is morally right." (Bykvist 2009, 48) This, at any rate, is what we should say if we believe in an aggregative consequentialist conception, such as CU. We can avoid this counter-intuitive verdict by incorporating a constraint against harming the innocent. But this is not the only way. If we accept a consequentialist view that is based on Leximin, we judge goodness by the well-being of the worst off, viz. the victims. They are obviously better off if they are not harmed. Hence, we can say that it would be wrong to torture and kill them in *The Roman Circus*. We do not need a constraint.

[180]For a helpful and comprehensive discussion of constraints, see Kagan (1998, 70–152).

[181]A number of authors have investigated which kinds of constraints a consequentialist moral framework can accommodate. On this issue, see, in particular, Kamm (1996, 239–243) and Scanlon (2001, 47) who draw the conclusion that there are certain kinds of constraints which cannot be represented in consequentialist terms.

restrict the way in which consequentialists can use constraints to defend their moral outlook against our objections. To explain how far and why this is the case, let us, first of all, clarify the notion of a negative duty.

The idea of a negative duty involves the distinction between actions and inactions (or "negative acts") that we made previously. In Sect. 1.2.1.2, we said that both actions and inactions are morally significant. Both in acting and in failing to act we can violate a moral duty. That is, we can *do* an act which is forbidden. And we can *fail* to do an act which is obligatory. As philosophical jargon has it, we violate a *negative duty* in the former case and a *positive duty* in the latter case (cf. Russell 1999, 249). In other words, negative duties forbid a given action and mandate inaction, while positive duties require an action and forbid inaction. For this reason, negative duties can always be fulfilled by doing nothing at all, while positive duties always require an action (cf. Kagan 1998, 131). This said, here is how constraints supposedly relate to negative duties.[182]

Connection Between Constraints and Negative Duties (CCND)
Constraints always generate negative duties, while negative duties can also be generated by other normative factors.

CCND says that, when there is a constraint, there will always be a negative duty, that is, a duty to refrain from the action that is specified by the constraint. We say, e.g., that moral agents are subject to a constraint against killing other persons. According to CCND, the presence of this constraint implies the existence of a negative duty. It is easy to see what this duty is. It is *not* to kill another person. However, on CCND, we cannot say that moral agents are subject to a constraint against not saving persons from being killed. Such a constraint would generate a positive duty to do an action, viz. save other people from being killed. And this, according to CCND, is not possible.

This said, let us investigate how CCND affects the way in which consequentialists can defend themselves against our arguments. For this purpose, recall our strategy. As we discussed in Sect. 3.2.2, we will try to produce an argument against consequentialism that proceeds *from the inside out*. That is, we will start by constructing a case which shows that the paradigmatic version of consequentialism, CU, violates provisional fixed points in our moral thinking. We will argue, say, that, in a given case, an option for acting, a_i, which, intuitively, is morally impermissible comes out as permissible, according to CU. Then, we will boldly claim that this constitutes an objection that affects all consequentialist theories. To defend themselves against this allegation, consequentialists will have to retreat to an alternative version of consequentialism on which a_i comes out as impermissible. One way of doing this is to argue that, unlike CU, their favoured consequentialist theory accommodates a constraint that generates a duty not to choose a_i. Now, here comes the point that we are seeking to make. If CCND is correct, then the consequentialist can plausibly give this kind of reply only if a_i is properly viewed

[182]This relation is pointed to, e.g., by Beauchamp and Childress (2001, 353).

as an action. If, however, a_i is the option to do *nothing*, then there cannot be a constraint that forbids it. For the duty not to do nothing, is a positive duty which, according to CCND, cannot be generated by a constraint. The bottom line is, then, that consequentialists cannot defend their moral outlook by claiming that a given option is subject to a consequentialist constraint if this option consists in doing nothing at all.

This conclusion holds, at any rate, if CCND is, in fact, correct. Perhaps, though, it can be challenged. To do this, one would have to come up with an example of a constraint that clearly generates a positive duty. One such example may be the constraint against promise-breaking. If I promise to Φ, this particular constraint gives rise to a *prima facie* duty to Φ. And if Φ is an action, it seems to create a positive duty, viz. the duty to Φ. This counter-example does not, however, stand up. Upon closer consideration, the constraint against promise-breaking, does not, by itself, generate the positive duty to Φ. Rather, it forbids me to promise that I will Φ and then not Φ (cf. Kagan 1998, 131). I can avoid violating the duty that this particular constraint generates by not doing anything at all. And that means that the constraint against promise-breaking, in fact, creates negative duties.

There are, however, counter-examples which, as far as I can see, may work. Consider, e.g., the obligation of gratitude which, intuitively, is a constraint.

> Suppose you freely bestow some good upon me, do me a favor (...). Then I seem to be under a "debt of gratitude" to you. But what exactly do I owe you? At the very least, no doubt, I owe you a certain kind of positive regard, a feeling of appreciation and thanks. But many people feel that obligations of gratitude go beyond this, that there is also a requirement that I *return the favor*. (Kagan 1998, 135; emphasis in the original)

In some circumstances, "returning the favour" may call for an action on my part. That means that the constraint that gratitude imposes upon me can give rise to a positive duty.

Something similar may be said about other constraints. Another obvious example is loyalty. Duties of loyalty, many people believe, bind us morally. You are expected, e.g., to "honour thy father and thy mother," as the Bible has it. This constraint may also give rise to positive duties because it is plainly impossible to honour somebody by doing nothing.

Perhaps, then, CCND is not fully accurate. I believe, nevertheless, that it is roughly correct. If the behaviour of a moral agent is subject to a moral constraint, that strongly suggests that the agent has a negative duty to refrain from a given act. To put it conversely, if the agent has a positive duty to do a particular act, that strongly suggests that this duty is not generated by a constraint, but by a different normative factor. This conclusion is all we need to make our argument in Chap. 5 plausible.

4.3 Summary

In Sect. 3.2.2, we laid out a methodic procedure for our investigation, viz. FRA_2. It comprised four steps. In this chapter, we completed (i), (ii), and (iii).

In the first section, we focused on (i), which consisted in factorizing a paradigmatic consequentialist theory into its component parts. We chose CU for this purpose. To start, we drew on a distinction that we had made very early on, viz. in Sect. 1.2. Every moral theory, we said, can be divided into a practical and a theoretical component. The former instructs us as to how we should go about making choices. The latter gives us principles for the evaluation of acts. In the case of consequentialist theories, which are monistic, the theoretical component, we established, can be represented as a criterion of rightness. CU, we said, can hence be factorized into a criterion of rightness and a practical instruction that tells us how to use it. We began our investigation with its criterion of rightness. It says that an act is right if and only if it maximizes the sum total of happiness of all sentient creatures. We broke this idea up into two further components, viz. a conception of the right (not to be confused with the criterion of rightness itself) and a conception of the good. CU's conception of the right is maximizing. It says that an act is right if and only if it maximizes the good. CU combines Maximization with the Classic Utilitarian Theory of Goodness (CUG). What is to be maximized, according to CU, is the sum total of happiness of all sentient creatures.

The first thing we noted about CUG is that it identifies goodness, G, with a value of overall welfare, W. We viewed W as a functional, f, that takes as its arguments only the appropriately weighted numerical representations of the well-being of the morally relevant individuals. Formally, the idea can be stated as the identity $G = W = f(u_1, \ldots, u_n; w_1, \ldots, w_n)$. We called it Welfarism (with a capital "W"). Having investigated CU's interpretation of W, we found that its abstract shape gives rise to four questions: Firstly, what is the nature of the functional f that combines $u_1, \ldots, u_n$ and $w_1, \ldots, w_n$ into a single value? Secondly, which weights $w_1, \ldots, w_n$ should be attached to the well-being $u_1, \ldots, u_n$ of the morally relevant individuals 1, $\ldots, n$? Thirdly, who is meant by $1, \ldots, n$? Fourthly, what do $u_1, \ldots, u_n$ represent?

CU's answer to the first question is Summation, viz. the idea that overall welfare is the sum of the appropriately weighted well-being of the morally relevant individuals. The second question is answered by Equal Treatment, that is, the idea that $w_1 = \ldots = w_n = 1$. Universalism is CU's answer to the third question. It says that $1, \ldots, n$ represent *all* sentient individuals. Finally, CU's answer to the fourth question is Welfare Hedonism. It says that individual well-being is measured by the balance of pleasures over pains. In conjunction, Maximization, Welfarism, Summation, Equal Treatment, Universalism, and Welfare Hedonism yield the view that an act is right if and only if it maximizes the sum total of happiness of all sentient creatures.

This *dictum*, we noted, leaves open whether we should interpret it in an objective or a subjective sense. CU is commonly interpreted as being committed to the idea of Objectivism, viz. that only the actual consequences of an act matter. When combined

with the aforementioned components, this yields the doctrine that an act is right if and only if it, *in fact*, maximizes the sum total of happiness of all sentient creatures.

Finally, we considered the practical component of CU. We said that CU, as it is commonly construed, subscribes to Directness. Directness is the view that the agent should use the criterion of rightness as a decision-making tool – in each and every choice situation. That is, she should investigate whether the acts available to her exhibit the features that, according to her moral theory, make it right.

Overall, then, CU can be construed as the conjunction of eight moral claims, viz. Maximization, Welfarism, Summation, Equal Treatment, Universalism, Welfare Hedonism, Objectivism, and Directness which are all logically independent of one another. This list of components leaves out some ideas that are commonly used to characterize CU, viz. impartiality, agent-neutrality, and aggregation. In a next step, we showed that these features are not components of CU, but are, rather, logical consequences of the components that appear in our factorization. Finally, we considered the motivation for each of the claims that make up CU and recognized that *prima facie* all of them have something to be said in their favour.

In the second section, we addressed steps (ii) and (iii) of FRA_2. Step (ii) required us to examine which determinable components can be put aside for the purpose of our investigation. We considered the distinction between Subjective and Objective Consequentialism, Direct and Indirect Consequentialism, and consequentialist theories with alternative theories of individual well-being. In all three cases, we concluded that the determinable components in whose domain these determinate components lie are unmotivated for our purposes. Our argument for putting aside the distinction between subjective and objective variants of consequentialism was that in trolley cases an objective consequentialist theory must have the same implications as a subjective consequentialist theory if the two are otherwise identical. We proposed an analogous argument in the case of direct and indirect variants of consequentialism. In the case of alternative theories of individual well-being, we pursued a slightly different strategy. We tried to show that it is possible to set up a trolley case, such that any two consequentialist theories must coincide in their moral implications concerning this case if they differ only in regards to their respective theories of well-being.

In step (iii) of FRA_2, we surveyed the alternatives to each of the five determinable components that we had not eliminated in step (ii). We noted that there were, logically speaking, innumerable such alternatives. But it seemed reasonable to suppose that only a few of these alternatives are relevant. These, we reasoned, should be those which can remedy the problems that have traditionally been associated with CU. It seemed, therefore, that a reasonable way to detect the most interesting variants of consequentialism was to consider the most pressing concerns that have incited moral theorists to abandon CU and to adopt a different moral view. Alternatives to the CU-components, it seemed, should suggest themselves in the process. This said, we began step (iii) with an investigation of the alternatives to Maximization.

The main alternative to Maximization is Satisficing. We investigated some its variants, examined which ones we can immediately put aside and which ones may

deserve closer examination. In the end, we concluded that two non-maximizing forms of consequentialism should be considered in step (iv) of our investigation, viz. Slote's Comparative Satisficing Consequentialism and Hurka's Maxificing Consequentialism. The former endorses the view that an act is right if and only if it is good enough as measured by a *relative* standard. To determine whether an act is right, we compare its goodness with the best available alternative. If it is above a given relative threshold, say 80 % as good as the best act, then it is right to do it. Hurka's version is considerably more complex. It is based on a distinction between two kinds of cases, viz. *A*-cases and *B*-cases. An *A*-case is one in which at least the best act alternative available achieves a given absolute standard of goodness. A *B*-case, in contrast, is one where no option for acting achieves that standard. Hurka's Maxificing principle says, then, that an act is right if and only if the following holds: (i) If the agent faces an *A*-case, her act produces enough good as defined by an absolute standard. (ii) If the agent faces a *B*-case, her act maximizes the good.

We proceeded by considering alternative conceptions of the good. In doing so, we examined alternatives to Summation, Equal Treatment, and Universalism. A few simple considerations sufficed to rule out any alternative to Universalism. We said that no version of non-universalist consequentialism can be even remotely plausible because a consequentialist who gives up Universalism eliminates certain individuals from the domain of morally relevant subjects *altogether*. This implies that their weal and woe is a matter of moral indifference, which is extremely implausible. It makes much more sense, we noted, to say that all sentient beings deserve to be given some regard though certain beings may have a higher moral status than others. This suggested that we should rather look towards alternatives to Equal Treatment. We recognized Prioritarianism and Self-Referential Altruism as interesting candidates. Furthermore, we identified Leximin, Multiplication, coarse-grained theories of the good, Average and Moderate Egalitarianism as interesting alternatives to Summation.

Finally, we examined alternatives to the CU-component Welfarism. We started by noting that the logical complement of Welfarism is Non-Welfarism, viz. the idea that goodness does not depend only on the extent to which it promotes overall welfare. Non-Welfarism can, hence, be interpreted, firstly, as Extra-Welfarism, that is, the idea that there are further factors *X, Y, ...* (over and above overall welfare *W*) that determine the goodness of an act and, secondly, as Anti-Welfarism, which is the idea that only other factors *X, Y, ...* count from the standpoint of goodness. Anti-Welfarism, we noted, is hardly credible since the well-being of the morally relevant individuals is obviously important. Thus, we put Anti-Welfarism aside and homed in on Extra-Welfarism. To start, we asked which possibilities lie in the sphere of Extra-Welfarism. We had previously considered a comprehensive list of items which may be seen as good in themselves, e.g. consciousness, health, pleasures, beatitude, and so on. On the encompassing interpretation of Welfarism that we had adopted, we can interpret all of these things as components of an individual's well-being. As we noted, there seem to be only two categories of factors that resist being construed as parts of a person's welfare. The first category comprises things whose intrinsic goodness does not depend on the existence of individuals who might appreciate

Table 4.1 Construction kit for consequentialist doctrines (after elimination)

Paradigmatic component	Alternative components			
Maximization	Comparative satisficing	Hurka's Maxificing		
Summation	Multiplication	Leximin	Plural egalitarianism	Coarse-grained theory of the good
Equal treatment	Prioritarianism	Self-referential altruism		
Welfarism	Incorporated constraints			

them. The second contains moral constraints. We put aside the first category of things and homed in on the second because the former are irrelevant in trolley cases, while the latter are not. We considered the consequentialist motivation for constraints which, contrary to received wisdom, appeared to be rather sound. After that, we made one concise point about constraints. We said that their existence strongly suggests the existence of a negative duty or, conversely, that we likely cannot explain the existence of positive duties by the existence of a constraint.

This, then, concludes steps (i), (ii), and (iii) of FRA_2. Table 4.1 summarizes their results.

Chapter 5
Joining the Dots

In this chapter, the rubber finally meets the road. Having completed steps (i), (ii), and (iii) of our methodic procedure, FRA_2, we will take on the pivotal step (iv). That is, we will put together a comprehensive case against consequentialism which shows, I hope, that all versions of the creed should be rejected. As we said in Sect. 3.2.2, we will develop our case *from the inside out*, that is, from one of its paradigmatic cases *viz*. CU, to the non-standard versions.

In Sect. 5.1, we will introduce and examine an initial case description to which we will refer as $Case_0$. Our aim in formulating it is to show that CU is an untenable moral doctrine because it violates some provisional moral fixed points. We will carefully consider our intuitive moral judgements about all choice options that are available to the agent and will record them in a table. Then, we will attempt to show that CU does, as a matter of fact, clash with our provisional fixed points about $Case_0$. To this end, we will analyse how a proponent of CU would have to judge the scenario and compare the result of this analysis with our carefully considered intuitive responses. Having shown that CU does, indeed, sin against our intuitions, we will formulate an objection against consequentialism that generalizes our finding. We will call it $Objection_0$. It maintains that no consequentialist moral theory can match our provisional fixed points about $Case_0$. To be sure, $Objection_0$ is false. However, as we discussed in Sect. 3.2.2, it is useful for methodic reasons. We can examine what a consequentialist would have to say to rebut $Objection_0$, and we can investigate the problems that she runs into if she gives the one reply or the other. Our aim is, of course, to show that, at some stage, *every* answer that the consequentialist might give to $Case_0$ runs into trouble.

In Sect. 5.2, we will consider and take apart various possible replies that a consequentialist moral philosopher may propose to rebut $Objection_0$. To illustrate, one of her answers may be that Maximization should be dropped in favour of a version of Satisficing. If indeed this is an adequate reply to $Objection_0$, we will try to rebut it by formulating a further case, $Case_1$, which shows that a consequentialist doctrine that adopts the proposed variant of Satisficing is also unacceptable. To this

N. Mukerji, *The Case Against Consequentialism Reconsidered*,
Theory and Decision Library A: 51, DOI 10.1007/978-3-319-39249-3_5

end, we will proceed, once again, from the inside out. That is, we will attempt to show that the paradigmatic satisficing doctrine (which accepts all CU-components except Maximization) yields implications that violate our moral intuitions in this new case. Then, we will generalize this finding once more. We will put forward a new claim, Objection$_1$, which says that no consequentialist theory that endorses the respective version of Satisficing can match our provisional moral fixed points in *Case*$_1$. After that, we will look for possible replies by the consequentialist and so on. We will proceed in that manner with all the other replies until we have eventually worked through all branches and sub-branches that grow out of the initial case, *Case*$_0$. As we noted previously, it looks like this process may carry on indefinitely. But that is not so. Every time we formulate a new charge against the consequentialist, she has to make a new commitment to a particular determinate component in the construction kit. At some point, then, she will inevitably run out of possible responses.

In Sect. 5.3, we finally sum up the main points made in this chapter before we conclude.

5.1 The Case Against Classic Utilitarianism

The starting point of our argument is the following case.[1]

> **Case$_0$**
> Jones is standing on a footbridge over a railway, as a runaway trolley carrying ten people is hurtling down the tracks. On every plausible theory of well-being, the lives of the ten, it shall be assumed, are as valuable to them as Jones's life is to him. Jones can tell that, if the trolley is not stopped, it will hit a massive rock at the end of the tracks. It is obvious that the impact will most certainly kill all ten people. If Jones does not do anything, this is what will happen. Jones has, however, two options that will avert this worst possible case. He can jump down onto the tracks. In that case, he would die. The trolley would run into him and squash him. But it would also come to a halt before it collides with the rock, thus saving the ten. Jones's second option is to throw a sandbag that is lying on the footbridge onto the tracks. This would not stop the trolley. But it would slow it down, such that on impact only the three people sitting at the very front of the trolley would be killed. Jones knows all of his options for acting and all of their respective consequences, and no other facts are morally relevant in this case.

Before we get started, we need to ascertain that this scenario is, in fact, a trolley case insofar as it possesses the respective characteristics that we identified in Sect. 2.4.1. This is important for an obvious reason. In Sect. 4.2.1, recall, we argued, following step (ii) of FRA$_2$, that we may justifiably ignore certain determinable components in our investigation. In particular, we excluded the distinction between subjective and objective consequentialist theories and the distinction between direct and indirect forms of consequentialism. Our argument for ignoring them relied on

[1] The inspiration for this case comes from a thought experiment in Mulgan (2001a).

the characteristics of trolley cases. We reasoned that, if we only use trolley cases in our case against consequentialism, we are permitted to neglect the distinction between them. After all, any two consequentialist doctrines that differ only insofar as they fall on one or the other side of this distinction will imply the same verdicts in a scenario that exhibits the characteristics of a trolley case. First things first, then, we should convince ourselves that $Case_0$ does, in fact, possess the features that ensure that.

- Firstly, the agent in a trolley case is supposed to face a tragic choice. That is clearly so in $Case_0$. No matter what Jones does, at least one person is going to die, and one valuable life is going to be lost.
- Secondly, Jones has a limited set of options. He can choose to do nothing, jump off the bridge or throw the sandbag.
- Thirdly, it can be assumed that there are no relevant normative factors, except for those explicitly stated.
- Fourthly, Jones's act uniquely determines the outcome of the case. If he does nothing, the ten people in the trolley will die. If he jumps, he will die. But this will save the ten. And if he throws the sandbag, the three people at the very front of the trolley will die. The description mentions no contingencies.
- Finally, Jones knows all these facts.

Thus, $Case_0$ does, indeed, possess all characteristics of trolley cases that we used as premises in previous steps of our argument.

Recall, furthermore, that we eliminated the distinction between consequentialist theories with alternative accounts of individual well-being from our investigation. Our argument for doing so was that we can detail trolley cases in such a way that any two consequentialist theories which are identical except regarding their theories of individual well-being yield the same implications. This, we said, will be the case whenever we assume that the lives of the morally relevant individuals are equally valuable on all plausible accounts of well-being. In $Case_0$, we make this assumption.

Let us record, then, that our exclusion of the distinction between Objective and Subjective Consequentialism, Direct and Indirect Consequentialism, and various alternative accounts of well-being is, indeed, legitimate, given that we deal only with cases like $Case_0$, which possess the five characteristics of a trolley case. Subsequent cases are variations of $Case_0$ and are also intended to have these features. We will, however, refrain from meticulously ascertaining this in each instance.

Without further ado, let us turn, then, to the substantive moral analysis of $Case_0$. As a first step, it might help to give it a more structured and concise presentation. Table 5.1 does that for us. It shows the options that are available to Jones, their overall consequences for the well-being of all individuals involved, and their implications for Jones's own well-being. Also, it contains the intuitive moral responses that the case elicits and which, I believe, can be taken to be provisional fixed points in our moral mindset.

The provisional fixed points should be immediately plausible as soon as we highlight some facts about the options in $Case_0$. As far as a_1 is concerned, it seems that the following should be said.

Table 5.1 $Case_0$

Options	Consequences (overall)	Consequences (for Jones)	Provisional fixed points
a_1 (=do nothing)	10 deaths	Alive	Impermissible
a_2 (=jump)	1 death	Dead	Permissible, but not required; supererogatory
a_3 (=throw sandbag)	3 deaths	Alive	Permissible, but not required; not supererogatory

In $Case_0$, Jones can, at no costs or risk to himself, save seven out of ten other persons from death. If he acts so as to save none, he acts wrongly on the face of it. Intuitively, he has, after all, a "duty to rescue," as it is recognized by many legal codes in various countries and cultures. For this reason, it appears to be clearly impermissible for Jones to choose a_1.

It seems, however, surely permissible for Jones to do a_2, that is, jump off the bridge to save the ten persons. In Sect. 4.2.2.1, we noted that, intuitively, moral agents have what Michael Slote has called an "agent-sacrificing option." That is, in a situation where they can sacrifice their own good for the sake of other people, they are morally permitted do so (cf. Slote 1985a, 11–12). Hence, Jones does not seem to do anything wrong if, indeed, he jumps. Now, this may seem doubt-worthy to some. What if Jones has children who need him? What if he is about to finish a project that is vital for the future of humanity? Perhaps he is about to find a cure for cancer? Sure enough, all these considerations might be important in a real-life case. But remember that we are dealing with a trolley case here. Its description mentions nothing of the sort. Hence, we may assume that no such facts obtain. The verdict that jumping is morally permissible for Jones can, therefore, be taken to be a provisional fixed point.

But is it *required* of Jones to jump? Most philosophers, I believe, would insist that it is not (cf., e.g., Scheffler 1982/1994, 23; Dorsey 2009, 144).[2] Jones may, of course, do the good deed to save all ten people instead of saving merely seven. However, given "that the only permissible means he has of doing the good deed is killing himself, (..) he may refrain from doing the good deed." (Thomson 2008, 365) This is particularly plausible because Jones has another option which will save seven persons at no risk or cost to himself. It appears reasonable to judge, therefore, that jumping is indeed optional.

We can convince ourselves that jumping is, in fact, most likely supererogatory. On page 137, recall, we formulated three (minimal) conditions for supererogation, viz. a Permissibility Requirement, a Self-Sacrifice Requirement, and an Altruism Requirement. Jumping fulfils the Permissibility Requirement as it is obviously permissible. On the assumption that the alternative a_3, viz. throwing the sandbag, is also permissible, it furthermore fulfils the Self-Sacrifice Requirement. For there is, in that

[2]Fishkin (1982, 14) generalizes this intuition to a principle he calls "Cutoff for Heroism." According to him, it is part of the "basic structure of individual morality."

case, at least one eligible option, viz. to throw the sandbag, which would be better from the self-interested standpoint of the agent. Finally, the Altruism Requirement is also fulfilled. From an altruistic point of view, it would, in fact, be preferable for Jones to jump. This would save three people who would die if he would merely throw the sandbag. Hence, jumping does, in fact, seem to be supererogatory. At least, there is not, as far as I can see, any reason to suspect that it is not.[3]

Finally, doing a_3, viz. throwing the sandbag, surely seems to be morally permissible for Jones. By doing it, he saves the lives of seven out of ten persons who would otherwise die. It is not morally required, however, for Jones to throw the sandbag. This is so for the obvious reason that there is another permissible option, viz. for him to jump off the bridge. Throwing the sandbag is not supererogatory, however, since it appears to fail both the Self-Sacrifice Requirement and Altruism Requirement.

Apparently then, there is from an intuitive standpoint no required act in $Case_0$. What is mandatory, however, is the disjunction of a_2 and a_3. It is clear that one of these acts Jones *ought*, morally, to do.

In the remainder of this chapter, we shall try to show that $Case_0$ presents a problem for consequentialism that they cannot overcome. To this end, we will proceed, as we said again and again, from the inside out. That is, we will start with a paradigmatic version of consequentialism, viz. CU, before considering its more peripheral cases. The starting point of our argument is, hence, the classic utilitarian analysis of $Case_0$.

CU, as we stated it above, says that an act is right to do if and only if it maximizes the sum total of happiness of all sentient creatures. Since the lives of all the persons involved in $Case_0$ have, by assumption, the same value (on all plausible accounts of individual well-being), jumping off the bridge, a_2, is the happiness-maximizing and, hence, uniquely best act. It saves 10 out of 11 persons' lives. In comparison, ten people die if Jones chooses option a_1 and does nothing. And three people die if Jones opts for a_3, that is, if he throws the sandbag. That means that, on CU, it is morally permissible for Jones to jump, while it is morally forbidden for him to do nothing or throw the sandbag. CU's moral verdicts are, hence, not in line with our firmly held moral intuitions about $Case_0$. In the case of a_2, it confuses what appears to be a merely supererogatory act with a morally required act. And with respect to a_3, it mistakenly judges that an act which seems to be clearly morally permissible is morally forbidden. CU, it appears, only gets it right about a_1, which it takes to be morally impermissible. Table 5.2 sums up these findings.

Of course, $Case_0$ does not show that we can reject consequentialism as a whole. It merely suggests that CU is unacceptable. However, remember what we said

[3]Note, however, that some philosophers would not concur with our judgement that the self-sacrificial suicide of Jones is morally commendable. Ayn Rand, e.g., was notorious for denouncing altruism and self-sacrifice (see Rand 1982, 58–76). Interestingly, Judith Jarvis Thomson, too, has expressed views which suggest that she might not agree. "A willingness to give up one's life," she writes, "*simply* on learning that five others will live if and only if one dies is a sign of a serious moral defect in a person." (Thomson 2008, 366; emphasis in the original).

Table 5.2 The classic utilitarian analysis of $Case_0$

Options	Consequences (overall)	Consequences (for Jones)	Provisional fixed points	Classic utilitarianism
a_1 (=do nothing)	10 deaths	Alive	Impermissible	Impermissible
a_2 (=jump)	1 death	Dead	Permissible, but not required; supererogatory	Required
a_3 (=throw sandbag)	3 deaths	Alive	Permissible, but not required; not supererogatory	Impermissible

above. CU is not just any consequentialist doctrine. It is the *paradigm* case of consequentialism. Hence, it seems reasonable to suspect that consequentialism more generally is affected by the problem that we just identified. At any rate, a proponent of consequentialism has, it appears, the *onus* to explain how a consequentialist doctrine can match our intuitions about $Case_0$. She has, in other words, the *onus* to rebut the following charge.

Objection$_0$
No consequentialist moral theory can match our intuitive responses about $Case_0$.

5.2 Consequentialist Replies

Having stated Objection$_0$, the question arises how a consequentialist may rebut it. Obviously, she has to demonstrate that it is false. To this end, she has to show that she can avoid the problem associated with CU in $Case_0$ by diverging from one of its components to an alternative. For this purpose, she must first choose a workable defence strategy. That is, she must decide which CU-component she wants to drop and which alternative component she wants to endorse instead. Then, she has to show that the doctrine which results from this modification does, in fact, enable her to come up with an appropriate moral evaluation of $Case_0$. There are some alternative components that consequentialists can choose. We summed them up in Table 4.1 at the end of Sect. 4.3. On the following pages, we shall consider each of them one by one.

5.2.1 Consequentialist Constraints

Let us first consider the possibility that consequentialists can rebut Objection$_0$ by diverging to a consequentialist theory that incorporates moral constraints. Above we noted that, initially, it seems as though this strategy is not at all congenial with the overall consequentialist position. However, as we discussed in Sect. 4.2.4.1, the

most common arguments against the idea of consequentialist constraints might, in fact, be untenable.[4] Hence, incorporating constraints may, indeed, be a good defence strategy for consequentialists. Let us ask, therefore, whether consequentialists can, in fact, apply this strategy to dodge Objection_0. For convenience, we will call consequentialists who endorse constraints "constrained consequentialists."[5]

To answer this question, let us remind ourselves, first of all, how constrained consequentialists can incorporate constraints into their moral theories. Above, we noted that they cannot simply say that there is a constraint against performing a given act. In doing that, they would posit an additional normative factor over and above the goodness of the action, which would violate the Core Idea of consequentialism. To model a constraint in a way that is consistent with the Core Idea of consequentialism, constrained consequentialists have to claim, rather, that a given act is *intrinsically bad*. And they have to claim, furthermore, that its intrinsic badness is reflected in the resulting state of affairs.

We have illustrated how this strategy might work by applying it to the scenario *Transplant* that we discussed on page 169. In this case, you, the surgeon, are faced with a choice between killing one person, thereby saving five others, and letting five persons die by refraining from killing the one. Most people agree that, in this situation, it is impermissible for you to kill the one to save the five and permissible to let the five die. A consequentialist moral theory, however, seems to imply the opposite. It appears to suggest that it is permissible for you to kill the one and impermissible to let the five die. After all, in the *Transplant* case, killing the one seems to have the best consequences. As we discussed, a constrained consequentialist can argue, however, that we have to take into consideration the fact that you would *kill* a person to save the five. The intrinsic badness of this act, she can claim, outweighs the goodness of the five being saved. Therefore, she can insist, the counter-intuitive conclusion that it is permissible for you to kill the one does not, in fact, follow from all consequentialist moral theories. At any rate, she may say, it does not follow from *her* theory.

At this point, we need not judge whether the way in which the constrained consequentialist weasels out of the *Transplant* case is ultimately convincing. (Non-consequentialists, I am convinced, will generally have strong doubts about that). As it turns out, we only need to assess whether she can reasonably give an analogous reply to dodge $\textit{Case}_0$. To this end, we should note, first of all, that $\textit{Case}_0$ is slightly different from the *Transplant* case. In *Transplant*, the challenge for the constrained consequentialist is to explain why killing the one is wrong and letting the five die right though her moral theory apparently implies the opposite. In $\textit{Case}_0$, the challenge is to explain why both a_2 (=throwing the sandbag) and a_3 (=jumping off the bridge) are permissible, while, on CU, only the latter comes out as permissible. The only way for the constrained consequentialist to do this is to claim that a_2 is

[4] Of course, if there is an argument which establishes that consequentialist constraints are impossible, this would only strengthen our case.

[5] In doing that, we follow a terminological suggestion by Ott (2004, 95).

an intrinsically bad act and that, once we take this fact into account, it ties with a_3 regarding goodness. This, in turn, would mean that both a_2 and a_3 are best acts and thus permissible which, intuitively, is the right view.

This argument of the constrained consequentialist is problematic for at least two reasons. Firstly, it is hard to see why it should be intrinsically bad for Jones to jump off the bridge. In and of itself, jumping off a bridge seems to be a morally neutral act. It is not the type of act that is subject to a moral constraint. But perhaps the problem does not lie in the fact that Jones jumps. Maybe the morally problematic aspect is that, in doing it, Jones would *kill himself*. Some philosophers – e.g. Immanuel Kant[6] – have indeed argued that moral agents have a duty to themselves not to commit suicide. It should be noted, however, that by jumping off the bridge Jones would not commit an ordinary suicide. His death would *save* three other people from their death. I believe, therefore, that his act would strike one, rather, as an heroic deed for a good cause. On some accounts of practical rationality, it may be seen as irrational. But it can hardly be seen as immoral. Thus, it seems entirely unmotivated to say that it is intrinsically bad.

However, even if we grant the constrained consequentialist that the act of jumping is, in fact, intrinsically bad, a second objection applies. To formulate it, we have to recall one point that we made in Sect. 3.2.2. We said that at certain stages in the argument it may be useful to augment our initial case, $Case_0$, with an additional case, $Case_{0*}$, and to press both of them against CU simultaneously. Having demonstrated that CU violates our provisional moral fixed points about both $Case_0$ and $Case_{0*}$, we can formulate a new charge, Objection_{0*}, which claims that no consequentialist theory can match our provisional fixed points about $Case_0$ and $Case_{0*}$. This procedure may be useful, we said, for the following reason. Perhaps consequentialists can dodge $Case_0$ by diverging from a paradigmatic component, C_{i1}, to an alternative component, C_{ij} (in this case, from Welfarism to a constraint-based view). If so, they can rebut Objection_0. But it may turn out that this strategy does not allow them to dodge both $Case_0$ and $Case_{0*}$ simultaneously. In that case, diverging to C_{ij} would not help them to rebut Objection_{0*} and we can put this component aside. Let us examine, therefore, whether we can formulate an Objection_{0*}. To this end, we have to introduce, first of all, a $Case_{0*}$.

Case_{0*}
The facts of the case are almost as they were in $Case_0$. However, this time Jones's third option has slightly different consequences. If he throws the sandbag, *four* persons die (as opposed to *three* in $Case_0$).

Like in $Case_0$, Jones has three options in $Case_{0*}$. He can do nothing, jump, or throw the sandbag. Let us again refer to these acts as a_1, a_2, and a_3, respectively. $Case_{0*}$ is of course very similar to $Case_0$. The only difference is that, in $Case_{0*}$, a_3, i.e. throwing the sandbag, would save only six out of ten persons' lives, while, in $Case_0$, this act would save seven out of ten. This difference does not, however, seem to affect our intuitive convictions about the permissibility of Jones's options.

[6] Kant's position can be found, e.g., in Kant (1785, 53–54).

In $Case_{0*}$, it still seems morally forbidden to do a_1, while both a_2 and a_3 appear to be permissible. That is, it still appears to be forbidden for Jones to do nothing, while both jumping and throwing the sandbag seem permissible. Furthermore, like in $Case_0$, a_2 seems to be supererogatory, as it is also altruistic and self-sacrificial. CU, however, implies that, in $Case_{0*}$, a_1 and a_3 are morally forbidden, while a_2 is the only morally permissible option and thus obligatory. Like in $Case_0$, CU only gets it right when it comes to a_1. Once again, it mistakes a supererogatory act, viz. a_2, for an obligatory act, and it forbids an act, viz. a_3, that is intuitively permissible. As it turns out, then, both $Case_0$ and $Case_{0*}$ seem to represent valid objections against CU.

In keeping with our inside-out methodology, we can generalize this finding into a more general objection that we direct at all forms of consequentialism.

Objection$_{0*}$
No consequentialist moral theory can match our intuitive responses about $Case_0$ and $Case_{0*}$ simultaneously.

At this point, we should ask whether there is any way for the constrained consequentialist to rebut Objection$_{0*}$. To this end, she would have to show that a constrained consequentialist is capable of matching our provisional fixed points about both $Case_0$ and $Case_{0*}$. Let us analyse, first of all, what she would have to claim to bring the implications of her theory in line with our intuitive convictions about $Case_0$.

It is clear that a constrained consequentialist has to make very specific assumptions about the amount of intrinsic badness in a_2 in $Case_0$. a_2 and a_3, mind you, have to tie in terms of their axiological value, such that both come out as morally permissible on a maximizing consequentialist view. If they do not tie, only one of them will be categorized as permissible, while the other will be classed as morally forbidden. Therefore, the constrained consequentialist has to make the following commitment in order to address $Case_0$.

Commitment 1
The amount of intrinsic badness associated with a_2 in $Case_0$ precisely offsets the difference in goodness between a_2 and a_3 that can be attributed to the respective difference in the numbers of deaths.

On the face of it, Commitment 1 seems entirely arbitrary. For argument's sake, however, let us grant the constrained consequentialist this assumption. The question we have to ask is whether it is possible for the constrained consequentialist, given Commitment 1, to match our intuitive convictions about $Case_{0*}$. It is immediately clear that we have to answer this question in the negative. As Table 5.3 shows, the constrained consequentialist is committed to the view that only a_2 is morally permissible in $Case_{0*}$. She has, after all, assumed that the intrinsic badness of a_2 is equal in value to precisely two deaths. If that is so, a_2 is axiologically equivalent to an act that results in three deaths. Since a_3 results in four deaths, a_2 comes out as the axiologically best act and is thus the only permissible option. Given her answer to $Case_0$, it is, hence, not possible for the constrained consequentialist to solve $Case_{0*}$. This means, in turn, that she cannot rebut Objection$_{0*}$.

Table 5.3 The constrained consequentialist analysis of Case$_{0*}$

Options	Consequences (overall)	Consequences (for Jones)	Provisional fixed points	Constrained consequentialism (+ Commitment 1)
a_1 (=do nothing)	10 deaths	Alive	Impermissible	Impermissible
a_2 (=jump)	1 death	Dead	Permissible, but not required; supererogatory	Required
a_3 (=throw sandbag)	4 deaths	Alive	Permissible, but not required; not supererogatory	Impermissible

Our case against the constrained consequentialist can, in fact, be strengthened by introducing yet another case.

Case$_{0**}$
The facts of the case are almost as they were in *Case*$_0$. However, this time Jones's third option has slightly different consequences. If he throws the sandbag, *two* persons die (as opposed to *three* in *Case*$_0$).

Obviously, *Case*$_{0**}$ also represents a valid objection against CU. As in *Case*$_0$ and *Case*$_{0*}$, CU implies that the only permissible option for Jones in *Case*$_{0**}$ is to jump off the bridge, although, intuitively, it is also permissible for him to throw the sandbag. Based on this finding, we can formulate an additional objection against consequentialism.

Objection$_{0**}$
No consequentialist moral theory can match our intuitive responses about *Case*$_0$ and *Case*$_{0**}$ simultaneously.

As Table 5.4 shows, a constrained consequentialist cannot rebut Objection$_{0**}$. To solve *Case*$_0$, recall, she has to endorse Commitment 1. She has to assume, that is, that the intrinsic badness of a_2 is equal in value to precisely two deaths, which means that a_2 is axiologically equivalent to an act that results in three deaths. Since in *Case*$_{0**}$ a_3 results only in two deaths, a_3 comes out as the axiologically best act. It becomes, thus, the only permissible option. This is intuitively implausible because the agent-sacrificing option, a_2, seems also permissible. Hence, the constrained consequentialist cannot rebut Objection$_{0**}$. She cannot give a plausible answer to both *Case*$_0$ and *Case*$_{0**}$.

Let us conclude, then, that consequentialists cannot persuasively rebut our case against their position by diverging to a version of their view that factors in moral constraints.

Table 5.4 The constrained consequentialist analysis of $Case_{0**}$

Options	Consequences (overall)	Consequences (for Jones)	Provisional fixed points	Constrained consequentialism (+ Commitment 1)
a_1 (=do nothing)	10 deaths	Alive	Impermissible	Impermissible
a_2 (=jump)	1 death	Dead	Permissible, but not required; supererogatory	Impermissible
a_3 (=throw sandbag)	2 deaths	Alive	Permissible, but not required; not supererogatory	Required

5.2.2 Slote's Comparative Satisficing

The second strategy that a consequentialist may choose to rebut $Objection_0$ is to diverge to a non-maximizing variant of consequentialism. As we established in Sect. 4.2.2.2, there are two alternatives in the domain of Non-Maximization that initially seem promising. A consequentialist may either opt for a variant of Slote's Comparative Satisficing or a variant of Hurka's Maxificing. Let us start with the former.

Slote's Comparative Satisficing purports that an act is morally permissible if and only if it is *good enough*, where the relevant standard of comparison for enoughness is the axiologically best act. As we have already worked out, from the standpoint of overall welfare, the best thing for Jones to do in $Case_0$ is to choose a_2, that is, jump off the bridge. This would save 10 out of 11 lives. Jones would be the only person who dies. The second best act is a_3, that is, to throw the sandbag. It would save 8 out of 11 lives. Seven out of ten persons in the trolley would survive if Jones did a_3 – and so would Jones himself. The third best act – or, rather, the worst act – in $Case_0$ is a_1, that is, not to do anything at all. If Jones were to choose that option, only he would survive, and all ten people in the trolley would die. Now, what would a consequentialist who decides to drop CU's Maximization in favour of Slote's Comparative Satisficing (a "comparative satisficing consequentialist", henceforth) have to argue in order to rebut $Objection_0$?

The comparative satisficing consequentialist would obviously have to make the following claims about $Case_0$.

- Firstly, since a_1 is intuitively impermissible, the comparative satisficing consequentialist would have to hold that the level of goodness that Jones achieves by doing it is too far below the amount of goodness that he might have achieved, had he chosen the best option, a_2. Therefore, she could argue, a_1 is morally wrong.
- Secondly, since a_2 does appear to be morally permissible, the comparative satisficing consequentialist would have to argue that it is, in fact, good enough.

This, she can claim, is evidently the case since the best possible act is always good enough and thus permissible on a comparative satisficing view.

- Thirdly, a_3 seems to be morally permissible, too. Hence, the comparative satisficing consequentialist has to claim that it is good enough in comparison with the best possible act, a_2, and therefore also permissible.
- Fourthly, the comparative satisficing consequentialist would have to explain our intuitive verdicts regarding the supererogatory status of a_2 and the non-supererogatory nature of a_3. On her comparative satisficing view, a_2 obviously fulfils all of the requirements for supererogation that we stated on page 137. It is permissible. According to the case description, it also involves a self-sacrifice, viz. the agent's death. Furthermore, it is altruistic since it is preferable, from an altruistic standpoint, to at least one permissible alternative, viz. a_3, that is, throwing the sandbag. In contrast, a_3 fails both the Self-Sacrifice Requirement and the Altruism Requirement. Amongst the permissible acts in $Case_0$, a_3 is most favourable from the perspective of Jones's self-interest. It does not involve any self-sacrifice. Neither is there a further permissible act which, from an altruistic standpoint, is dispreferred *vis-a-vis* a_3.

As it turns out, then, a comparative satisficing consequentialist can successfully rebut Objection_0. Note, however, that she can do so only on the condition that she makes the following two commitments since these seem to be the only reasonable, principled explanations for the four claims that we just stated.

Commitment 2
In a case where 11 lives are at stake, and an agent can save, at best, 10 out of these 11 lives, an act which saves one life is not good enough and, hence, morally impermissible.

Commitment 3
In a case where 11 lives are at stake, and an agent can save, at best, 10 out of these 11 lives, an act which saves eight lives is good enough and, hence, morally permissible.

Let us consider how we may come up with a follow-up objection that rebuts a comparative satisficing consequentialist view that endorses Commitments 2 and 3. To this end, it seems helpful to draw on an objection to Slote's Comparative Satisficing that has been raised by Tim Mulgan. Drawing on a case similar to the following $Case_1$, he argues that "satisficers get away with murder."[7] (Mulgan 2001a, b, 139).

Case_1
Jones is standing next to Smith on a footbridge over a railway, as a runaway trolley carrying nine people is hurtling down the tracks. On any plausible account of well-being, the lives of the nine, it shall be assumed, are as valuable to them as Jones's and Smith's lives are to Jones and Smith, respectively. Jones can tell that, if the trolley is not stopped, it will hit a massive rock at the end of the tracks. The impact will most certainly kill all nine people in the trolley. If Jones does not do anything, this is what will happen. Jones has, however, three options that will avert this worst possible case. As in $Case_0$, Jones can jump down onto the

[7] In this connection, see also the exchange between Turri (2005) and Mulgan (2005) that followed the original paper by Mulgan (2001a).

> tracks or throw a sandbag. If he jumps, the trolley will be stopped, and none of the people in it will die. Only Jones himself will die. If Jones throws the sandbag, this will not halt the trolley. But it will slow it down, such that, on impact, only the two people sitting at the very front will be killed. In addition, Jones has a fourth option. He can push Smith over the rim of the bridge and onto the tracks. This will not bring the trolley to a halt (as when Jones jumps) because Smith's body is lighter than Jones's. Like throwing the sandbag, however, it will slow down the trolley, such that, on impact, only the two people sitting at the very front of the trolley will be killed. But Smith will obviously die, too.

In $Case_1$, Jones's options are very similar to his options in $Case_0$. He can do nothing, jump, throw the sandbag, or push Smith over the rim of the footbridge. Let us refer to these options as a_1, a_2, a_3, and a_4, respectively. Now, how should these acts be evaluated? If Jones chooses a_1, the nine people in the trolley will die. So it appears evident, for one thing, that choosing this option is morally impermissible. There is, after all, an alternative act, a_3, which would save seven out of those nine people and would harm nobody. Furthermore, a_2 has the best outcome and would harm only Jones. The harm to Jones does, however, seem to be morally acceptable since Jones is himself the agent. If he freely chooses to suffer it, there appears to be no objection to option a_2 from the moral point of view. It should, hence, be permissible for him to do a_2. In fact, this act seems to be supererogatory because it is not only permissible and self-sacrificial but also altruistic. There is, after all, a permissible alternative which is dispreferred from an altruistic point of view, viz. a_3, that is, throwing the sandbag. Finally, a_4 seems to be morally forbidden. If Jones pushes Smith off the bridge, he would thereby save the same seven people that he would have saved, had he thrown the sandbag. However, he would thus kill Smith and, hence, cause one unnecessary death. This means, furthermore, that a_3 cannot be supererogatory. With a_1 and a_4 being ruled out as impermissible, no permissible option exists that is less preferable from an altruistic point of view than a_3.

Having laid bare these provisional fixed points, the following all-important question arises: Is it possible for the comparative satisficing consequentialist who has made Commitment 2 and Commitment 3 to match them?

Our comparative satisficing consequentialist can obviously match our provisional fixed points about a_1. She can claim that it is not good enough for Jones to save 2 out of 11 people (viz. himself and Smith), given that he could, at best, save ten people, viz. Smith and the nine people in the trolley. This is compatible with her comparative satisficing view and with Commitments 2 and 3. Commitment 2 says that, in a case such as $Case_1$ where 11 lives are at stake and an agent can save, at best, 10 out of these 11 lives, an act which saves one life is not good enough and, hence, morally impermissible. This is perfectly consistent with the view that saving two lives in such a situation is not good enough either.

The comparative satisficing consequentialist can, furthermore, match our provisional fixed points about a_2. It seems that a_2 is permissible and supererogatory. Permissibility follows from her Commitment 3. Commitment 3 says that in a case, such as $Case_1$, where 11 lives are at stake and the agent can save, at best, 10 out of these 11 lives, an act which saves eight lives is good enough and, hence, morally permissible. If Jones does a_2 in $Case_1$, 10 out of 11 lives will be saved.

By Commitment 3, a_2 is, therefore, morally permissible. Since it is, furthermore, self-sacrificial and altruistic, this makes it also supererogatory.

In addition, the comparative satisficing consequentialist can match our provisional fixed points about a_3. She assumes that saving eight lives is good enough and morally permissible in a situation like $Case_1$. a_3 saves nine lives. By Commitment 3, it, too, should therefore be permissible. This is intuitively correct.

Finally, what about a_4, i.e. the option of pushing Smith? As we just said, it is intuitively morally impermissible. It does not make sense for Jones to kill Smith in order to save seven out of nine people in the trolley if he could save the same seven people by throwing the sandbag. Throwing the sandbag, after all, would not harm Smith. Now, the comparative satisficing consequentialist runs into trouble. For she must judge that a_4 is permissible. $Case_1$ is, after all, a situation in which 11 lives are at stake, and the agent can save, at best, 10 out of these 11 lives. By Commitment 3, an act which saves 8 out of these 11 lives is good enough and, hence, morally permissible. Given her reply to $Case_0$, the comparative satisficing consequentialist seems, then, to be logically committed to the view that it is morally okay for Jones to kill Smith in $Case_1$. This is why Tim Mulgan says that "satisficers get away with murder." Table 5.5 records our findings.

This does not mean, of course, that the comparative satisficing consequentialist is trapped. She can insist that we misconstrued her view and diverge from one of the components that we assumed she would endorse. To see which options she has, we should, in keeping with our inside-out method, formulate a new objection.

Objection$_1$
No comparative satisficing consequentialist moral theory can match our intuitive responses about $Case_1$.

It seems, however, that we should not go down this path since it gives the comparative satisficing consequentialist an obvious way out. She might insist that

Table 5.5 The comparative satisficing analysis of Case$_1$

Options	Consequences (overall)	Consequences (for Jones)	Provisional fixed points	Comparative satisficing consequentialism (+ Commitments 2 and 3)
a_1 (=do nothing)	9 deaths	Alive	Impermissible	Impermissible
a_2 (=jump)	1 death	Dead	Permissible, but not required; supererogatory	Permissible, but not required; supererogatory
a_3 (=throw sandbag)	2 deaths	Alive	Permissible, but not required; not supererogatory	Permissible, but not required; not supererogatory
a_4 (=push Smith)	3 deaths	Alive	Impermissible	Permissible, but not required; not supererogatory

she is not only a comparative satisficing consequentialist but also a constrained consequentialist. As we saw in the previous section, constrained consequentialists insist that the intrinsic nature of an act has to be taken into consideration when evaluating consequences. Now, a_4 undoubtedly involves a *murder*. Since murders are arguably intrinsically bad, a comparative satisficing constrained consequentialist might claim that the goodness of a_4 is reduced to a level that may no longer count as good enough. There seems, however, to be an alternative scenario which rules out this reply and speeds up the process. Suppose we had not introduced $Case_1$ to follow up on $Case_0$. Suppose, instead, that we had used the following case.

Case$_{1*}$
The facts of the case are almost as they were in $Case_0$. However, Jones has an additional option for acting. He can, fourthly, hit a switch. If he does that, the trolley will turn left at the next turnout and onto a track that runs up a slight slope. At the end of this track there is also a massive rock that the trolley will hit. But the slope will slow it down such that, on impact, only the two persons who are sitting closest to the front of the trolley will die.

In $Case_{1*}$, Jones has the same options that he had in $Case_0$, viz. a_1, a_2, and a_3. That is, he can do nothing, jump, or throw the sandbag, respectively. The consequences of these options are also identical to their consequences in $Case_0$. If Jones does nothing, the ten people in the trolley die. If he jumps, only he dies, saving the ten. And if he throws the sandbag, the three people who are sitting closest to the front end of the trolley die. In addition, Jones has a fourth option a_4, viz. to hit the switch. If he does that, only two out of the three people sitting in front of the trolley will die.

Now, what are our intuitive verdicts about $Case_{1*}$? I take it that it is still impermissible for Jones to do nothing. And it still seems to be permissible (and supererogatory) for Jones to jump. So far so good. But now comes the twist. While the only remaining alternative in $Case_0$ was a_3, in $Case_{1*}$ there is an additional option, a_4, whose existence pertains, intuitively, to the evaluation of a_3. By doing a_3, Jones would save seven out of ten people, viz. those seven who are sitting farthest towards the back of the trolley. By doing a_4, however, Jones could save one *additional* person. He could do this, furthermore, at no costs to himself and without doing any harm to anybody else. This comparison between a_3 and a_4 seems to rule out that a_3 can be permissible. It leaves a_4 as the only other permissible option in $Case_{1*}$.

Can the comparative satisficing consequentialist match these provisional fixed points? She can obviously match those about a_1 and a_2 since Commitments 2 and 3 imply, of course, that a_1 is impermissible and that a_2 is permissible (as well as supererogatory). She can also match our provisional fixed point about a_4. Commitment 3 says that in a case where 11 lives are at stake, and the agent can save, at best, 10 out of these 11 lives, an act that saves eight lives may count as good enough and is, therefore, classed as permissible. a_4 saves even more lives than that, viz. nine. It should, hence, be permissible by Commitment 3. When it comes to the evaluation of a_3, however, a problem arises for the comparative satisficing consequentialist. As we just established, she should judge that it is impermissible. However, Commitment 3 implies that it is permissible. It saves eight lives which,

Table 5.6 The comparative satisficing analysis of Case$_{1*}$

Options	Consequences (overall)	Consequences (for Jones)	Provisional fixed points	Comparative satisficing consequentialism (+ Commitments 2 and 3)
a_1 (=do nothing)	10 deaths	Alive	Impermissible	Impermissible
a_2 (=jump)	1 death	Dead	Permissible, but not required; supererogatory	Permissible, but not required; supererogatory
a_3 (=throw sandbag)	3 deaths	Alive	Impermissible	Permissible, but not required; not supererogatory
a_4 (=hit switch)	2 deaths	Alive	Permissible, but not required; not supererogatory	Permissible, but not required; not supererogatory

by Commitment 3, should be good enough and, thus, permissible in a case where 11 lives are at stake, and the best act saves ten lives. Table 5.6 summarizes these findings about *Case*$_{1*}$.

Having established this result, we can, hence, formulate a further criticism.

Objection$_{1*}$
No comparative satisficing consequentialist moral theory can match our intuitive responses about *Case*$_{1*}$.

Is there anything the comparative satisficing consequentialist can say in reply to Objection$_{1*}$? It seems that there is not. She has already committed herself to Slote's Comparative Satisficing and, in particular, to Commitment 2 and 3. This leaves her the option to look for alternative components to Summation, Equal Treatment, and Welfarism. Let us start with the first possibility.

It can be ruled out, it seems, that alternatives to Summation are helpful. To see this, we do not have to go through all options one by one since there appears to be a principled reason for this. To match the provisional fixed points about the previous cases, the comparative satisficing consequentialist would have to explain why a_3 is good enough in *Case*$_0$, but not good enough in *Case*$_{1*}$. Given Comparative Satisficing, whether or not a_3 is good enough (in any given case) depends, firstly, on its value and, secondly, on the value of the best act available. Since the best act, a_2, is available both in *Case*$_0$ and in *Case*$_{1*}$, the question whether a_3 is or is not good enough in *Case*$_0$ and *Case*$_{1*}$ turns, hence, only on *its* own value. To explain, then, that a_3 is good enough in *Case*$_0$ but not in *Case*$_{1*}$, the comparative satisficing consequentialist would have to choose a function which attaches a greater value to a_3 in *Case*$_0$ than in *Case*$_{1*}$. Since a_3, however, is the same act and has the same consequences in both cases, this function would have to attach two different values to one and the same input value. No function can do that.

It is not clear, furthermore, how any version of Unequal Treatment might help the comparative satisficing consequentialist out of the impasse. To begin with, there is no plausible basis for a moral discrimination between the individuals involved in $Case_0$ and $Case_{1*}$. From Jones's perspective, all these persons are indistinguishable from the moral point of view, such that anything other than perfect impartiality on his part would seem morally arbitrary and unjustifiable. Above that, however, there is no possible weighting method that would explain why, in $Case_0$, a_3 counts as good enough, while in $Case_{1*}$ it does not. In both cases, a_3 affects the *same* individuals in the *same* way. No matter how we weigh its impact on their respective welfares, its total value should turn out the same in both cases.

Finally, it is not clear how embracing a consequentialist theory that incorporates constraints would help the comparative satisficing consequentialist to explain why a_3 should be seen as good enough in $Case_0$ and not $Case_{1*}$. First of all, it is hard to see which kind of constraint a_3 is supposed to violate. Secondly, to explain why a_3 is good enough in the one case but not in the other, the comparative satisficing consequentialist would have to argue that it violates a constraint in $Case_{1*}$, but not $Case_0$. Since a_3 has the same description and the same consequences in both cases, it is not clear how this would be possible.

As it seems, then, the comparative satisficing consequentialist runs into a dead end. There appears to be no way for her to rebut Objection_{1*}.

5.2.3 Hurka's Maxificing

The second variant of Non-Maximization that we found worthy of further scrutiny was the version of Maxificing proposed by Tom Hurka. Before we examine it, we should, first of all, remind ourselves what it says.

As we have worked out in Sect. 4.2.2.2, Hurka's Maxificing involves a differentiation between two types of cases, viz. *A*-cases and *B*-cases. An *A*-case is one where at least the morally best act alternative achieves a given absolute standard of goodness. A *B*-case is one where, no matter what the agent does, she cannot achieve the that standard of goodness. Drawing on this distinction, the criterion of rightness of Hurka's Maxificing Consequentialism says that an act is right if and only if the following holds: (i) If the agent faces an *A*-case, her act produces enough goodness, as defined by the respective absolute standard. (ii) If the agent faces a *B*-case, her act maximizes the good.

Now, what can a proponent of Hurka's idea (a maxificing consequentialist, henceforth) say to rebut Objection_0? Obviously, since her view is, as it were, partly maximizing, she has to ensure that she does not run into the same problem as advocates of plain-vanilla maximizing doctrines, such as CU. On any maxificing view of the type proposed by Hurka, Jones is required to maximize the good in $Case_0$, if that case is interpreted as a *B*-case. In a *B*-case, Maxificing Consequentialism implies, after all, the same verdicts as does CU. To ensure that she overcomes CU's problem

in $Case_0$, the maxificing consequentialist has to insist, therefore, that $Case_0$ is an A-case rather than a B-case.

Commitment 4
$Case_0$ is an A-case.

For the sake of argument, then, let us grant this assumption and ask whether, given Commitment 4, it is possible for a maxificing consequentialist to match our intuitive verdicts regarding Jones's options a_1, a_2, and a_3 in $Case_0$. It is clear that, if $Case_0$ is an A-case, a_1 comes out as morally wrong if and only if the state of affairs associated with this act does not reach the absolute standard of goodness that is a requirement for rightness. Hence, the maxificing consequentialist can explain why, on her view, a_1 is morally wrong only if she assumes that, if ten people die in $Case_0$, the goodness in the world falls below the morally tolerable level. She can also argue that a_2 and a_3 are both morally right. To this end, she has to assume that the axiological value of the respective states of affairs associated with these acts is above the threshold of goodness that is required for rightness.[8] In addition, she can explain the supererogatory status of a_2.[9] We can conclude, then, that the maxificing consequentialist can rebut Objection_0. She can do so, however, only on the condition that she endorses the following commitment.

Commitment 5
In $Case_0$, the absolute level of goodness associated with a_1 is below the threshold that is required for rightness, while the absolute level of goodness associated with a_2 and a_3 is above that threshold.[10]

Before we go to the lengths of formulating a follow-up objection, let us reflect on the maxificer's reply and the two commitments that she needs to make to rebut Objection_0. It is indeed striking that both of her assumptions appear to be rather *ad hoc*. Her verdict that a_1 is morally impermissible rests on the premise that the level of goodness associated with this act lies below the morally tolerable minimum. The verdict that a_3 is morally permissible rests on the assumption that the level of goodness that is associated with that deed is above the threshold. Now, remember that the difference in goodness between the outcomes of these two alternatives is equivalent in value to precisely seven lives. The maxificer has to assume, therefore, that, in $Case_0$, the world contains a very specific amount of goodness. She has to assume that it is just above the critical threshold, such that, if ten people die, the level of goodness gets intolerable. However, if only three people die, that level is still tolerable. This assumption is highly dubious and may certainly be questioned. But we need not even go that far. To prevent the maxificing consequentialist's reply from getting off the ground, we can once again augment $Case_0$ with a new case.

[8] The assumption that $Case_0$ is an A-case already implies, of course, that a_2 is above the threshold. An A-case is defined, after all, as a case in which the best act achieves the threshold.

[9] The reasoning is exactly analogous to the reasoning of the comparative satisficing consequentialist on page 192.

[10] Note that Commitment 4 is, strictly speaking, redundant as it is implied by Commitment 5.

Case$_{0***}$
The facts of the case are almost as they were in *Case*$_0$, except that, prior to *Case*$_{0***}$ happening, ten people (whose lives had the same subjective value as the lives of the people involved in *Case*$_{0***}$) have just died in a terrible trolley car accident somewhere in the world.

The difference between *Case*$_0$ and *Case*$_{0***}$, I presume, does not change our intuitive verdicts. That is, in *Case*$_{0***}$, a_1 should still be impermissible, while a_2 and a_3 should still be permissible. CU, however, implies that only a_2 is permissible in *Case*$_{0***}$. Therefore, we can formulate a new criticism of consequentialism which generalizes our finding that CU conflicts with our considered moral judgements about both *Case*$_0$ and *Case*$_{0***}$.

Objection$_{0***}$
No consequentialist moral theory can match our intuitive responses about *Case*$_0$ and *Case*$_{0***}$.

As Table 5.7 shows, the maxificing consequentialist cannot rebut Objection$_{0***}$. Given her commitments about *Case*$_0$, she has to conclude that in *Case*$_{0***}$ a_3 is impermissible, which conflicts with our intuitive moral judgement about that act.

Here is why. To solve *Case*$_0$, the maxificing consequentialist has to accept Commitments 4 and 5. That is, she has to assume that *Case*$_0$ is an *A*-case. And she has to stipulate that the goodness of a_1 is below the morally tolerable minimum, while a_2 and a_3 are above it. Now, the absolute difference in goodness between a_2 and a_1 in *Case*$_0$ is equivalent to the value of ten persons' lives. Since we assume that, prior to the occurrence of *Case*$_{0***}$, ten people have just died, the absolute moral value that is associated with the outcome of a_2 in *Case*$_{0***}$ must be equal to the absolute moral value that is associated with the outcome of a_1 in *Case*$_0$. Since the maxificing consequentialist must assume, in keeping with Commitment 5, that, in *Case*$_0$, a_1 is below the threshold, she must judge that a_2 is also below the threshold in *Case*$_{0***}$. That, in turn, means that she must interpret *Case*$_{0***}$ as a *B*-case. Now, according to her moral view, the only act that is permissible in a *B*-case is the maximizing act. Since a_2 is the uniquely maximizing act in *Case*$_{0***}$, it is the only permissible act. In other words, her view implies that a_3 is morally wrong, which contradicts our intuitive judgements. As it turns out, then, the maxificing consequentialist can solve *Case*$_0$ only if she makes assumptions that lead

Table 5.7 The maxificing analysis of Case$_{0***}$

Options	Consequences (overall)	Consequences (for Jones)	Provisional fixed points	Maxificing consequentialism (+Commitments 4 and 5)
a_1 (=do nothing)	10 deaths	Alive	Impermissible	Impermissible
a_2 (=jump)	1 death	Dead	Permissible, but not required; supererogatory	Required
a_3 (=throw sandbag)	3 deaths	Alive	Permissible	Impermissible

to a counter-intuitive result in $Case_{0***}$. She cannot solve both cases simultaneously. Hurka's Maxificing is unsuited, therefore, to overcome $Objection_{0***}$. So it, too, is off the table.

5.2.4 Alternative Welfarist Conceptions of the Good

Let us check, then, whether there are further possible replies that might allow consequentialists to dodge $Case_0$. CU also comprises, as we know, the component Summation. It purports that the relative goodness of two options, a_i and a_j, is determined by the sum total of the appropriately weighted well-being of all morally relevant subjects. In Sect. 4.2.3.2, we considered alternatives to this component. These were Multiplication, Leximin, Plural Egalitarianism, and Coarse-Grained Utilitarianism (CGU). First up, let us consider what a proponent of Multiplication can say to rebut $Objection_0$.

The first thing that we should note about Multiplication is that it is not clear how it is supposed to work since the numbers that are attached to the welfare of individuals are somewhat arbitrary. It is clear, however, that the (future) life of every individual whose life is at stake in $Case_0$ has, by assumption, the same value. And it should be a finite and positive value x. It is not clear which value should be attached to a life lost. There are three possibilities, though.

- We can stipulate, firstly, that the value of a life lost carries some non-numerical value. In that case, it is not clear how we could conceivably multiply.
- We may assume, secondly, that the value of a life lost is 0. If we do that, the product of the values of the individuals' future lives is also 0 for all choice options in $Case_0$. After all, at least one person always dies. That would mean that in all these cases all of Jones's choice options are either right or wrong on a maximizing consequentialist view.

  Both possibilities would be radically counter-intuitive.

- The third option is to allow a lost life to carry a value that is still positive – i.e. in the interval between 0 and x.

Perhaps consequentialists could justify this assumption by the arbitrariness of the neutral point. In that case, however, Multiplication would yield the same axiological ranking as Summation. We can conclude, therefore, that dropping Summation in favour of Multiplication would not help the consequentialist to solve the problematic $Case_0$.

The next option we should consider is Leximin. It judges that alternatives are to be ranked based on the well-being of the least well-off individuals in the respective states of affairs. If two states tie in this comparison, Leximin judges that the one which is better for the second-worst-off individual is better, and so on. Now, let us represent the well-being of an individual with 0, if she dies in $Case_0$ and with 1, if she survives. The states of affairs, $c(a_1)$, $c(a_2)$, and $c(a_3)$ that are associated with acts a_1, a_2, and a_3, respectively, are then as follows:

$$c\,(a_1) = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1)$$
$$c\,(a_2) = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0)$$
$$c\,(a_3) = (0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1)$$

In this representation, the well-being levels of the ten persons in the trolley occupy the first ten places while Jones's welfare is represented by the last numerical value in the sequence. To apply Leximin, we have to permute the representation the numerical values so that they appear in an ascending order – i.e. from worst to best. Let $c(a_i)^P$ be such a transformation of $c(a_i)$.

$$c(a_1)^P = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1)$$
$$c(a_2)^P = (0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$$
$$c(a_3)^P = (0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1)$$

Leximin ranks a_2 over a_3 and a_1 and places a_3 above a_1, such that we get: $a_2 > a_3 > a_1$.[11] It implies, thus, the same ranking as Summation. Diverging to Leximin, therefore, does not solve the problematic $Case_0$. a_2 comes out as the only morally permissible act. a_3 which, intuitively, is also permissible comes out as morally forbidden. Leximin, hence, cannot help consequentialists to rebut Objection_0.

What about Plural Egalitarianism? As we stated it previously, Plural Egalitarianism judges the relative goodness of alternatives based on a plural index that takes into account both the relative equality (or inequality, as the case may be) of the respective distribution of well-being and its overall value. Various combinations are possible within Plural Egalitarianism. We can attach different weights to the two dimensions of evaluation. Moreover, we can choose different inequality measures. Above, we said nowhere near enough about this issue since this would have required that we enter into the complexities and intricacies of the theory of inequality measurement. This, we said, is a matter that is surely beyond the scope of this inquiry. I suspect, nevertheless, that no plural egalitarian variant of consequentialism can satisfactorily address the problems that are raised by $Case_0$.

Here is why. In order to match our intuitive convictions about $Case_0$, a plural egalitarian consequentialist would have to explain why a_2 and a_3 are *both* morally permissible. The only way to do this is to come up with an axiology which judges that a_3 is morally exactly as good as a_2. Otherwise, it would not imply that a_2 and a_3 are *both* morally permissible. Apparently though, the state of affairs associated with a_2 is better than the state of affairs that is associated with a_3 – both in terms of equality *and* regarding its overall well-being. The latter is obvious. The former can be justified as follows. Intuitively, the best states of affairs in terms of equality should lie at the opposite ends of a spectrum. In $Case_0$, they are the ones in which all 11 people survive and in which all 11 die. The equality-wise worst states are in

[11] Transitivity of the betterness relation implies that a_2 is better than a_1.

the middle of the spectrum. They are the ones in which five survive and six die and *vice versa*. The state in which one dies and ten survive which is associated with a_2 seems to be better, then, in terms of equality, than the state in which three die and eight survive, which is associated with a_3. The former is, after all, closer to the end of the spectrum than the latter. Hence, the state of affairs related to a_2 is better than the state that a_3 brings. It is better *in both respects*. It is, therefore, hard to see how a plural egalitarian consequentialist could come up with an axiology that can produce the right result. It is hard to see, that is, how she could come up with the ranking that is necessary to explain the intuitively correct verdicts regarding Jones's options in $Case_0$. Hence, we can conclude, I believe, that Plural Egalitarianism cannot help consequentialists to rebut Objection_0 as it cannot solve $Case_0$ in a satisfactory manner.

A further possible strategy for the consequentialist to rebut Objection_0 is to adopt a coarse-grained account of the good, as proposed by Vallentyne (2006). The idea is to resort to an axiology which makes only rough discriminations between the good that attaches to the various options for acting. How does this proposal fare?

To show that both a_2 and a_3 are morally permissible, the coarse-grained consequentialist has to establish that we cannot distinguish these acts from an axiological point of view. She can claim that this is in fact so if she assumes that they are associated with roughly the same level of goodness. At the same time though, the coarse-grained consequentialist has to argue that a_1 is morally impermissible. She can do this consistently if she assumes that the state of affairs that is connected with this act is much worse than the state of affairs related to a_2 and a_3 and that, therefore, a_1 is to be axiologically distinguished from and ranked below these acts. The latter acts, she can claim, are, therefore, morally permissible, while the former is morally impermissible.

So far so good. What about the status of a_2, however? Can the coarse-grained consequentialist also match our intuition that this act is supererogatory? It seems as though she can. To see this, let us go through the minimal requirements for supererogation, viz. the Permissibility Requirement, the Altruism Requirement, and the Self-Sacrifice Requirement. On the coarse-grained view, a_2 fulfils the Permissibility Requirement, as we just established. As for the second requirement, the coarse-grained consequentialist can argue as follows: If we ignore Jones's well-being in an axiological comparison between a_2 and a_3, the former is as valuable as an act which saves ten out of ten lives, whereas the latter is as valuable as an act that saves only seven out of ten. The coarse-grained consequentialist can consistently hold that this difference is significant. And she can insist that a_2 is indeed the most altruistic act. As for the third condition, the coarse-grained consequentialist can acknowledge that a_2 involves a great sacrifice. In choosing a_2 over a_3, Jones sacrifices, after all, his *own* life which, from his *personal* point of view, can be seen as a significant sacrifice. As it turns out, then, the coarse-grained consequentialist can account for our common-sense moral judgements about $Case_0$.

We can, however, easily formulate a follow-up objection to this reply. To this end, let us state a crucial commitment that the coarse-grained consequentialist makes in putting forward her defence against Objection_0. To argue that, on the coarse-grained

consequentialist view, both a_2 and a_3 are to be seen as morally permissible, she has to make the assumption that saving 10 out of 11 lives is roughly as good as saving 8 out of 11 lives.

Commitment 6
In a situation like $Case_0$ where 11 lives are at stake, saving 10 out of 11 lives is roughly as good as saving 8 out of 11 lives.

It is easy to see that Commitment 6 leads to very problematic conclusions in other cases. Consider, e.g., the following $Case_2$.

Case$_2$
Jones is standing on a footbridge over a railway. A runaway trolley carrying 11 people is hurtling down the tracks. Jones can tell that, if the trolley is not stopped, it will hit a massive rock at the end of the tracks. The impact will most certainly kill 3 out of those 11 people, viz. those who are sitting at the very front of the trolley. Jones has, however, one option that will avert this worst possible case. He can hit a switch. If he does that, the trolley will be diverted onto a side-track at the next turnout. At the end of the side-track, there is also a massive rock that the trolley will hit. But the side-track runs up a slight slope. This will slow the trolley down, such that only the one person at the very front of the trolley will be killed on impact.

Jones has two options in this case. He can choose a_1, that is, do nothing. Alternatively, he can do a_2, that is, hit the switch. If he does a_1, the three people sitting at the very front of the trolley will die. Or, to put it differently, 8 out of 11 people will be saved. If Jones does a_2, only one of those three individuals will die, viz. the person who is sitting at the very front of the trolley. Sadly, this individual will die either way – i.e. no matter what Jones does. So the crucial difference between a_1 and a_2 is that, by doing a_2, Jones can save an additional two persons. This, I believe, warrants the intuitive verdict that it is permissible for Jones to do a_2 but impermissible for him to choose a_1. In other words, Jones clearly ought to save the greater number in $Case_2$.[12]

Now, can the coarse-grained consequentialist match these intuitive verdicts? As Table 5.8 shows, given that she has made Commitment 6 in $Case_0$, it seems that she cannot. Here is why. On a maximizing view, at least one of the alternatives, a_1 or a_2, must be permissible since at least one of these alternatives must be maximally good. Furthermore, by Commitment 6, saving 10 out of 11 lives is roughly as good as saving 8 out of 11 lives. So a_1 is to be seen as roughly equally good as a_2. Therefore, a_2 is permissible if and only if a_1 is, too. Given Commitment 6, it follows, therefore,

[12]This would be accepted even by "number sceptics." Number sceptics (e.g. Taurek 1977) deny that it is, in general, morally obligatory to save the greater number of lives. E.g., in a case where I can save, say, the lives of persons 2, 3, ..., 11 or the life of person 1 it is not morally obligatory for me to do the former. This is a "conflict case," (Lübbe 2008, 71) where the interests of person 1 conflict with the interests of persons 2, 3, ..., 11. Number sceptics can accept, however, that it is morally obligatory to save the lives of persons 1, 2, ..., 10 rather than the lives of persons 1, 2, ..., 9 since there is no conflict involved (cf. Kamm 1993, 97, fn. 12; Lübbe 2008, 71). $Case_2$ is of the latter kind.

Table 5.8 The coarse-grained consequentialist analysis of $Case_2$

Options	Consequences (overall)	Consequences (for Jones)	Provisional fixed points	Coarse-grained consequentialism (+ Commitment 6)
a_1 (=do nothing)	3 deaths	Alive	Impermissible	Permissible
a_2 (=hit switch)	1 death	Alive	Permissible	Permissible

on the coarse-grained view, that both a_1 and a_2 are permissible which, intuitively, is incorrect. We can, hence, formulate the following $Objection_2$.

$Objection_2$
No coarse-grained consequentialist theory can match our intuitive responses about $Case_2$.

Can the coarse-grained consequentialist rebut $Objection_2$ by diverging to a further alternative component? To check this, we have to investigate coarse-grained theories which also diverge from Maximization, Equal Treatment, and Welfarism. Let us first look towards alternatives to Maximization.

It is hard to see how dropping Maximization in favour of Comparative Satisficing or Maxificing should help the coarse-grained consequentialist to rebut $Objection_2$. The coarse-grained consequentialist has the problem that her doctrine is too permissive. Both a_1 and a_2 come out as permissible in $Case_2$, although, intuitively, a_1 should be regarded as impermissible. Both Comparative Satisficing and Maxificing, however, tend to make a consequentialist moral theory more, rather than less, permissive. Hence, it is hard to see how these components can conceivably help the coarse-grained consequentialist to rebut $Objection_2$.

The second way in which the coarse-grained consequentialist might rebut $Objection_2$ is to look for alternatives to Equal Treatment. But I cannot see how this might help her either. For it is difficult to understand how a version of Unequal Treatment would explain why a_1 is impermissible, while a_2 is permissible. To show this, she would have to argue that, in fact, the value of a_1 is markedly below the value of a_2. By Commitment 6, however, these two options have roughly the same axiological value. Commitment 6 says, after all, that saving 10 out of 11 lives is roughly as good as saving 8 out of 11 lives. a_1 saves eight lives and a_2 saves ten. So it clearly follows, by Commitment 6, that they are roughly equally good. Hence, both are to be judged either as permissible or as impermissible. Since at least one must be permissible on a maximizing view, both must be permissible. The only way to block this implication is to claim that, given the special circumstances of $Case_2$, Commitment 6 does not apply. The coarse-grained consequentialist may claim that the lives of the two additional individuals who would die if Jones was to do a_1 as opposed to a_2 are more valuable than the lives of "normal" individuals. Hence, a_1 cannot be seen as roughly as good as a_2. Therefore, a_2 is morally permissible but a_1 is forbidden.

This move would, in fact, bring the implications of the coarse-grained consequentialist theory in line with our intuitions. But it is clearly a case of special pleading. There is simply no basis for such a differential weighting in $Case_2$. Sure enough, if the two people in question were, say, Jones's brothers, sisters, parents,

friends, or what have you, it may seem justified, from Jones's personal point of view, to make the requisite moral discriminations. As we discussed in Sect. 4.2.3.2, this could be done, e.g., by endorsing a version of Self-Referential Altruism. Note, however, that none of these facts obtain in $Case_2$. Therefore, the unequal weighting of the individuals' lives that the coarse-grained consequentialist needs to stipulate in order to bring her moral outlook in line with our intuitions is without foundation. We should conclude, therefore, that Unequal Treatment cannot help the coarse-grained consequentialist to rebut Objection_2.

The third option is to look towards constrained variants of coarse-grained consequentialism. But this does not seem to work either. To argue that there is a difference in goodness between a_1 and a_2, the coarse-grained constrained consequentialist would have to maintain that, in $Case_2$, there is a consequentialist constraint against doing a_1. More precisely, she would have to argue that a_1 is intrinsically bad and that the state of affairs which results from doing a_1 is, therefore, significantly worse than the state of affairs that results from a_2. a_1, however, consists in doing nothing. It is an *inaction* (or "negative act"). In Sect. 4.2.4.2, we said that, in all likelihood, a constraint generates a negative duty, that is, a duty to refrain from an action. In this case, though, the constraint would create a positive duty, that is, a duty to act, which is very implausible. Therefore, going constrained does not seem to help the coarse-grained consequentialist to rebut Objection_2 either. Hence, let us conclude that consequentialists who endorse a coarse-grained view run into a dead end.

Consequentialists may have one last ace up their sleeves. It might be possible for them to dodge $Case_0$ by diverging to an alternative to Equal Treatment. Equal Treatment is the view that the well-being of every morally relevant individual is to be given equal weight in working out what to do. Consequentialists, however, can adopt, in place of this component, a plausible version of Unequal Treatment, such as Prioritarianism or Self-Referential Altruism (or a combination of the two).

Now, does this give the consequentialist a way to reply to Objection_0? To rebut Objection_0, it is once again necessary for the consequentialist to argue that a_2 and a_3 are equally good because she must explain why they are both permissible. She can do that only if she gives Jones's well-being increased moral weight. To be precise, she would have to commit to the following claim.

Commitment 7
Jones's well-being is to be given three times as much weight as the well-being of the other individuals.

a_3 saves 8 out of 11 lives in $Case_0$. But it saves Jones's life, in particular. Hence, it is axiologically as valuable as an act that saves 10 out of 13 lives. a_2 saves 10 out of 11 lives. However, it costs Jones his life which, by Commitment 7, is worth three times as much as a 'normal' life. So it is also axiologically as valuable as an act which saves 10 out of 13 lives. Hence, a_2 and a_3 come out as equally good. Since they are both better than the only alternative, a_1, they are both maximally good and, therefore, morally permissible, while a_1 is not. Under this reconstruction of the case, a_2 also comes out as supererogatory. As we just established, a_2 is morally permissible. It is also self-sacrificial since there is a permissible option, a_3, that

Table 5.9 The consequentialist analysis of $\text{Case}_{0^{**}}$ with unequal treatment

Options	Consequences (overall)	Consequences (for Jones)	Provisional fixed points	Consequentialism (+ Commitment 7)
a_1 (=do nothing)	10 deaths	Alive	Impermissible	Impermissible
a_2 (=jump)	1 death	Dead	Permissible	Impermissible
a_3 (=throw sandbag)	2 deaths	Alive	Permissible	Permissible

does not cost Jones his life. It is, furthermore, altruistic. It appears, then, that a consequentialist who swaps the Equal Treatment contained in CU for a version of Unequal Treatment can answer the challenge posed by *Case*$_0$.

At this point, we can make use of the same tactic that we employed previously. We can augment *Case*$_0$ with another case. In fact, we have already formulated one which can serve this purpose, viz. *Case*$_{0^{**}}$ that we introduced on page 190. In this case, recall, Jones has three options, viz. a_1, a_2, and a_3. That is, he can do nothing, jump, or throw the sandbag, respectively, as Table 5.9 shows.

As we have already noted above, it seems obvious that it is morally forbidden, in *Case*$_{0^{**}}$, to do a_1. If, after all, Jones chooses a_1, all persons in the trolley will die. He can avoid this by doing a_3, that is, throw the sandbag. This will save eight out of the ten persons at no costs to anyone. It seems equally clear that no reasonable moral theory should require Jones to do a_2, that is, jump off the bridge. Also, no reasonable moral theory should hold that it is impermissible for Jones to jump. For as we have mentioned a number of times, moral agents seem to have an agent-sacrificing option. Jones may jump, then, if he wishes to. This would be morally laudable for him to do. But it is not morally required. Finally, it appears to be surely permissible for Jones to do a_3, that is, throw the sandbag, though this seems not to be supererogatory.

As we already established above, CU fails both in *Case*$_0$ and in *Case*$_{0^{**}}$. This means that we can formulate an augmented Objection$_{0^{**}}$ against consequentialism, as we have done above. It claims that no consequentialist theory can match our intuitive moral responses to *Case*$_0$ and *Case*$_{0^{**}}$ simultaneously. Now, it turns out that consequentialists who subscribe to Commitment 7 cannot rebut Objection$_{0^{**}}$ because they cannot match our intuitive moral verdicts in *Case*$_{0^{**}}$. By Commitment 7, Jones's life is three times as valuable as the lives of each of the ten. Therefore, a_1 has the same value as an act that saves 3 out of 13 persons. a_2 is as valuable as an act that saves 10 out of 13 people. Finally, a_3 is as valuable as an alternative that saves 11 out of 13 lives. By Commitment 7, a_3 is the maximizing act, which means that, on a maximizing consequentialist view, it is the only permissible option, while a_1 and a_2 come out as morally forbidden. Intuitively, however, a_2 is also morally permissible. In fact, it appears especially laudable. By doing a_2, Jones after all voluntarily sacrifices his life to save two fellow humans. Let us conclude, then, that there appears to be no version of Unequal Treatment that can help consequentialists to rebut Objection$_{0^{**}}$. We can hence put this alternative aside.

In conclusion, we can record that no consequentialist moral theory appears to be able to give an ultimately satisfactory reply to all of the various objections – Objection_0, Objection_{0*}, Objection_{0**} and Objection_{0***} – that we have levelled against consequentialism. I would tentatively suggest, therefore, that we should reject the whole family of consequentialist moral theories.

5.3 Summary

In this chapter, we tied up the loose ends of our case against consequentialism. In keeping with step (iv) of our methodic procedure, FRA_2, we made a case against consequentialism from the inside out. To this end, we introduced an initial case, $Case_0$, in Sect. 5.1. It was intended to show that CU is an untenable doctrine. $Case_0$ involved a story about a moral agent, Jones, who is standing on a footbridge over a railway, as a trolley carrying ten people approaches. He sees that, if the trolley is not stopped, it will collide with a massive rock, and he can tell that this will kill all of the ten people in it. In this situation, we assumed, Jones has three options. He can do nothing. He can jump down onto the tracks. And he can throw a sandbag. If he does nothing, all ten persons in the trolley will die, as predicted. If he jumps onto the tracks, his body will stop the trolley and save all ten individuals. But Jones himself will be squashed and killed. If he throws the sandbag, seven out of the ten trolley passengers will survive. Intuitively, we said, it is impermissible for Jones to do nothing because by throwing the sandbag he can save seven people at no costs to anyone. He should, therefore, at least throw the sandbag. It is also permissible for Jones to jump. In fact, this option seems to be supererogatory, as it is self-sacrificial and altruistic.

We established that the implications of CU conflict with these intuitive judgements. CU judges that the only permissible option for Jones is to jump while it regards the other two options as impermissible. We concluded, therefore, that CU violates our provisional fixed points about $Case_0$ and generalized this finding by putting forward Objection_0. Objection_0 is the claim that no consequentialist theory can match our provisional fixed points about $Case_0$. In the course of the chapter, we augmented this charge with three further criticisms, viz. Objection_{0*}, Objection_{0**}, and Objection_{0***}, which draw on additional cases that are also effective objections against CU.

In Sect. 5.2, we proceeded by analysing replies that consequentialists may give to rebut Objection_0. As we had established earlier, consequentialists can diverge from some of the CU-components to alternative components. The first possibility that we considered was that consequentialists might endorse a theory that factors in moral constraints. We analysed what a consequentialist of this stripe – a constrained consequentialist, as we called her – would have to say in order to show that Objection_0 was, in fact, mistaken. It turned out that the only way for a constrained consequentialist to deny Objection_0 is to claim that it is intrinsically bad for Jones to sacrifice his life for the sake of three other people. In and of itself, we noted, this

idea is very implausible. However, even if we grant this assumption, the reply of the constrained consequentialist does not seem satisfactory. To show this, we augmented our initial case against CU, $Case_0$, with a further case, $Case_{0*}$, in which we slightly alter the number of deaths. Then, we formulated a new charge, $Objection_{0*}$, which says that it is not possible for consequentialists to match our intuitive responses to $Case_0$ and $Case_{0*}$ simultaneously. As we showed, it is not possible for a constrained consequentialist to rebut this new $Objection_{0*}$.

In a next step, we considered non-maximizing forms of consequentialism, starting with the comparative satisficing version proposed by Slote (1984). It turned out that it was, indeed, possible for a comparative satisficing consequentialist to rebut $Objection_0$. But it became obvious that her solution was short-lived. We were able to show that the comparative satisficing consequentialist can solve $Case_0$ only if she makes some specific commitments. Using a further case, $Case_1$, we were able to show that these commitments force her to accept the verdict that it is morally okay for Jones to kill an innocent bystander, Smith, for no reason at all. We went on, therefore, to formulate a further $Objection_1$. It purports that no comparative satisficing consequentialist theory can match our intuitive responses to $Case_1$. In order to block all possible replies, we introduced a modified version of $Case_1$, $Case_{1*}$, and formulated a further $Objection_{1*}$.

After that, we considered a conceivable response to $Objection_0$ based on Tom Hurka's idea of Maxificing. Maxificing is the view that an act is permissible if and only if it is either the best the agent can do in a given situation or is good enough as judged by a certain absolute standard. To prove $Objection_0$ wrong, the maxificing consequentialist has to show that, on her view, a_1 is impermissible while the maximizing act, a_2, and a_3 are both permissible. As we found, she can draw this conclusion only if she assumes that both a_2 and a_3 are above the critical threshold while a_1 is under that threshold. We noted that these stipulations seemed totally *ad hoc*. Furthermore, we argued that it is possible to invoke another case, $Case_{0***}$, which is identical to $Case_0$ except in one respect. We assume that, in $Case_{0***}$, the level of goodness in the world has dropped, such that the absolute goodness associated with a_1, a_2, and a_3 is below the critical threshold. In this case, we observed, our considered intuitive verdicts remain the same. That is, we still judge that a_2 and a_3 are both permissible, while a_1 is not. We were able to show, however, that the maxificing consequentialist has to judge that only a_2 is permissible in $Case_{0***}$, given the commitments she has made in $Case_0$. We recognized that she is, hence, not able to rebut $Objection_{0***}$, viz. the claim that no consequentialist theory can match our intuitive moral verdicts about $Case_0$ and $Case_{0***}$ simultaneously.

Finally, we considered the possibility that an alternative welfarist conception of the good may help consequentialists to dispel $Objection_0$. In particular, we looked towards alternatives to the CU-components Summation and Equal Treatment, starting with the former. We quickly realized that diverging to Leximin does not help consequentialists. For, as it turned out, a Leximin-based consequentialist theory would imply the same problematic verdicts as CU in $Case_0$. Much the same can be said regarding the component Multiplication. Plural Egalitarianism, we came to believe, does not solve the problem either. It is based on the idea that an act

is better than another if its outcome provides a better overall mixture of equity and aggregate well-being. Since a_2 seems to fare better than a_3 both regarding equity and in terms of overall welfare, it was hard to see how a plural egalitarian consequentialist might justify the idea that these two acts were equally good. On the plural egalitarian consequentialist view, the latter, however, is a necessary condition for the permissibility of a_3. We found, furthermore, that diverging to a coarse-grained form of consequentialism did not provide an ultimately satisfactory answer to $Case_0$ either. Though we found that consequentialists may be able to rebut Objection_0 by endorsing a coarse-grained view, it became apparent that the commitments they have to take on to solve $Case_0$ seemed to imply absurd judgements in other cases. One such case was Case_2 in which the agent, Jones, has a choice between a_1 and a_2. By doing a_1, Jones would save 8 out of 11 people. By doing a_2, he could, at no costs or risk to himself, save the same eight persons *plus* an additional two individuals. As we noted, it seems clear that, in this situation, a_1 is impermissible while only a_2 is permissible. Given her commitments in $Case_0$, the coarse-grained consequentialist has to judge, however, that both options, a_1 and a_2, are morally permissible. The last possible reply to Objection_0 that we examined had the consequentialist endorse an alternative to Equal Treatment. We found that adopting a particular version of Unequal Treatment does, in fact, help consequentialists to overcome the complication that CU runs into in $Case_0$. To match our intuitive judgements about that case, they have to assume that Jones's life carries triple weight. As we found, however, this move has the unwelcome consequence that consequentialists commit themselves to a very counter-intuitive view in other cases. To show this, we looked at a modified version of $Case_0$, viz. $Case_{0**}$. In that case, the idea that Jones's life is three times as valuable as the lives of other persons implies that it is morally forbidden for Jones to sacrifice himself for the sake of others. This is counter-intuitive because, far from being morally forbidden, such an act strikes us as especially laudable.

This, then, concluded step (iv) of our investigation. As a number of philosophers before us, we attempted to show that there is good reason to reject the whole family of consequentialist moral theories. The argument that we gave provides, I hope, further support for that conclusion.

Chapter 6
Conclusion

In this book, we covered a lot of ground. We looked at an important issue in moral philosophy that has been a point of contention for quite a long time: the question whether there is a form of consequentialism that stands up to the objections that its critics have levelled against it as a moral view. We tried to make progress on this issue by devising a new methodological approach – the Family Resemblance Approach (FRA) – which is immune to a problem that looms large in the debate between consequentialists and their critics, viz. the Humpty Dumpty Defence (HDD). In implementing our new approach, we then constructed a comprehensive case against consequentialism that we built with the aid of our predecessors on whose shoulders we stood.

Now is the time to review what we said and to bring into clear view the various components of our argument. However, we should do more than merely sum up the steps of our reasoning. As I said in the introduction, intellectual honesty commands that we make clear which assumptions we made when we pieced together our case against consequentialism, how we might have erred in doing so and how this would affect the conclusion we drew. Of course, we cannot recapitulate all assumptions critics may doubt. But we can, at least, go through the most obvious points. I should, perhaps, emphasize that there lies no embarrassment in conceding that we used some controversial suppositions. Every philosophical piece has to take certain assumptions for granted, and some of them are bound to be doubted at least by some philosophers. It would, however, violate the philosophical ethos – at least as I understand it – if we pretended to stand on absolutely solid ground when, in fact, we do not. With this in mind, let us conclude with a few reflections on our argument.

In Chap. 1, we began our discussion with a brief investigation of its normative-ethical foundations. In Sect. 1.1, we started by addressing the idea of normative ethics per se, which we defined as the study of moral theories. This rendition of the concept seems rather innocuous. However, when we proceeded to explain what a moral theory is, we did step on controversial territory. In Sect. 1.2, we analysed this notion as containing a theoretical and a practical component where the former was

N. Mukerji, *The Case Against Consequentialism Reconsidered*,
Theory and Decision Library A: 51, DOI 10.1007/978-3-319-39249-3_6

taken to be a device that attaches a particular *moral status* to the *acts* available in a given *choice situation* to a *moral agent* as a function of certain *normative factors*. Not only is this statement a mouthful. Many moral philosophers – and, in particular, those of the consequentialist stripe – might take issue with it.

First of all, they might deny that acts are to be seen as possessing a definite moral status – say, *right* or *wrong* – or that deontic status is an all-or-nothing affair.[1] Norcross (2006), in particular, argues that "Scalar Consequentialism," which does not assign a definite moral status to the acts of a moral agent, might avoid many of the flaws of conventional consequentialist theories. And Peterson (2012) suggests that "Multi-Dimensional Consequentialism," which holds that rightness and wrongness are non-binary concepts might be a considerable improvement over traditional consequentialist theories. Since our argument relied on the assumption that moral theories do assign a definite, discrete moral status to acts, it does not cover views like the ones proposed by Norcross and Peterson. We would have to say more than we did above to show that these forms should also be rejected.

The second point at issue is our stipulation that moral theories are devices that evaluate *acts*. Many ethicists believe that they can have different "primary evaluative focal points," as Kagan (1992, 239) puts it. And consequentialists, in particular, have noted that their moral theories may cover not just acts, but also "desires, dispositions, beliefs, emotions, the colour of our eyes, the climate and everything else." (Parfit 1986, 25) I believe that a good case can be made for the limitation that we imposed on our discussion (see, e.g., Howard-Snyder 1993). But it must be conceded, nevertheless, that we did not explicitly address this issue. Perhaps theories like the Rule Consequentialism proposed by Hooker (2003) or the Virtue Consequentialism held by Driver (2001/2003) should also be considered as genuine forms of consequentialism. In that case, our argument would be unsuccessful unless we could show that it can be extended to these types, too.

Thirdly, in the context of our discussion of the idea of a choice situation, we assumed that acts are not individuated by their consequences, as was suggested by Savage (1954/1972, 14). After all, this would, in effect, terminate the discussion. If Savage's view were correct, all non-consequentialist theories would violate classical deontic logic. They allow, after all, that two acts that have the same consequences may have different moral *stati*. On Savage's criterion of act identity, this, of course, would not make sense since it would mean that *one and the same* act can have different moral *stati*. Now, it may turn out that there are, in fact, persuasive metaphysical arguments that support Savage's idea. I, for one, do not believe that this is very likely. But if it were, all non-consequentialist theories would have to be ruled out as non-sensical, and consequentialism would have a walk-over.

Fourthly, the test cases that we looked at in our argument only concerned the implications of consequentialist moral theories for a particular type of agent, viz. individual human actors. These cases show, I believe, that every form of consequentialism violates certain provisional moral fixed points. But consequentialism may

[1] Critical views about moral status can be found in various places in the literature. For a brief overview, see footnote 10 in Chap. 1.

still carry the day if its proponents could demonstrate that it comes out as the most plausible view once all types of agents are taken into account. Mendola (2005a), e.g., has suggested that there may be an argument for consequentialism which is based on the idea of "group agents."[2]

Finally, there might be a problem for us that has to do with the nature of normative factors. It may turn out that all morally relevant facts about an act are most aptly seen as consequences. Now, of course, this is highly doubtful. However, if it were true, it would mean that the best moral theory is consequentialist because *every* moral theory is consequentialist. We did not address this possibility.[3]

In Chap. 2, we went into the metaethical foundations of our argument. We investigated the question how moral theories can be evaluated and considered a particular moral-epistemological conception that we called the "Rawlsian Approach." The rough idea behind it is, as we discussed in Sect. 2.1, that moral theories should be evaluated according to their fit with our moral sense or, to be more precise, according to their fit with the intuitive moral verdicts that issue from this moral sense. We assumed that some version of this idea was correct. And we did so, as we should concede, without argument. Upon further investigation, it may turn out, of course, that this idea is completely mistaken. In this case, our argument, too, would collapse! This point is important to note because there are versions of consequentialism that use a different justificatory method. As we noted in Sect. 2.1, Hare (1981), e.g., rejects the Rawlsian Approach. Instead, he justifies his version of consequentialism, which is a form of preference utilitarianism, through an analysis of the meaning of moral verdicts. Given that this particular consequentialist theory is wedded to a different justificatory approach, we should add that the argument we developed above begs the question against Hare's theory as well as related versions of consequentialism.

This said, we should add that our case against consequentialism does not only rely on the assumption that *some* version of the Rawlsian Approach is correct. It relies, rather, on the idea that a *particular* interpretation of it is correct, viz. one that allows moral intuitions to play an important role in theory evaluation. We factorized the Rawlsian Approach into three evaluative sub-criteria which we referred to as intuitive fit, consistency, and systematicity. And we observed that the Rawlsian Approach, in effect, recommends that we choose the moral theory which fulfils these criteria best overall. It is evident that consistency is a knock-out criterion. However, there may be various possible trade-offs between systematicity and intuitive fit. We assumed an interpretation of the Rawlsian Approach that allows intuitive considerations to play an important role. But perhaps there are convincing

[2]On the idea of group agency, see, also, the recent work done by List and Pettit (2011).

[3]It should be noted, however, that this would be a very radical claim. Some moral theorists have, in fact, claimed that all moral theories can be "consequentialized." (e.g. Portmore 2007) They claim that for every non-consequentialist moral theory there is a consequentialist analogue which implies the exact same moral verdicts. The claim that all moral theories are, in fact, forms of consequentialism goes much farther than this and has not, as far as I know, been proposed by anyone in the moral-philosophical debate.

arguments to the effect that systematicity is more important than intuitive fit and that the latter should merely be accorded the role of a tie-breaker. In fact, if that were so, our argument would be untenable.

Suppose, however, that the Rawlsian Approach is, in fact, the most appropriate moral-epistemological view and that, on the best interpretation of it, intuitive fit is, in fact, assigned a significant weight. These assumptions are still not sufficient for our argument to go through. As we discussed in Sect. 2.2, there are various interpretations of the sub-criterion of intuitive fit. We relied on the supposition that its best interpretation requires our moral theories to match our low-level intuitions – at least to a certain extent. We provided arguments that were supposed to support this idea. But they did not, of course, establish it conclusively.

In Sect. 2.3, we discussed how our interpretation of the Rawlsian Approach can be applied to the evaluation of moral theories. We observed that it gave us only a philosophical ideal. That is, it only tells us that the best moral theory is the one that scores best by the criteria consistency, systematicity, and intuitive fit (appropriately weighted). But it does not provide a "pass-or-fail test" that we can apply to the evaluation of a moral theory. For this reason, we introduced the Provisional Fixed Point Approach (PFPA). It is based on the idea that there are certain intuitive moral verdicts – "provisional moral fixed points" – that seem extremely reliable, such that we may, for all intents and purposes, assume that any reasonable moral doctrine should match them. A theory which does not match them, we may conclude by PFPA, fails by the criterion of intuitive fit. In Chap. 5, we applied PFPA. We showed that no consequentialist moral theory can match our firmly held moral verdicts about a series of trolley cases. Based on that result, we rejected all forms of consequentialism. As we discussed in Sect. 2.3, however, we have to qualify this conclusion with the following *proviso*. We said that, if it should turn out to be impossible to match the respective provisional fixed points, we may have to reconsider our dismissal of consequentialism. PFPA presupposes, after all, that the best moral theory can, in fact, accommodate the moral verdicts that are taken to be provisional fixed points. If it turns out that no non-trivial moral theory[4] is capable of matching them, the criticism that consequentialism violates some of them would not be fair. To make the case against consequentialism ironclad, we would have to demonstrate that there is, in fact, a non-trivial moral theory that implies the moral fixed points in all the trolley cases that we considered above.[5] We, of course, did not provide such a demonstration.

[4]Note that it is always possible to devise a moral theory that trivially matches all provisional moral fixed points. This can be done by drawing up principles which simply record them on a case-by-case basis. Such a trivial moral theory, however, would certainly not be acceptable since it would not have anything to offer in the way of systematicity.

[5]Strictly speaking, even if it was possible to demonstrate this, there would still be a chance that the best moral theory is consequentialist. Here is why. In our case against consequentialism we looked only at a narrow range of provisional fixed points. It may be that all consequentialist moral theories violate some of them and that there is an alternative moral view that does not. It may, nevertheless, turn out that, on the whole, consequentialist doctrines can accommodate comparatively more (or more important) provisional fixed points than that alternative theory. In that case, they would, of course, be preferable to the latter.

Furthermore, our argument crucially depends on the premise that trolleyology and the specific use that we made of trolley cases is a justifiable method of moral inquiry. In Sect. 2.4, we looked at some objections to this idea, and we responded to them as best we could. We noted, in particular, that trolley cases can, broadly speaking, serve us in two ways. And we showed that the way we use them is, in fact, the less controversial of the two. This distinction helped us to respond to certain kinds of criticisms. A more fundamental worry about the use of highly stylized hypothetical cases in ethics may, however, remain. We certainly did not say nearly enough to establish, beyond a reasonable doubt, that this worry is without foundation.

In Chap. 3, we then addressed important methodological issues and made further assumptions that at least some moral philosophers may question. In Sect. 3.1, we looked at the procedure that is commonly used to criticize consequentialism, viz. the Definitional Method (DM). In the context of DM, we came to talk about the Core Idea of consequentialism. It is the idea that whether an act is right depends only on the goodness that it produces. We assumed, all throughout the discussion, that this notion is, in fact, shared by all consequentialist doctrines. And we discussed moral theories only to the extent that they were compatible with it. Some moral theorists will, no doubt, object to this. They may argue that, by making this assumption, we, in effect, limited our attention to what has been called Evaluative Consequentialism (cf. Sinnott-Armstrong 2011). They may insist, in other words, that we excluded all forms of Non-Evaluative Consequentialism .[6] Now, for all we know, some version of the latter view may, in fact, be able to overcome the challenges that we presented in Chap. 5, such that consequentialism is saved from the attacks that we launched. This criticism is, of course, correct. Our argument does not cover non-evaluative forms of consequentialism, such as the one proposed by Portmore (2011). Why did we not include them, then? We had, in fact, a twofold reason for this. Firstly, we had to make sure that we would not bite off more than we could chew. Ensuring the tractability of our discussion was key, and excluding Non-Evaluative Consequentialism served this purpose. Secondly, there seemed to be a good reason to be sceptical towards this particular strand of consequentialism. At first glance, it appears completely unmotivated to regard any non-evaluative theory as a form of consequentialism. But this impression may, of course, be mistaken. It may turn out that there is, in fact, a persuasive case for Non-Evaluative Consequentialism. In that case, we would have to say more than we did to demonstrate that consequentialism should be rejected.

In Sect. 3.2, recall, we introduced our methodical procedure – the second formulation of the Family Resemblance Approach (FRA_2). We would go on to use it to formulate a comprehensive argument against consequentialism in Chap. 5. In

[6]Non-evaluative forms of consequentialism do not evaluate the moral status of an act based on a single measure of goodness. Rather, they rely on two such measures. The state of affairs that results from an act is evaluated, firstly, in terms of its moral goodness. Secondly, it is assessed from the perspective of the agent's self-interest. Then, a non-evaluative, supplementary "value" is computed from the results of both evaluations. It, in turn, is taken to be the basis of moral evaluation. See, also, footnote 19 in Chap. 3, which briefly discusses non-evaluative forms of consequentialism.

step (i) of FRA_2 that we addressed in Sect. 4.1, we factorized a paradigmatic form of consequentialism, viz. CU, into its logical building blocks. In Sect. 4.2.1, we then took on step (ii) and eliminated certain theoretical distinctions. In particular, we put aside the distinction between subjective and objective forms of consequentialism, direct and indirect versions as well as the distinction between consequentialist theories with different theories of individual well-being. Our justification for doing this had to do with our use of trolley cases. We argued that these distinctions are irrelevant in trolley cases and concluded that we can, therefore, ignore them. Our final argument would, after all, depend entirely on the methodical use of these cases. Above, we said that our argument depends on the premise that trolleyology is a justifiable method of moral inquiry. This is true. But we should add to this that the dependency is, in fact, twofold. Our reasoning relies on the assumption that trolley cases are justifiable, firstly, because we actually used these cases in our argument and, secondly, because we eliminated various theoretical distinctions that may, in fact, be highly relevant outside the narrow confines of trolley cases. Should it turn out that trolleyology is unjustifiable, our dismissal of the distinctions might, of course, also be indefensible.

This said, let us also talk about step (iii) of FRA_2, which we tackled in Sect. 4.2.2. We took stock of the various alternatives to the CU-components. Since there are, innumerable logical possibilities, we had to limit our attention to a subset of them. As we discussed, it seemed to make sense to look only at those components that appeared to be motivated. We noted that it seems reasonable to suppose that most variants of consequentialism are motivated by certain problems that attach to CU. That is, the motivation to look for alternative components of consequentialist doctrines has commonly been an attempt to remedy faults of CU. It appeared sensible, therefore, to consider alternative components only to the extent that they could be interpreted as an improvement (of sorts) over CU. Our method for finding them consisted, hence, in going through significant criticisms of CU. We considered objections to that doctrine and examined which alternative components might help consequentialists to dodge them. In the process of doing that, we may, of course, have overlooked certain interesting possibilities. It may be that, once we add them to the list, our argument will not go through anymore.

Finally, let us consider step (iv) that we took on in Chap. 5. Stringing together a series of trolley cases, we made a case against consequentialism. As we discussed in Sect. 3.2.2, the only way to make sure that we can justifiably reject all consequentialist theories of morality is to show that all possible combinations of components give rise to a defective moral theory. As we noted, though, this task was unmanageable since there are simply too many possible combinations. Hence, we opted for a methodological shortcut. We made a case against consequentialism from the inside out. That is, we started by formulating an initial case, $Case_0$, which we used to demonstrate that CU is defective. Then, we put forward a charge, $Objection_0$, against consequentialism that generalized our finding. We claimed that no consequentialist theory can give us the right answer in $Case_0$. $Objection_0$ was, of course, false. But it was useful for methodological reasons. For it raised the question how consequentialists can defend themselves against it. We considered

their options. To this end, we consulted the construction kit for consequentialist doctrines that we introduced at the end of Sect. 4.3. It shows all the standard and non-standard components of consequentialist theories that consequentialists can embrace. We worked out which of these consequentialists can diverge to in order to rebut Objection$_0$, and we formulated further cases which were meant to demonstrate that the modified theories gave rise to new objections. We kept on going until we had shown that all of the answers that consequentialists might give to the initial case, *Case*$_0$, should ultimately be rejected.

As we already discussed in Sect. 3.2.2, this procedure is problematic for the following reason. With our inside-out argument we first attacked CU. Then, we looked for possible replies that consequentialists might give. But we did so only in the immediate vicinity of CU. That is, we would allow the consequentialist to modify only one component of CU, e.g. Maximization. The others were kept constant. Therefore, we cannot exclude the possibility that consequentialists might eventually find a combination of components which helps them to clear the hurdles that we set up for them. It was for this reason that we noted in Sect. 3.2.2 that the best we can achieve with our argument is to shift the burden of proof to the consequentialist.

This is the note on which, I believe, we should finish. Our discussion did, I hope, shed new light on the debate about consequentialism and contained new and interesting ideas as to how the case against this moral outlook should be made. In the end, however, we have to conclude that our argument comes with many ifs and buts. It depends on numerous assumptions that critics may justifiably question. And the procedure that we used is certainly not capable of delivering any final results. I think, however, that anyone who seriously attempts to defend a philosophical claim has to be prepared to arrive at such a conclusion. It lies in the nature of our discipline. It would be naïve to assume that we could ever settle a philosophical issue once and for all since "philosophical questions are," as Martin Heidegger said, "in principle never settled as if some day one could set them aside." (Heidegger 1953/2000, 44) The most we can ever do when we argue for or against an established philosophical position is to shift the burden of proof to our objectors. That, I hope, we did accomplish.

References

Adams, Robert M. 2002. *Finite and infinite goods: A framework for ethics*. New York: Oxford University Press.

Adorno, Theodor W. 2006. *Minima Moralia: Reflections on a Damaged Life*. Trans. Edmund F. N. Jephcott. London/New York: Verso Books.

Ainslie, George. 1975. Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin* 82(4): 463–496.

Ainslie, George. 1986/2000. Beyond microeconomics: Conflict among interests in a multiple self as a determinant of value. In *The multiple self*, Studies in rationality and social change, ed. Jon Elster, 133–176. Cambridge: Cambridge University Press.

Ainslie, George. 1999. The dangers of willpower. In *Getting hooked: Rationality and addiction*, ed. Jon Elster and Ole-Jorgen Skog, 65–92. Cambridge: Cambridge University Press.

Alexander, Larry, and Moore Michael. 2008. Deontological ethics. In *The Stanford encyclopedia of philosophy. Fall 2008 edition*, ed. Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information (Stanford University).

Allen, Harold J. 1967. A logical condition for the redescription of actions in terms of their consequences? *The Journal of Value Inquiry* 1(2): 132–134.

Anderson, J.R., and R. Milson. 1989. Human memory: An adaptive perspective. *Psychological Review* 96: 703–719.

Anscombe, Gertrude E.M. 1958. Modern moral philosophy. *Philosophy* 33(124): 1–19.

Appiah, Kwame A. 2008. *Experiments in ethics*. Cambridge: Harvard University Press.

Arneson, Richard. 2000. Perfectionism and politics. *Ethics* 111: 37–63.

Arneson, Richard. 2004. Moral limits on the demands of beneficence? In *The ethics of assistance: Morality and the distant needy*, ed. Deen K. Chatterjee, 33–58. Cambridge: Cambridge University Press.

Arrhenius, Gustaf. 2009. Can the person affecting restriction solve the problems in population ethics. In *Harming future persons: Ethics, genetics and the nonidentity problem*, International Library of Ethics, Law, and the New Medicine, vol. 35, ed. Melinda A. Roberts and David T. Wasserman, 289–314. Dordrecht: Springer.

Arrow, Kenneth. 1951/1963. *Social choice and individual values*. New York: Wiley.

Atkinson, Anthony B. 1970. On the measurement of inequality. *Journal of Economic Theory* 2: 244–263.

Atwell, John E. 1969. Oldenquist on rules and consequences. *Mind* 78: 576–579.

Ayer, Alfred J. 1952. *Language, truth, and logic*. New York: Dover Publications.

Bach, Kent. 1980. Actions are not events. *Mind* LXXXIX(353): 114–120.

N. Mukerji, *The Case Against Consequentialism Reconsidered*,
Theory and Decision Library A: 51, DOI 10.1007/978-3-319-39249-3

Bach, Kent. 2010. Refraining, omitting, and negative acts. In *A companion to the philosophy of action*, ed. Timothy O'Connor and Constantine Sandis, 50–56. Malden: Wiley-Blackwell.
Bales, R.E. 1971. Act-utilitarianism: Account of right-making characteristics or decision-making procedure? *American Philosophical Quarterly* 8(3): 257–265.
Barnes, W.H.F. 1934. A suggestion about value. *Analysis* 1(3): 45–46.
Barry, Brian. 1995. *Justice as impartiality*. Oxford: Clarendon Press.
Bazerman, Max H., and Ann E. Tenbrunsel. 2011. *Blind spots: Why we fail to do what's right and what to do about it*. Princeton: Princeton University Press.
Beauchamp, Tom L., and James F. Childress. 2001. *Principles of biomedical ethics*, 5th ed. Oxford: Oxford University Press.
Bell, David Q. 2007. *Casuistry: Towards a more complete approach*. Ann Arbor: University of Illinois at Urbana-Champaign.
Bennett, Jonathan F. 1998. *The act itself*. Oxford: Oxford University Press.
Bentham, Jeremy. 1838. *The collected works of Jeremy Bentham: Published under the superintendence of his exectutor John Bowring*, vol. 1. Edinburgh: Williams Tait.
Bergström, Lars. 1982. Interpersonal utility comparisons. *Grazer Philosophische Studien* 16(17): 283–312.
Bergström, Lars. 1966. *The alternatives and consequences of actions: An essay on certain fundamental notions in teleological ethics*. Stockholm: Almqvist & Wiksell.
Bergström, Lars. 1996. Reflections on consequentialism. *Theoria* 62(1–2): 74–94.
Berker, Selim. 2009. The normative insignificance of neuroscience. *Philosophy and Public Affairs* 37(4): 293–329.
Blackorby, Charles, David Donaldson, and John A. Weymark. 1984. Social choice with interpersonal utility comparisons: A diagrammatic introduction. *International Economic Review* 25(2): 327–356.
Blum, Lawrence A. 1988. Gilligan and Kohlberg: Implications for moral theory. *Ethics* 98(3): 472–491.
Brandt, Richard B. 1992. *Morality, utilitarianism, and rights*. Cambridge: Cambridge University Press.
Brickman, Philip, Dan Coates, and Ronnie Janoff-Bulman. 1978. Lottery winners and accident victims: Is happiness relative? *Journal of Personality and Social Psychology* 36(8): 917–927.
Brink, David O. 1984. Moral realism and the sceptical arguments from disagreement and queerness. *Australasian Journal of Philosophy* 62(2): 111–125.
Brink, David O. 1986. Utilitarian morality and the personal point of view. *The Journal of Philosophy* 83(8): 417–438.
Brink, David O. 1989. *Moral realism and the foundations of ethics*, Cambridge Studies in Philosophy. Cambridge: Cambridge University Press.
Brink, David O. 2006. Some forms and limits of consequentialism. In *The Oxford handbook of ethical theory*, ed. David Copp, 380–423. Oxford: Oxford University Press.
Broad, Charlie D. 1971. Self and others. In *Broad's critical essays in moral philosophy*, 262–282. London: Allen and Unwin.
Broome, John. 1991. *Weighing goods: Equality, uncertainty and time*. Oxford: Basil Blackwell.
Broome, John. 1999. *Ethics out of economics*. Cambridge: Cambridge University Press.
Broome, John. 2004. *Weighing lives*. Oxford: Oxford University Press.
Brown, Campbell. 2011. Consequentialize this. *Ethics* 121(4): 749–771.
Buchanan, James M. 1975. *The limits of liberty: Between anarchy and leviathan*. Chicago: University of Chicago Press.
Buchanan, James M., and Geoffrey Brennan. 2000. *The reason of rules*. Indianapolis: Liberty Fund.
Bykvist, Kirster. 2002. Alternative actions and the spirit of consequentialism. *Philosophical Studies* 107(1): 45–68.
Bykvist, Kirster. 2009. *Utilitarianism: A guide for the perplexed*. London: Continuum.
Carlson, Erik. 1995. *Consequentialism reconsidered*. Dordrecht: Kluwer Academic Publishers.
Carlson, Erik. 1999a. Consequentialism, alternatives, and actualism. *Philosophical Studies* 96: 253–268.

Carlson, Erik. 1999b. The oughts and cans of objective consequentialism. *Utilitas* 11(1): 91–96.
Carroll, Lewis. 1871/1990. *Through the looking-glass, and what Alice found there.* London: Random House Value Publishing.
Castaneda, Hector-Neri. 1968. A problem for utilitarianism. *Analysis* 28(4): 141–142.
Caws, Peter. 1995. Minimal consequentialism. *Philosophy* 70(273): 313–339.
Chisholm, Roderick M. 1964. The descriptive element in the concept of action. *The Journal of Philosophy* 61(20): 613–625.
Chong, Chong K. 1992. Ethical egoism and the moral point of view. *The Journal of Value Inquiry* 26: 23–36.
Cocking, Dean, and Justin Oakley. 1995. Indirect consequentialism, friendship, and the problem of alienation. *Ethics* 106(1): 86–111.
Cowell, Frank A. 2000. Chapter 2: Measurement of inequality. In *Handbook of income distribution*, ed. Anthony B. Atkinson and François Bourguignon, 87–166. Amsterdam: Elsevier.
Crisp, Roger. 2005. Deontological ethics. In *The Oxford companion to philosophy*. ed. Ted Honderich, 200–201. Oxford: Oxford University Press.
Crisp, Roger. 2006. *Reasons and the good*. Oxford: Clarendon Press.
Culyer, A.J. 1989. The normative economics of health care finance and provision. *Oxford Review of Economic Policy* 5(1): 34–58.
Cummiskey, David. 1990. Kantian consequentialism. *Ethics* 100(3): 586–615.
Cummiskey, David. 1996. *Kantian consequentialism*. New York: Oxford University Press.
Cyert, R.M., and J.G. March. 1963/1992. *A behavioral theory of the firm.* Malden: Blackwell Business.
D'Aspremont, Claude, and Louis Gevers. 1977. Equity and the informational basis of collective choice. *Review of Economic Studies* 44: 199–209.
Daniels, Norman. 1979. Wide reflective equilibrium and theory acceptance in ethics. *The Journal of Philosophy* 76(5): 256–282.
Daniels, Norman. 1996. *Justice and justification: Reflective equilibrium in theory and practice*. Cambridge: Cambridge University Press.
Daniels, Norman. 2011. Reflective equilibrium. In *The Stanford encyclopedia of philosophy. Spring 2011 edition*, ed. Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information (Stanford University).
Danto, Arthur C. 1965. Basic actions. *American Philosophical Quarterly* 2(2): 141–148.
Danto, Arthur C. 1969. Complex events. *Philosophy and Phenomenological Research* 30(1): 66–77.
Darwall, Stephen. 2003. Agent-centred restrictions from the outside in. In *Deontology*, ed. Stephen Darwall, 112–138. Oxford: Blackwell Publishers.
Darwall, Stephen. 2004. *Welfare and rational care*. Princeton: Princeton University Press.
Davidson, Donald. 1963. Actions, reasons, and causes. *The Journal of Philosophy* 60(23): 685–700.
Dennett, Daniel C. 1984. *Elbow room: The varieties of free will worth wanting*. Cambridge: MIT Press.
Dinwiddy, John R., and William L. Twining. 2004. *Bentham: Selected writings of John Dinwiddy*. Standford: Stanford University Press.
Dorsey, Dale. 2005. Global justice and the limits of human rights. *The Philosophical Quarterly* 55(221): 562–581.
Dorsey, Dale. 2009. Aggregation, partiality, and the strong beneficence principle. *Philosophical Studies* 146(1): 139–157.
Dorsey, Dale. 2012. Consequentialism, metaphysical realism and the argument from cluelessness. *The Philosophical Quarterly* 62(246): 48–70.
Dreier, James. 1993. Structures of normative theories. *The Monist* 76: 22–40.
Dreier, James. 2004. Why ethical satisficing makes sense and rational satisficing doesn't. In *Satisficing and maximizing: Moral theorists on practical reason*, ed. Michael Byron, 131–154. Cambridge: Cambridge University Press.

Dreier, James. 2006. Introduction. In *Contemporary debates in moral theory*, ed. Dreier James, x–xxiv. Malden: Blackwell.
Driver, Julia. 2001/2003. *Uneasy virtue.* Cambridge Studies in Philosophy. Cambridge: Cambridge University Press.
Driver, Julia. 2009. Gertrude Elizabeth Margaret Anscombe. In *The Stanford encyclopedia of philosophy. Fall 2009 edition*, ed. Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information (Stanford University).
Driver, Julia, and Don Loeb. 2008. Moral heuristics and consequentialism. In *Moral psychology (volume 2): The cognitive science of morality: Intuition and diversity*, ed. Walter Sinnott-Armstrong, 31–40. Cambridge, MA: MIT Press.
Dwyer, Susan. 1999. Moral competence. In *Philosophy and linguistics*, ed. K. Murasugi and R. Stainton, 169–190. Boulder: Westview Press.
Elliot, Robert. 1997. *Faking nature: The ethics of environmental restoration*. London: Routledge.
Elliot, Robert. 2003. Normative ethics. In *A companion to environmental philosophy*, ed. D. Jamieson, 177–191. Oxford: Blackwell.
Elster, Jon. 1983. *Sour grapes: Studies in the subversion of rationality*. Cambridge: Cambridge University Press.
Elster, Jakob. 2011. How outlandish can imaginary cases be? *Journal of Applied Philosophy* 28(3): 241–258.
Feldman, Fred. 1986. *Doing the best we can: An essay in informal deontic logic*. Dordrecht: D. Reidel Pub. Co.
Feldman, Fred. 1995. Adjusting utility for justice: A consequentialist reply to the objection from justice. *Philosophy and Phenomenological Research* 55(3): 567–585.
Feldman, Fred. 1997. *Utilitarianism, hedonism, and desert: Essays in moral philosophy*. Cambridge: Cambridge University Press.
Feldman, Fred. 2004. *Pleasure and the good life: Concerning the nature, varieties, and plausibility of hedonism*. Oxford: Clarendon Press.
Feldman, Fred. 2006. Actual utility, the objection from impracticality, and the move to expected utility. *Philosophical Studies* 129(1): 49–79.
Fishkin, James S. 1982. *The limits of obligation*. New Haven: Yale University Press.
Flanagan, Owen J. 1993. *Varieties of moral personality: Ethics and psychological realism*. Cambridge: Harvard University Press.
Fleurbaey, Marc. 1996. *Théories Économiques de la Justice*. Paris: Economica.
Fodor, Jerry A. 1998. *Concepts: Where cognitive science went wrong*, Oxford Cognitive Science Series. Oxford: Oxford University Press.
Foot, Philippa. 1978. *Virtues and vices and other essays in moral philosophy*. Berkeley: University of California Press.
Foot, Philippa. 1985. Utilitarianism and the virtues. *Mind* 94(374): 196–209.
Forschler, Scott. 2009. Truth and acceptance conditions for moral statements can be identical: Further support for subjective consequentialism. *Utilitas* 21(3): 337–346.
Forster, Michael. 2010. Wittgenstein on family resemblance concepts. In *Wittgenstein's philosophical investigations: A critical guide*, ed. Arif Ahmed, 66–87. Cambridge: Cambridge University Press.
Frankena, William K. 1963/1973. *Ethics.* 2nd ed. Englewood Cliffs: Prentice-Hall.
Frankfurt, Harry G. 1978. The problem of action. *American Philosophical Quarterly* 15(2): 157–162.
Frederick, Shane, George Loewenstein, and Ted O'Donoghue. 2002. Time discounting and time preference: A critical review. *Journal of Economic Literature* 40(2): 351–401.
French, Peter A. 1979. The corporation as a moral person. *American Philosophical Quarterly* 16(3): 207–215.
French, Peter A. 1984. *Collective and corporate responsibility*. New York: Columbia University Press.
Fried, Barbara H. 2012. What does matter? The case for killing the trolley problem (or letting it die). *The Philosophical Quarterly* 62(248): 1–25.

Gaertner, Wulf. 2006. *A primer in social choice theory*. Oxford: Oxford University Press.
Gauthier, David. 1987. *Morals by agreement*. New York: Oxford University Press.
Geach, Peter T. 1969. *God and the soul*. London: Routledge and Kegan Paul.
Geirsson, Heimir, and Margaret R. Holmgren. 2000. *Ethical theory: A concise anthology*. Peterborough: Broadview Press.
Gert, Bernard. 1998/2005. *Morality: Its nature and justification.* Oxford: Oxford University Press.
Gert, Bernard. 1993. Transplants and trolleys. *Philosophy and Phenomenological Research* 53(1): 173–179.
Gert, Bernard. 2004. *Common morality: Deciding what to do*. New York: Oxford University Press.
Gettier, Edmund L. 1963. Is justified true belief knowledge? *Analysis* 23(6): 121–123.
Gewirth, Alan. 1988. Ethical universalism and particularism. *The Journal of Philosophy* 85(6): 283–302.
Gibbard, Allen. 1990. *Wise choices, apt feelings: A theory of normative judgment*. Cambridge: Harvard University Press.
Gigerenzer, Gerd. 2007. *Gut feelings*. New York: Viking.
Gigerenzer, Gerd. 2008. Moral intuition – Fast and frugal heuristics? In *Moral psychology (volume 2): The cognitive science of morality: Intuition and diversity*, ed. Walter Sinnott-Armstrong, 1–26. Cambridge, MA: MIT Press.
Gigerenzer, Gerd. 2010. Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in Cognitive Science* 2(3): 528–554.
Gigerenzer, Gerd, and Peter M. Todd. 1999. Fast and frugal heuristics: The adaptive toolbox. In *Simple heuristics that make us smart*, ed. Gerd Gigerenzer, Peter M. Todd, and ABC Research Group, 3–34. Oxford: Oxford University Press.
Glover, Jonathan. 1977. *Causing death and saving lives*. London: Penguin.
Godwin, William. 1793. *An enquiry concerning political justice and its influence on general virtue and happiness*. London: G. G. J. and J. Robinson.
Goldman, Holly S. 1978. Doing the best one can. In *Values and morals: Essays in honor of William Frankena, Charles Stevenson, and Richard Brandt*, ed. Alvin I. Goldman and Jaegwon Kim, 185–214. Dordrecht: Reidel.
Goldman, Alan H. 2003. *Practical rules: When we need them and when we don't*. Cambridge: Cambridge University Press.
Goodin, Robert E. 1995. *Utilitarianism as a public philosophy*. Cambridge: Cambridge University Press.
Graafland, Johan J. 2007. *Economics, ethics and the market: Introduction and applications*. Abingdon: Routledge.
Greene, Joshua D. 2008. The secret joke of Kant's soul. In *Moral psychology (volume 3): The neuroscience of morality: Emotion, brain disorders, and development*, ed. Walter Sinnott-Armstrong, 35–80. Cambridge, MA: MIT Press.
Greene, Joshua D. 2010. Notes on 'The Normative Insignificance of Neuroscience' by Selim Berker. https://static1.squarespace.com/static/54763f79e4b0c4e55ffb000c/t/54cb945ae4b001aedee69e81/1422627930781/notes-on-berker.pdf. Accessed 11 Apr 2016.
Greene, Joshua D., B.R. Sommerville, L.E. Nystrom, and J.D. Cohen. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science* 293(5537): 2105–2108.
Greene, Joshua D., Fiery A. Cushman, Lisa E. Stewart, Kelly Lowenberg, Leigh E. Nystrom, and Jonathan D. Cohen. 2009. Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition* 111(3): 364–371.
Griffin, James. 1986. *Well-being: Its meaning, measurement and moral importance*. Oxford: Clarendon Press.
Griffin, James. 1994. The distinction between criterion and decision procedure: A reply to Madison Powers. *Utilitas* 6(2): 177–182.
Grisez, Germain. 1978. Against consequentialism. *The American Journal of Jurisprudence* 23: 21–72.
Haidt, Jonathan, and Jonathan Baron. 1996. Social roles and the moral judgement of acts and omissions. *European Journal of Social Psychology* 26: 201–218.

Hardin, Russell. 1988. *Morality within the limits of reason*. Chicago: The University of Chicago Press.
Hare, Richard M. 1961. *The language of morals*. Oxford: Oxford University Press.
Hare, Richard M. 1981. *Moral thinking: Its levels, method and point*. Oxford: Clarendon Press.
Harrison, Jonathan. 1967. Ethical objectivism. In *The encyclopedia of philosophy: Epictetus to Hilbert*, ed. Paul Edwards, 71–75. London: Macmillan.
Harrod, R.F. 1936. Utilitarianism revised. *Mind* 45(178): 137–156.
Harsanyi, John C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *The Journal of Political Economy* 63(4): 309–321.
Harsanyi, John C. 1977a. Morality and the theory of rational behavior. *Social Research* 44(4): 623–656.
Harsanyi, John C. 1977b. Rule utilitarianism and decision theory. *Erkenntnis* 11(1): 25–53.
Hauser, Marc. 2008. *Moral minds: How nature designed our universal sense of right and wrong*. London: Little, Brown Book Group.
Hauser, Marc, Fiery Cushman, Young Liane, R.K.-X. Jin, and John Mikhail. 2007. A dissociation between moral judgments and justifications. *Mind and Language* 22(1): 1–21.
Haybron, Dan. 2011. Happiness. In *The Stanford encyclopedia of philosophy. Fall 2011 edition*, ed. Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information (Stanford University).
Heidegger, Martin. 1953/2000. *Introduction to metaphysics*. New Haven: Yale University Press.
Henson, Richard G. 1971. Utilitarianism and the wrongness of killing. *The Philosophical Review* 80(3): 320–337.
Hirose, Iwao. 2004. Aggregation and numbers. *Utilitas* 16(1): 62–79.
Hirose, Iwao. 2007. Review article: Aggregation and non-utilitarian moral theories. *Journal of Moral Philosophy* 4(2): 273–284.
Hirose, Iwao. 2011. *Moral aggregation: Unpublished book manuscript (Januar 2011)*. McGill University.
Hodgson, David H. 1967. *Consequences of utilitarianism: A study in normative ethics and legal theory*. Oxford: Clarendon Press.
Holtug, Nils. 2003. Welfarism – The very idea. *Utilitas* 15(2): 151–174.
Holtug, Nils, and Kasper Lippert-Rasmussen. 2007. An introduction to comtemporary egalitarianism. In *Egalitarianism: New essays on the nature and value of equality*, ed. Nils Holtug and Kasper Lippert-Rasmussen, 1–38. Oxford: Clarendon Press.
Hooker, Brad. 1990. Rule-consequentialism. *Mind* 99(393): 67–77.
Hooker, Brad. 1994. Is rule-consequentialism a rubber duck? *Analysis* 54(2): 92–97.
Hooker, Brad. 2000. Rule consequentialism. In *The Blackwell guide to ethical theory*, ed. Hugh LaFollette, 183–204. Oxford: Blackwell Publishers.
Hooker, Brad. 2003. *Ideal code, real world: A rule-consequentialist theory of morality*. Oxford: Oxford University Press.
Hooker, Brad. 2009a. The demandingness objection. In *The problem of moral demandingness: New philosophical essays*, ed. T.D.J. Chappell, 148–162. Hampshire: Palgrave Macmillan.
Hooker, Brad. 2009b. Up and down with aggregation. *Social Philosophy and Policy* 26(1): 126–147.
Hörster, Norbert. 1973. Is act-utilitarian truth-telling self-defeating? *Mind* 82: 413–416.
Howard-Snyder, Frances. 1993. Rule consequentialism is a rubber duck. *American Philosophical Quarterly* 30(3): 271–278.
Howard-Snyder, Frances. 1994. The heart of consequentialism. *Philosophical Studies* 76(1): 107–129.
Howard-Snyder, Frances. 1996. A new argument for consequentialism? A reply to Sinnott-Armstrong. *Analysis* 56(2): 111–115.
Howard-Snyder, Frances. 1997. The rejection of objective consequentialism. *Utilitas* 9(2): 241–248.
Howard-Snyder, Frances. 1999. Response to Carlson and Qizilbash. *Utilitas* 11(1): 106–111.
Howard-Snyder, Frances. 2005. It's the thought that counts. *Utilitas* 17(3): 265–281.

Huebner, B., S. Dwyer, and M. Hauser. 2009. The role of emotion in moral psychology. *Trends in Cognitive Sciences* 13(1): 1–6.
Hume, David. 1888/1960. *A treatise of human nature,* ed. L. A. Selby-Bigge. Oxford: Clarendon Press.
Hurka, Thomas. 1990. Two kinds of satisficing. *Philosophical Studies* 59: 107–111.
Hurka, Thomas. 1992. Consequentialism and content. *American Philosophical Quarterly* 29(1): 71–78.
Hurka, Thomas. 1993. *Perfectionism*. Oxford: Oxford University Press.
Hurka, Thomas. 2004. Satisficing and substantive values. In *Satisficing and maximizing: Moral theorists on practical reason*, ed. Michael Byron, 71–76. Cambridge: Cambridge University Press.
Hursthouse, Rosalind. 2012. Virtue ethics. In *The Stanford encyclopedia of philosophy*. Summer 2012 edition, ed. Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information (Stanford University).
Ichikawa, Jonathan J., and Matthias Steup. 2012. The analysis of knowledge. In *The Stanford encyclopedia of philosophy. Winter 2012 edition*, ed. Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information (Stanford University).
Jamieson, Dale. 1984. Utilitarianism and the morality of killing. *Philosophical Studies* 45(2): 209–221.
Jeske, Diane. 2008. Special obligations. In *The Stanford encyclopedia of philosophy. Fall 2008 edition*, ed. Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information (Stanford University).
Jevons, William Stanley. 1871. *The theory of political economy 1*. London: Macmillan.
Jollimore, Troy. 2008. Impartiality. In *The Stanford encyclopedia of philosophy. Fall 2008 edition*, ed. Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information (Stanford University).
Joyce, Richard. 2009. Moral anti-realism. In *The Stanford encyclopedia of philosophy. Summer 2009 edition*, ed. Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information (Stanford University).
Kagan, Shelly. 1988. The additive fallacy. *Ethics* 99(1): 5–31.
Kagan, Shelly. 1989. *The limits of morality*. Oxford: Clarendon Press.
Kagan, Shelly. 1992. The structure of normative ethics. *Philosophical Perspectives* 6: 223–242.
Kagan, Shelly. 1998. *Normative ethics*. Boulder: Westview Press.
Kagan, Shelly. 2001. Thinking about cases. *Social Philosophy and Policy* 18(2): 44–63.
Kahane, Guy. 2013. The armchair and the trolley: An argument for experimental ethics. *Philosophical Studies* 162: 421–445.
Kamm, Frances M. 1983. Killing, letting die: Methodological and substantive issues. *Pacific Philosophical Quarterly* 64: 297–312.
Kamm, Frances M. 1993. *Morality, mortality: Death and whom to save from it 1*. New York: Oxford University Press.
Kamm, Frances M. 1996. *Morality, mortality: Rights, duties, and status 2*. Oxford: Oxford University Press.
Kamm, Frances M. 2007. *Intricate ethics: Rights, responsibilities, and permissible harm*, Oxford Ethics Series. Oxford: Oxford University Press.
Kant, Immanuel. 1785. *Grundlegung zur Metaphysik der Sitten*. Riga: Johann Friedrich Hartknoch.
Kant, Immanuel. 1799. Über ein vermeintliches Recht aus Menschenliebe zu lügen (1797). In *Immanuel Kant's vermischte Schriften*, 357–368. Riga: in der Rengerschen Buchhandlung.
Kant, Immanuel. 1803. *Die Metaphysik der Sitten*, vol. 1–2. Königsberg: F. Nicolovius.
Kaplow, Louis, and Steven Shavell. 2001. Any non-welfarist method of policy assessment violates the Pareto principle. *The Journal of Political Economy* 109(2): 281–286.
Kappel, Klemens. 2006. The meta-justification of reflective equilibrium. *Ethical Theory and Moral Practice* 9(2): 131–147.
Keller, Simon. 2009. Welfarism. *Philosophy Compass* 4(1): 82–95.

Kim, Jaegwon. 1976. Events as property exemplifications. In *Action theory: Proceedings of the Winnipeg conference on human action, held at Winnipeg, Manitoba, Canada, 9–11 May 1975*, ed. M. Brand and D.N. Walton, 159–178. Dordrecht: D. Reidel Pub. Co.

Knobe, Joshua, and Shaun Nichols (eds.). 2008. *Experimental philosophy*. Oxford: Oxford University Press.

Korsgaard, Christine M. 2008. *The constitution of agency: Essays on practical reason and moral psychology*. Oxford: Oxford University Press.

Kymlicka, Will. 2002. *Contemporary political philosophy: An introduction*, 2nd ed. Oxford: Oxford University Press.

Lakatos, Imre. 1970. Falsification and the methodology of scientific research programmes. In *Criticism and the growth of knowledge*, ed. Imre Lakatos and Alan Musgrave, 91–196. Cambridge: Cambridge University Press.

Lang, Gerald R. 2013. Should utilitarianism be scalar? *Utilitas* 25(1): 80–95.

Lawlor, Rob. 2009a. *Shades of goodness: Gradability, demandingness and the structure of moral theories*. Hampshire: Palgrave Macmillan.

Lawlor, Rob. 2009b. The rejection of scalar consequentialism. *Utilitas* 21(1): 100–116.

Lenman, James. 2008. Moral naturalism. In *The Stanford encyclopedia of philosophy. Winter 2008 edition*, ed. Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information (Stanford University).

Lillehammer, Hallvard. 2011. The epistemology of ethical intuitions. *Philosophy* 86(02): 175–200.

Lippert-Rasmussen, Kasper. 2005. *Deontology, responsibility, and equality*. Copenhagen: Dept. of Media, Cognition and Communication, Univ. of Copenhagen.

List, Christian, and Philip Pettit. 2011. *Group agency: The possibility, design, and status of corporate agents*. Oxford: Oxford University Press.

Little, Ian M.D. 1950. *A critique of welfare economics*. Oxford: Clarendon Press.

Loewenstein, George. 2005. Hot-cold empathy gaps and medical decision making. *Health Psychology* 24(4): 49–56.

Lübbe, Weimar. 2008. Taurek's no worse claim. *Philosophy and Public Affairs* 36(1): 69–85.

Luetge, Christoph, and Nikil Mukerji, eds. 2016. *Order Ethics: An ethical framework for the social market economy*. Dordrecht: Springer.

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

Macklin, Ruth. 1967a. A rejoinder. *The Journal of Value Inquiry* 1: 135–138.

Macklin, Ruth. 1967b. Actions, consequences and ethical theory. *The Journal of Value Inquiry* 1: 72–80.

Mason, Elinor. 2002. Against blameless wrongdoing. *Ethical Theory and Moral Practice* 5: 287–303.

McDermott, Michael. 1982. Utility and distribution. *Mind* 91(364): 572–578.

McElwee, Brian. 2010. Should we de-moralize ethical theory? *Ratio* 23(3): 308–321.

McElwee, Brian. 2011. Impartial reasons, moral demands. *Ethical Theory and Moral Practice* 14(4): 457–466.

McKerlie, Dennis. 1994. Equality and priority. *Utilitas* 6(1): 25–42.

McMahan, Jeff. 2002. *The ethics of killing: Problems at the margins of life*. Oxford: Oxford University Press.

McNamara, Paul. 2010. Deontic logic. In *The Stanford encyclopedia of philosophy. Fall 2010 edition*, ed. Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information (Stanford University).

McNaughton, David. 1996. An unconnected heap of duties? *The Philosophical Quarterly* 46(185): 433–447.

McNaughton, David, and Piers Rawling. 1991. Agent-relativity and the doing-happening distinction. *Philosophical Studies* 63(2): 167–185.

McNaughton, David, and Piers Rawling. 1995. Value and agent-relative reasons. *Utilitas* 7(1): 31–47.

McNaughton, David, and Piers Rawling. 2009. Benefits, holism, and the aggregation of value. *Social Philosophy and Policy* 26(1): 354–374.

Mendola, Joseph. 2005a. Consequentialism, group acts, and trolleys. *Pacific Philosophical Quarterly* 86: 64–87.
Mendola, Joseph. 2005b. Intuitive maximin. *Canadian Journal of Philosophy* 35(3): 429–440.
Mendola, Joseph. 2006. *Goodness and justice: A consequentialist moral theory*, Cambridge Studies in Philosophy. Cambridge: Cambridge University Press.
Mikhail, John. 2007. Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences* 11(4): 143–152.
Mill, John Stuart. 1863. *Utilitarianism*. London: Parker, Son and Bourn.
Mill, John Stuart. 1882. *A system of logic, ratiocinative and inductive, being a connected view of the principles of evidence, and the methods of scientific investigation*, 8th ed. New York: Harper & Brothers.
Mill, John Stuart. 2004. *Autobiography*, With the assistance of. J. Manis. Electronic Classics Series. Hazleton: Pennsylvania State University.
Miller, David. 2008. Political philosophy for earthlings. In *Political theory: Methods and approaches*, ed. David Leopold and Marc Stears, 29–48. Oxford: Oxford University Press.
Moore, G.E. 1903/1959. *Principia ethica.* Cambridge: Cambridge University Press.
Moore, Michael. 2008. Patrolling the borders of consequentialist justifications: The scope of agent-relative restrictions. *Law and Philosophy* 27(1): 35–96.
Moore, Andrew, and Roger Crisp. 1996. Welfarism in moral theory. *Australasian Journal of Philosophy* 74(4): 598–613.
Mossel, Benjamin. 2009. Negative actions. *Philosophia* 37(2): 307–333.
Mukerji, Nikil. 2009. *Das Differenzprinzip von John Rawls und seine Realisierungsbedingungen*. Münster: LIT Verlag.
Mukerji, Nikil. 2013a. Consequentialism, deontology and the morality of promising. In *Business ethics and risk management*, ed. Johanna Jauernig and Christoph Luetge, 111–126. Dordrecht: Springer.
Mukerji, Nikil. 2013b. The case against consequentialism: Methodological issues. In *GAP.8 Proceedings: Was dürfen wir glauben? Was sollen wir tun?,* ed. Miguel Hoeltje, Thomas Spitzley, and Wolfgang Spohn, 654–665. Duisburg-Essen: DuEPublico.
Mukerji, Nikil. 2013c. Utilitarianism. In *Handbook of the philosophical foundations of business ethics*, vol. 1, ed. Christoph Luetge, 297–312. Dordrecht: Springer.
Mukerji, Nikil. 2014. Intuitions, experiments, and armchairs. In *Experimental ethics – Toward an empirical moral philosophy*, vol. 1, ed. Christoph Luetge, Hannes Rusch, and Matthias Uhl, 227–243. London: Palgrave Macmillan.
Mukerji, Nikil. 2015. Experimentelle Ethik. In *Handbuch Philosophie und Ethik*, vol. 2, ed. Julian Nida-Rümelin, Irina Spiegel, and Markus Tiedemann, 93–101. Ferdinand Schöning: Paderborn.
Mukerji, Nikil, and Christoph Luetge. 2014. Responsibility, order ethics, and group agency. *Archiv für Rechts- und Sozialphilosophie* 100(2): 176–186.
Mulgan, Timothy. 2001a. How satisficers get away with murder. *International Journal of Philosophical Studies* 9(1): 41–46.
Mulgan, Timothy. 2001b. *The demands of consequentialism*. Oxford: Clarendon Press.
Mulgan, Timothy. 2005. Reply to John Turri. *International Journal of Philosophical Studies* 13(4): 493–496.
Mulgan, Timothy. 2007. *Understanding Utilitarianism*. Stocksfield: Acumen Publishing Limited.
Müller-Lyer, Franz Carl. 1889. Optische Urtheilstäuschungen. *Archiv für Anatomie und Physiologie, Physiologische Abteilung* 2(Supplement): 263–270.
Nagel, Thomas. 1978. *The possibility of altruism*. Chichester: Princeton University Press.
Nagel, Thomas. 1986. *The view from nowhere*. Oxford: Oxford University Press.
Nagel, Thomas. 1991. *Mortal questions*. Cambridge: Cambridge University Press.
Narveson. 1967. Utilitarianism and new generations. *Mind* 76(301): 62–72.
Narveson. 2004. Maxificing: Life on a budget; or, if you would maximize, then satisfice! In *Satisficing and maximizing: Moral theorists on practical reason*, ed. Michael Byron, 59–70. Cambridge: Cambridge University Press.

Ng, Yew-Kwang. 1981. Welfarism: A defence against Sen's attack. *The Economic Journal* 91(362): 527–530.
Ng, Yew-Kwang. 1990. Welfarism and utilitarianism: A rehabilitation. *Utilitas* 2(2): 171–193.
Ng, Yew-Kwang. 2004. *Welfare economics: Towards a more complete analysis*. Hampshire: Palgrave Macmillan.
Nida-Rümelin, Julian. 1993. *Kritik des Konsequentialismus*. München: Oldenbourg Verlag.
Nida-Rümelin, Julian. 1997a. *Economic rationality and practical reason*. Dordrecht: Springer.
Nida-Rümelin, Julian. 1997b. Why consequentialism fails. In *Contemporary action theory*, ed. G. Holmström-Hintikka and R. Tuomela, 295–308. Dordrecht: Kluwer Academic Pub.
Nida-Rümelin, Julian. 2002. *Ethische essays*. Frankfurt am Main: Suhrkamp.
Nida-Rümelin, Julian. 2005. Why rational deontological action optimizes subjective value. *Protosociology* 21: 182–193.
Nida-Rümelin, Julian. 2011a. *Die Optimierungsfalle: Philosophie einer humanen Ökonomie*. München: Irisiana.
Nida-Rümelin, Julian. 2011b. *Verantwortung*. Stuttgart: Reclam.
Nielsen, Kai. 1978. Class and justice. In *Justice and economic distribution*, ed. John Arthur and William Shaw, 225–245. Englewood Cliffs: Prentice-Hall.
Nielsen, Kai. 1981. Impediments to radical egalitarianism. *American Philosophical Quarterly* 18(2): 121–129.
Niiniluoto, Ilkka. 2011. Scientific progress. In *The Stanford encyclopedia of philosophy. Summer 2011 edition*, ed. Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information (Stanford University).
Norcross, Alastair. 1995. Should utilitarianism accommodate moral dilemmas? *Philosophical Studies* 79: 59–83.
Norcross, Alastair. 1998. Great harms from small benefits grow: How death can be outweighed by headaches. *Analysis* 58(2): 152–158.
Norcross, Alastair. 2006. The scalar approach to utilitarianism. In *Mill's utilitarianism*, ed. Henry R. West, 217–232. Oxford: Blackwell Publishing.
Norcross, Alastair. 2008. Off Her Trolley? Frances Kamm and the metaphysics of morality. *Utilitas* 20(1): 65–80.
Norcross, Alastair. 2009. Two dogmas of deontology: Aggregation, rights, and the separateness of persons. *Social Philosophy and Policy* 26(1): 76.
Nozick, Robert. 1974. *Anarchy, state, and Utopia*. New York: Basic Books.
Nozick, Robert. 1993. *The nature of rationality*. Princeton: Princeton University Press.
Oddie, Graham. 2001. Axiological atomism. *Australasian Journal of Philosophy* 79(3): 313–332.
Oddie, Graham, and Peter Menzies. 1992. An objectivist's guide to subjective value. *Ethics* 102(3): 512–533.
Oddie, Graham, and Peter Milne. 1991. Act and value: Expectation and the representability of moral theories. *Theoria* 57(1–2): 42–76.
Oldenquist, Andrew. 1966. Rules and consequences. *Mind* 75(298): 180–192.
Österberg. 1988. *Self and others: A study of ethical egoism*. Dordrecht: Kluwer Academic Publishers.
Otsuka, Michael. 2006. Saving lives, moral theory, and the claims of individuals. *Philosophy and Public Affairs* 34(2): 109–135.
Ott, Konrad. 2004. Essential components of future ethics. In *Ökonomische Rationalität und praktische Vernunft: Gerechtigkeit, ökologische Ökonomie und Naturschutz: Eine Festschrift anlässlich des 60. Geburtstags von Ulrich Hampicke*, ed. Ralf Döring and Michael Rühs, 83–108. Würzburg: Königshausen und Neumann.
Parfit, Derek. 1986. *Reasons and persons*. Oxford: Oxford University Press.
Parfit, Derek. 2003. Equality and priority. In *Debates in contemporary political philosophy – An anthology*, ed. Derek Matravers and Jon Pike, 115–132. London: Routledge.
Parfit, Derek. 2011. *On what matters I*. Oxford: Oxford University Press.
Pears, David. 1971. In *Agent, action and reason*, ed. Robert W. Binkley, Richard Bronaugh, and Ausonio Marras. Toronto: University of Toronto.

Percival, Philip. 2002. Epistemic consequentialism. *Proceedings of the Aristotelian Society, Supplementary Volumes* 76: 121–151.
Person, Ingmar. 1985. The universal basis of egoism. *Theoria* 51(3): 137–158.
Peterson, Martin. 2010. A royal road to consequentialism? *Ethical Theory and Moral Practice* 13(2): 153–169.
Peterson, Martin. 2012. Multi-dimensional consequentialism. *Ratio* 25(2): 177–194.
Petrinovich, Lewis, and Patrica O'Neill. 1996. Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology* 17(3): 145–171.
Pettit, Philip. 1984. Satisficing consequentialism. *Proceedings of the Aristotelian Society, Supplementary Volumes* 58: 165–176.
Pettit, Philip. 1987. Rights, constraints and trumps. *Analysis* 47(1): 8–14.
Pettit, Philip. 1988a. The consequentialist can recognize rights. *The Philosophical Quarterly* 38(150): 42–55.
Pettit, Philip. 1988b. The paradox of loyalty. *American Philosophical Quarterly* 25(2): 163–171.
Pettit, Philip. 1993. Introduction. In *Consequentialism*, ed. Pettit Philip, xiii–xix. Aldershot: Dartmouth.
Pettit, Philip. 1997/2007. The consequentialist perspective. In *Three methods of ethics: A debate*, ed. Marcia W. Baron, Philip Pettit, and Michael A. Slote. 10th print., 92–174. Malden: Blackwell.
Pettit, Philip. 2000. Non-consequentialism and universalizability. *The Philosophical Quarterly* 50(199): 175–190.
Pettit, Philip, and Geoffrey Brennan. 1986. Restrictive consequentialism. *Australiasian Journal of Philosophy* 64(4): 438–455.
Pettit, Philip, and Michael Smith. 2000. Global consequentialism. In *Morality, rules, and consequences: A critical reader*, ed. Elinor Mason, Dale E. Miller, and Brad Hooker, 121–133. Lanham: Rowan & Littlefield.
Pogge, Thomas. 1990. The effects of prevalent moral conceptions. *Social Research* 57(3): 649–663.
Pogge, Thomas. 2000. On the site of distributive justice: Reflections on cohen and murphy. *Philosophy and Public Affairs* 29(2): 137–169.
Popper, Karl Raimund. 1947. *The open society and its enemies: Volume I: The spell of Plato*, 2nd ed. London: George Routledge & Sons, Ltd.
Popper, Karl Raimund. 1959/2005. *The logic of scientific discovery.* e-Library Edition. London: Routledge.
Portmore, Douglas W. 1998. Can consequentialism be reconciled with our common-sense moral intuitions? *Philosophical Studies* 91(1): 1–19.
Portmore, Douglas W. 2001. Can an act-consequentialist theory be agent relative? *American Philosophical Quarterly* 38(4): 363–377.
Portmore, Douglas W. 2005. Combining teleological ethics with evaluator relativism: A promising result. *Pacific Philosophical Quarterly* 86(1): 95–113.
Portmore, Douglas W. 2007. Consequentializing moral theories. *Pacific Philosophical Quarterly* 88: 39–73.
Portmore, Douglas W. 2008. Dual-ranking act-consequentialism. *Philosophical Studies* 138(3): 409–427.
Portmore, Douglas W. 2011. *Commonsense consequentialism: Wherein morality meets rationality*. New York: Oxford University Press.
Prichard, Harold A. 2002. In *Moral writings*, ed. Jim MacAdam. Oxford: Oxford University Press.
Prinz, Jesse. 2010. Ethics and psychology. In *The Routledge companion to ethics*, Routledge Philosophy Companions, ed. John Skorupski, 384–396. Abingdon: Routledge.
Prior, Arthur N. 1960. The autonomy of ethics. *Australasian Journal of Philosophy* 38(3): 199–206.
Putnam, Hilary. 1981. *Reason, truth, and history*. Cambridge: Cambridge University Press.
Putnam, Hilary. 2002. *The collapse of the fact/value dichotomy and other essays*. Cambridge, MA: Harvard University Press.

Qizilbash, Mozaffar. 1999. The rejection of objective consequentialism: A comment. *Utilitas* 11(1): 97–105.
Quinn, Philip. 1990. The recent revival of divine command ethics. *Philosophy and Phenomenological Research* 50(Supplement): 345–365.
Rachels, James. 1997. *Can ethics provide answers? And other essays in moral philosophy*. Lanham: Rowman & Littlefield.
Railton, Peter. 1984. Alienation, consequentialism, and the demands of morality. *Philosophy and Public Affairs* 13(2): 134–171.
Rand, Ayn. 1982. *Philosophy: Who needs it*. New York: Signet.
Rawls, John. 1951. Outline of a decision procedure in ethics. *The Philosophical Review* 60(2): 177–197.
Rawls, John. 1955. Two concepts of rules. *The Philosophical Review* 64(1): 3–32.
Rawls, John. 1958. Justice as fairness. *The Philosophical Review* 67(2): 164–194.
Rawls, John. 1971/1999. *A theory of justice.* Cambridge: Belknap Press of Harvard University Press.
Rawls, John. 1974–1975. The independence of moral theory. *Proceedings and Adresses of the American Philosophical Association* 48: 5–22.
Rawls, John. 2003. *Justice as fairness: A restatement. 3rd printing*. Cambridge: Harvard University Press.
Rechenauer, Martin. 2003. *Philosophical and technical welfarism: Some important distinctions.* Unpublished Paper Manuscript. München, Ludwig-Maximilians-Universität.
Regan, Donald H. 1983. Against evaluator relativity: A response to Sen. *Philosophy and Public Affairs* 12(2): 93–112.
Rescher, Nicholas. 1966. *Distributive justice: A constructive critique of the utilitarian theory of distribution*. Indianapolis: The Bobbs-Merrill Company.
Resnik, Michael D. 1987/2000. *Choices: An introduction to decision theory.* Minneapolis: University of Minnesota Press.
Richardson, Henry S. 1997. *Practical reasoning about final ends*. Cambridge: Cambridge University Press.
Ridge, Michael. 2005. Review of pleasure and the good life by Fred Feldman. *Mind* 114(2): 414–417.
Ridge, Michael. 2011. Reasons for action: Agent-neutral vs. agent-relative. In *The Stanford encyclopedia of philosophy. Winter 2011 edition*, ed. Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information (Stanford University).
Robbins, Lionel. 1932/2007. *An essay on the nature and significance of economic science.* London: Macmillan and Co.
Robbins, Lionel. 1938. Interpersonal comparisons of utility: A comment. *The Economic Journal* 48(192): 635–641.
Roberts, Melinda A. 2002. A new way of doing the best that we can: Person−based consequentialism and the equality problem. *Ethics* 112(2): 315–350.
Roberts, Melinda A., and David T. Wasserman. 2009. Harming future persons: Introduction. In *Harming future persons: Ethics, genetics and the nonidentity problem*, International Library of Ethics, Law, and the New Medicine, vol. 35, ed. Melinda A. Roberts and David T. Wasserman, xiii–xxxviii. Dordrecht: Springer.
Roemer, John E. 1998. *Theories of distributive justice*. Cambridge, MA: Harvard University Press.
Rosebury, Brian. 1995. Moral responsibility and "Moral Luck". *The Philosophical Review* 104(4): 499–524.
Ross, William D. 1930/2002. *The right and the good,* ed. Philip Stratton-Lake. Oxford: Clarendon Press.
Ross, William D. 1939. *Foundations of ethics*. Oxford: Clarendon Press.
Russell, Bruce. 1999. Duty. In *The Cambridge dictionary of philosophy*, ed. Robert Audi, 248–249. Cambridge: Cambridge University Press.
Ryan, Alan. 1966. Mill and the naturalistic fallacy. *Mind* 75(299): 422–425.

Samuelson, Paul A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.
Sandberg, Joakim, and Niklas Juth. 2011. Ethics and intuitions: A reply to singer. *The Journal of Ethics* 15(3): 209–226.
Savage, Leonard J. 1954/1972. *The foundations of statistics.* 2nd ed. Dover Publications: New York.
Sayre-McCord, Geoffrey. 1985. Coherence and models for moral theorizing. *Pacific Philosophical Quarterly* 6: 170–190.
Sayre-McCord, Geoffrey. 1986. Deontic logic and the priority of moral theory. *Noûs* 20: 179–197.
Sayre-McCord, Geoffrey. 2001. Mill's "Proof" of the principle of utility: A more than half-hearted defense. *Social Philosophy and Policy* 18(2): 330–360.
Sayre-McCord, Geoffrey. 2011. Moral realism. In *The Stanford encyclopedia of philosophy. Summer 2011 edition*, ed. Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information (Stanford University).
Scanlon, Thomas M. 1973. Rawls' theory of justice. *University of Pennsylvania Law Review* 121(5): 1020–1069.
Scanlon, Thomas M. 1998. *What we owe to each other*. Cambridge, MA: Belknap Press of Harvard University Press.
Scanlon, Thomas M. 2001. Symposium on Amartya Sen's philosophy: 3. Sen and consequentialism. *Economics and Philosophy* 17(1): 39–50.
Scarre, Geoffrey. 1996. *Utilitarianism*. London: Routledge.
Scheffler, Israel. 1954. On justification and commitment. *The Journal of Philosophy* 51(6): 180–190.
Scheffler, Samuel. 1982/1994. *The rejection of consequentialism: Philosophical investigation of the considerations underlying rival moral conceptions.* 2nd Ed. Oxford: Clarendon Press.
Scheffler, Samuel. 1985. Agent-centred restrictions, rationality and the virtues. *Mind* 94(375): 409–419.
Schneewind, Jerome B. 1963. First principles and common sense morality in Sidgwick's ethics. *Archiv für Geschichte der Philosophie* 45: 137–156.
Schroeder, Mark. 2006. Not so promising after all: Evaluator-relative teleology and common-sense morality. *Pacific Philosophical Quarterly* 7: 348–356.
Schroeder, Mark. 2007. Teleology, agent-relative value, and 'Good'. *Ethics* 117: 265–295.
Schroth, Jörg. 2008. Distributive justice and welfarism in utilitarianism. *Inquiry* 51(2): 123–146.
Schroth, Jörg. 2009. Deontologie und die moralische Relevanz der Handlungskonsequenz. *Zeitschrift für Philosophische Forschung* 63(1): 55–75.
Schurz, Gerhard. 1997. *The is-ought problem: An investigation in philosophical logic*. Dordrecht: Kluwer.
Searle, John R. 1983. *Intentionality: An essay in the philosophy of mind*. Cambridge: Cambridge University Press.
Sen, Amartya Kumar. 1970a. *Collective choice and social welfare*. San Francisco: Holden-Day.
Sen, Amartya Kumar. 1970b. The impossibility of a Paretian liberal. *Journal of Political Economy* 78(1): 152–157.
Sen, Amartya Kumar. 1973/1997. *On economic inequality.* Oxford: Clarendon Press.
Sen, Amartya Kumar. 1977. On weights and measures: Informational constraints in social welfare analysis. *Econometrica* 45(7): 1539–1572.
Sen, Amartya Kumar. 1979. Utilitarianism and welfarism. *The Journal of Philosophy* 76(9): 463–489.
Sen, Amartya Kumar. 1980–1981. Plural utility. *Proceedings of the Aristotelian Society* 81: 193–215.
Sen, Amartya Kumar. 1982. Rights and agency. *Philosophy and Public Affairs* 11(1): 3–39.
Sen, Amartya Kumar. 1983. Evaluator relativity and consequential evaluation. *The Journal of Philosophy* 2(12): 113–132.
Sen, Amartya Kumar. 1985. *Commodities and capabilities*. Amsterdam: North-Holland.
Sen, Amartya Kumar. 1987/2004. *On ethics and economics.* Malden: Blackwell Publishing.

Sen, Amartya Kumar. 1993. Positional objectivity. *Philosophy and Public Affairs* 22(2): 126–145.
Sen, Amartya Kumar. 1997. Maximization and the act of choice. *Econometrica* 65(4): 745–779.
Sen, Amartya Kumar. 1997b. Interpersonal comparisons of welfare. In *Choice, welfare and measurement*, 264–282. Cambridge: Harvard University Press.
Sen, Amartya Kumar. 2000. Consequential evaluation and practical reason. *The Journal of Philosophy* 97(9): 477–502.
Sen, Amartya Kumar. 2009. *Social choice and individual values. Second annual Kenneth Arrow lecture.* Columbia University. http://vimeo.com/8244363. Accessed 11 Apr 2016.
Sen, Amartya Kumar. 2010. *The idea of justice*. London: Penguin.
Sen, Amartya Kumar, and James E. Foster. 1973. *On economic inequality*. Oxford: Clarendon Press.
Shaver, Robert. 1999. The appeal of utilitarianism. *Utilitas* 16(3): 235–250.
Shaver, Robert. 2010. Egoism. In *The Stanford encyclopedia of philosophy. Winter 2010 edition*, ed. Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information (Stanford University).
Shaw, William. 2006. The consequentialist perspective. In *Contemporary debates in moral theory*, ed. Dreier James, 5–20. Malden: Blackwell.
Sher, George. 1983. Antecedentialism. *Ethics* 94(1): 6–17.
Shrader-Frechette, Kristin. 1991. *Risk and rationality: Philosophical foundations for populist reforms*. Berkeley: University of California Press.
Shue, Henry. 2006. Torture in dreamland: Disposing of the ticking bomb. *Case Western Reserve Journal of International Law* 37(2–3): 231–239.
Sider, Theodore. 1993. Asymmetry and self-sacrifice. *Philosophical Studies* 70(2): 117–132.
Sidgwick, Henry. 1879. The establishment of ethical first principles. *Mind* 4(13): 106–111.
Sidgwick, Henry. 1907. *The methods of ethics*. London: Hackett Publishing.
Simon, Herbert A. 1955. A behavioral model of rational choice. *The Quarterly Journal of Economics* 69(1): 99–118.
Simon, Herbert A. 1959. Theories of decision-making in economics and behavioral science. *The American Economic Review* 49: 253–283.
Singer, Peter. 1972. Famine, affluence, and morality. *Philosophy and Public Affairs* 1(3): 229–243.
Singer, Peter. 1974. Sidgwick and reflective equilibrium. *The Monist* 58(3): 490–517.
Singer, Peter. 1979/1993. *Practical ethics.* Cambridge: Cambridge University Press.
Singer, Peter. 2005. Ethics and intuitions. *The Journal of Ethics* 9(3–4): 331–352.
Singer, Peter. 2009. *The life you can save: Acting now to end world poverty*. New York: Random House.
Sinnott-Armstrong, Walter. 2001. What is consequentialism? A reply to Howard-Snyder. *Utilitas* 13(3): 342–349.
Sinnott-Armstrong, Walter. 2006. *Moral skepticisms*. New York: Oxford University Press.
Sinnott-Armstrong, Walter. 2008. Framing moral intuitions. In *Moral psychology (volume 2): The cognitive science of morality: Intuition and diversity*, ed. Walter Sinnott-Armstrong, 47–76. Cambridge, MA: MIT Press.
Sinnott-Armstrong, Walter. 2011. Consequentialism. In *The Stanford encyclopedia of philosophy. Winter 2011 edition*, ed. Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information (Stanford University).
Skorupski, John. 1995. Agent-neutrality, consequentialism, utilitarianism . . . : A terminological note. *Utilitas* 7(1): 49–54.
Skorupski, John. 2000. *Ethical explorations*. New York: Oxford University Press.
Slote, Michael A. 1984. Satisficing consequentialism. *Proceedings of the Aristotelian Society* 58: 139–163.
Slote, Michael A. 1985a. *Common-sense morality and consequentialism*. London: Routledge & Kegan.
Slote, Michael A. 1985b. Utilitarianism, moral dilemmas, and moral cost. *American Philosophical Quarterly* 22(2): 161–168.

Slote, Michael A. 1989. *Beyond optimizing: A study of rational choice*. Cambridge: Harvard University Press.
Slote, Michael A. 1992/1995. *From morality to virtue.* New York: Oxford University Press.
Slote, Michael A. 1997/2007. Virtue ethics. In *Three methods of ethics: A debate*, ed. Marcia W. Baron, Philip Pettit, and Michael A. Slote. 10th print., 175–238. Malden: Blackwell.
Sluga, Hans. 2006. Family resemblance. *Grazer Philosophische Studien* 71: 1–21.
Smart, John Jamieson Carswell. 1956. Extreme and restricted utilitarianism. *The Philosophical Quarterly* 6(25): 344–354.
Smart, John Jamieson Carswell. 1973. An outline of a system of utilitarian ethics. In *Utilitarianism: For and against*, 3–74. Cambridge: Cambridge University Press.
Sorensen, Roy A. 1998. *Thought experiments*. Oxford: Oxford University Press.
Spielthenner, Georg. 2005. Consequentialism or deontology. *Philosophia* 33(1–4): 217–235.
Stevenson, Charles L.S. 1937. The emotive meaning of ethical terms. *Mind* 46(181): 14–31.
Stigler, George J. 1961. The economics of information. *Journal of Political Economy* 69(3): 213–225.
Stocker, Michael. 1969. Consequentialism and its complexities. *American Philosophical Quarterly* 6(4): 276–289.
Stroud, Sarah. 1998. Moral overridingness and moral theory. *Pacific Philosophical Quarterly* 70: 170–189.
Sumner, Leonard W. 1987. *The moral foundation of rights*. Oxford: Clarendon Press.
Sumner, Leonard W. 1996. *Welfare, happiness, and ethics*. Oxford: Oxford University Press.
Suzumura, Kotaro. 2001. Pareto principles from inch to ell. *Economic Letters* 70: 95–98.
Sverdlik, Steven. 1996. Motive and rightness. *Ethics* 106(2): 327–349.
Tännsjö, Torbjörn. 1995. Blameless wrongdoing. *Ethics* 106(1): 120–127.
Tännsjö, Torbjörn. 2002. *Understanding ethics: An introduction to moral theory*. Edinburgh: Edinburgh University Press.
Tännsjö, Torbjörn. 2011. *Shalt thou sometimes murder? On the ethics of killing*. Uppsala: Unpublished Book Manuscript.
Taurek, John M. 1977. Should the numbers count? *Philosophy and Public Affairs* 6(4): 293–316.
Taylor, Charles. 1995. *Philosophical arguments*. Cambridge: Harvard University Press.
Temkin, Larry S. 1993. *Inequality*. Oxford: Oxford University Press.
Tersman, Folke. 2008. The reliability of moral intuitions. *Australasian Journal of Philosophy* 86(3): 389–405.
Thomann, Marius. 2010. *Die Logik des Könnens*. Berlin: Logos Verlag.
Thompson, Leigh, and George Loewenstein. 1992. Egocentric interpretations of fairness and interpersonal conflict. *Organisational Behaviour nd Human Decision Processes* 51: 176–197.
Thomson, Judith Jarvis. 1976. Killing, letting die, and the trolley problem. *The Monist* 59(2): 204–217.
Thomson, Judith Jarvis. 1985. The trolley problem. *The Yale Law Journal* 94(6): 1395–1415.
Thomson, Judith Jarvis. 1990. *The realm of rights*. Cambridge: Harvard University Press.
Thomson, Judith Jarvis. 1993. Reply to commentators. *Philosophy and Phenomenological Research* 53(1): 187–194.
Thomson, Judith Jarvis. 2008. Turning the trolley. *Philosophy and Public Affairs* 36(4): 359–374.
Timmermann, Jens. 2007. *Kant's groundwork of the metaphysics of morals: A commentary*. Cambridge: Cambridge University Press.
Timmons, Mark. 2002. *Moral theory: An introduction*. Maryland: Rowman & Littlefield.
Timmons, Mark. 2008. Toward a sentimentalist deontology. In *Moral psychology (volume 3): The neuroscience of morality: Emotion, brain disorders, and development*, ed. Walter Sinnott-Armstrong, 93–104. Cambridge, MA: MIT Press.
Trapp, Rainer W. 1988. *Nicht-klassischer Utilitarismus: Eine Theorie der Gerechtigkeit*. Frankfurt am Main: V. Klostermann.
Tungodden, Bertil. 2003. The value of equality. *Economics and Philosophy* 19: 1–44.

Tungodden, Bertil, and Peter Vallentyne. 2007. Who are the least advantaged? In *Egalitarianism: New essays on the nature and value of equality*, ed. Nils Holtug and Kasper Lippert-Rasmussen, 174–195. Oxford: Clarendon Press.

Turri, John. 2005. You can't get away with murder that easily: A response to Timothy Mulgan. *International Journal of Philosophical Studies* 13(4): 489–492.

Tversky, Amos, and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science* 211(4481): 453–458.

Unger, Peter. 1996. *Living high and letting die*. Oxford: Oxford University Press.

Urmson, James O. 1958. Saints and heroes. In *Essays in moral philosophy*, ed. A.I. Melden, 198–216. Washington, DC: University of Washington Press.

Vallentyne, Peter. 1987. The teleological/deontological distinction. *The Journal of Value Inquiry* 21: 21–32.

Vallentyne, Peter. 1988. Teleology, consequentialism, and the past. *The Journal of Value Inquiry* 22: 89–101.

Vallentyne, Peter. 1989. Two types of moral dilemmas. *Erkenntnis* 30(3): 301–318.

Vallentyne, Peter. 2006. Against maximizing act-consequentialism. In *Contemporary debates in moral theory*, ed. Dreier James, 21–37. Malden: Blackwell.

Vallentyne, Peter. 2007. Of mice and men: Equality and animals. In *Egalitarianism: New essays on the nature and value of equality*, ed. Nils Holtug and Kasper Lippert-Rasmussen, 211–237. Oxford: Clarendon Press.

Vallentyne, Peter, and Shelly Kagan. 1997. Infinite value and finitely additive value theory. *The Journal of Philosophy* 94(1): 5–26.

van Roojen, Mark. 2004. The plausibility of satisficing and the role of good in ordinary thought. In *Satisficing and maximizing: Moral theorists on practical reason*, ed. Michael Byron, 155–175. Cambridge: Cambridge University Press.

Vermazen, Bruce. 1985. Negative acts. In *Essays on Davidson: Actions and events*, ed. B. Vermazen and M.B Hintikka, 93–104. Oxford: Clarendon Press.

von Neumann, John, and Oskar Morgenstern. 1944/1955. *The theory of games and economic behavior.* Third Edition. Princeton: Princeton University Press.

von Wright, George Henrik. 1963. *Norm and action: A logical enquiry.* London: Routledge & Kegan Paul.

Wall, Edmund. 2000. The problem of group agency. *The Philosophical Forum XXXI* 2: 187–197.

Walsh, Vivien C. 1996. *Rationality, allocation, and reproduction*. Oxford: Clarendon Press.

Wellman, Carl. 1972. Ethics since 1950. *The Journal of Value Inquiry* 6(2): 83–90.

Williams, Bernard Arthur Owen. 1970. The self and the future. *The Philosophical Review* 79(2): 161–180.

Williams, Bernard Arthur Owen. 1973. A critique of utilitarianisms. In *Utilitarianism: For and against*, 75–155. Cambridge: Cambridge University Press.

Williams, Bernard Arthur Owen. 1981. Persons, character and morality. In *Moral luck: Philosophical papers, 1973–1980*, 1–19. Cambridge: Cambridge University Press.

Williams, Bernard Arthur Owen. 1985. *Ethics and the limits of philosophy*. London: Fontana Press.

Williamson, Timothy. 2000. *Knowledge and its limits*. Oxford: Oxford University Press.

Wilson, George. 2009. Action. In *The Stanford encyclopedia of philosophy. Fall 2009 edition*, ed. Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information (Stanford University).

Wittgenstein, Ludwig. 1953/1986. *Philosophical investigations: Translated by G.E.M. Anscombe.* Oxford: Basil Blackwell.

Wittgenstein, Ludwig. 1960. *Preliminary studies for the "Philosophical investigations": Generally known as the Blue and Brown Books*. New York: Harper.

Wood, Allen W. 2008. *Kantian ethics*. Cambridge: Cambridge University Press.

Wood, Allen W. 2011. Humanity as an end in itself. In *On what matters II*, ed. Derek Parfit, 58–82. Oxford: Oxford University Press.

Woodruff, Paul. 2010. Plato's shorter ethical works. In *The Stanford encyclopedia of philosophy. Summer 2010 edition*, ed. Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information (Stanford University).
Wright, Robert. 1996. *The moral animal: The new science of evolutionary psychology*. London: Abacus.

Index

N. Mukerji, *The Case Against Consequentialism Reconsidered*,
Theory and Decision Library A: 51, DOI 10.1007/978-3-319-39249-3