Tomasz Herok

Biting the bullet. Counterintuitive conclusions vs the myth of relying on intuition in contemporary ethics.

This thesis is submitted for the degree of Doctor of Philosophy.

Lancaster University

Department of Politics, Philosophy and Religion

July 2023

Abstract

According to the current methodological orthodoxy, intuitions are relied on, appealed to, or used as evidence in ethics. My main point is that this claim, as commonly understood, is false – it refers to a practice that simply does not exist. In Chapter 1 I explain what the orthodoxy is, what it is not, and how to test it. In Chapter 2 I examine seven arguments in favour of the orthodoxy, and find all of them wanting. In Chapter 3 I introduce a new argument against the orthodoxy: the argument from counterintuitive conclusions. The idea behind it is that because philosophers routinely dismiss intuitions, intuitions cannot be treated as evidence. To this it can be replied that it is not impossible to dismiss intuitions and rely on them at the same time. I therefore describe eight theories that allow for such reconciliation. In Chapter 4 I examine three case studies – Michael Tooley's defence of infanticide, David Benatar's defence of antinatalism and John Taurek's attack on the idea of moral quantification – to show that none of the reconciliation theories works in practice. In Chapter 5 I discuss arguably the most significant practical consequence of the orthodoxy: experimental philosophy. I argue that since intuitions are never used as evidence, the project is largely pointless. In the final chapter I explain what is new in the thesis and describe differences between myself and others who have challenged the orthodoxy before.

Declaration

I declare that this thesis is my own work and that it has not been previously submitted for the award of a higher degree elsewhere. I confirm that the main text does not exceed the prescribed limit of 80,000 words. Parts of chapters 1 and 3 have been published, with modifications, as:

Herok, T. Intuitions are never used as evidence in ethics. *Synthese* 201, 42 (2023). https://doi.org/10.1007/s11229-022-04031-z .

Table of Contents

PREFACE

Philosophy is often advertised as a discipline whose goal is to question *everything*. It does not matter how deeply certain views are held, how emotionally attached we are to them, how undeniably true they seem, how important we think they are for the functioning of the society, or how bizarre or gloomy the world would be if they turned out to be false. Philosophers subject every view to scrutiny and if they find no grounds for holding it, they reject it. But one can also come across a different, equally popular picture. According to it, there is a distinct class of judgments, usually called *intuitions*, that philosophers try to explain, appeal to, account for, or use as evidence in their theories. This is not to say that these judgments cannot under any circumstances be dismissed, however it is clear they have a special evidential status. The clash between these two pictures seems obvious: surely philosophy is either about questioning everything indiscriminately or about accounting for our intuitions.

Many academic philosophers seem to endorse both pictures, or at least think the pictures are compatible with each other. One might think that there is a straightforward solution to the problem. But this, as I have learned, is simply not the case. My search for the solution eventually grew into a full-fledged dissertation project. The outcome is the conclusion that – at least in the area of contemporary ethics, but most likely outside it as well – intuitions are not treated as something to be explained by a theory. The intuition-centred view is simply a widespread myth.

I am hardly the first one to make this point. The intuition-centred view is still prevalent, however in recent years several dissidents have argued against it. The standard approach of their work is to examine the alleged paradigm cases of relying on intuitions in philosophy and demonstrate that no intuitions are in fact treated as evidence in a given text. My approach, however, is different: in addition to examining some of the paradigm cases, I look into how intuitions end up being dismissed in conclusions of philosophical arguments. In other words, the main question of my project is: “If philosophers use intuitions as evidence, why are their conclusions so counterintuitive?”

The structure of the thesis is as follows. In chapter 1, “Intuitions as evidence”, I try to determine what philosophers mean when they talk about explaining intuitions, appealing to intuitions, relying on intuitions etc. I argue that the typical understanding is what is sometimes called “descriptive

evidentialism": the idea that intuition-states are treated as evidence of their propositional contents in the context of justification. I then argue that descriptive evidentialism is false – on any account of what intuitions are. That said, I admit that philosophers can rely on intuitions to clarify, persuade, discover, or to support things other than the intuitions' contents. In chapter 2, "Defending the orthodoxy", I offer replies to seven arguments for descriptive evidentialism. In chapter 3, "The argument from counterintuitive conclusions", I discuss eight hypotheses that reconcile the fact that philosophers dismiss intuitions with the claim that philosophers rely on intuitions as evidence. In chapter 4, "Case studies", I test my reconciliation hypotheses against concrete examples of philosophical practice, and conclude that none of them involves any appeals to intuition in the typical sense. In chapter 5, "Experimental philosophy", I contend that one of the most significant practical consequences of endorsing the intuition orthodoxy has been the birth of experimental philosophy. I argue that because it rests on a mistake, the project is largely pointless. In the final chapter I summarise my points and explain how my position differs from that of others who have rejected the intuition dogma.

CHAPTER 1. Intuitions as evidence

1. The dogma

The view that philosophy somehow relies on intuitions has become a prominent part of the profession's self-image in recent decades. Here are some examples of how it is expressed:

> One of the favourite argumentative methods of present-day analytic philosophers is to appeal to intuitions. (Hintikka 1999, p. 127)

> One thing that distinguishes philosophical methodology from the methodology of the sciences is its extensive and avowed reliance on intuition. (Goldman 2007, p. 1)

> Most philosophers do it openly and unapologetically, and the rest arguably do it too, although some of them would deny it. What they all do is appeal to intuitions in constructing, shaping, and refining their philosophical views. (Kornblith 1998, p. 129)

> We ask philosophical intuitions – what we would say or how things seem to us to be – to do a lot of work for us. We advance philosophical theories on the basis of their ability to explain our philosophical intuitions, defend their truth on the basis of their overall agreement with our philosophical intuitions, and justify our philosophical beliefs on the basis of their accordance with our philosophical intuitions. (Alexander 2012, p. 1)

> Contemporary work in philosophy is shot through with appeals to intuition. When a philosopher wants to understand the nature of knowledge or causation or free will, the usual approach is to begin by constructing a series of imaginary cases designed to elicit prereflective judgments about the nature of these phenomena. These prereflective judgments are then treated as important sources of evidence. (Knobe et al. 2012, p. 82)

> Philosophers frequently appeal to intuitions in constructing and arguing for philosophical theories. A theory is commonly judged lacking when it fails to "capture" our intuitions and judged acceptable insofar as it captures more of our intuitions than other theories. (Bealer 1998, p. 179)

> Analytic philosophers frequently appeal to intuitions. In the method of cases, vivid scenarios elicit intuitive responses that speak directly for or against a philosophical claim. (De Cruz 2015, p. 233)

> From Plato to the present, philosophers have relied on intuitive judgments as evidence for or against philosophical theories. (Stich 2010, p. 110)

> From its beginning in Greek philosophy right through to the present, intuitions have always played an extremely important role in Western philosophy. Of course, the use of intuitions has been criticized from time to time, but in spite of the criticisms, philosophers have continued to rely heavily on intuitive judgments in pretty much the way they always have. (Gutting 1998, p. 8)

> Most philosophers who cut their teeth on Russell and Wittgenstein would rather not have to rely on intuitions, but there is no clear and compelling alternative. (Miller 2000, p. 231)

Many single out *ethics,* sometimes as the area where the practice of appealing to intuitions is particularly important:

> The most common method in normative ethics is piecemeal appeal to intuition. 'It follows from what you say that it would be all right to do such and such, but that's counter-intuitive, so you're wrong.' (Griffin 1988, p. 1)

> Anyone who reflects on the way we go about arguing for or against moral claims is likely to be struck by the central importance we give to thinking about cases. Intuitive reactions to cases—real or imagined—are carefully noted, and then appealed to as providing reason to accept (or reject) various claims. (Kagan 2001, p. 44)

> Philosophers these days frequently elicit "our intuitions" about this or that and appeal, implicitly or explicitly, to our feelings and sentiments, and to moral consensus. They invent imaginary cases and tell us bizarre stories which are intended to illuminate these intuitions. Pick up any recent journal or Moral Problems anthology, and it seems as if everyone is going about ethics in a similar way. (Shaw 1980, p. 127)

> The appeal to intuitions is a pervasive strategy in contemporary philosophical discourse. A good philosophical theory is widely taken to be one that gives an adequate account of our intuitions. Ethical theory is no exception. (Audi 1993, p. 295)

> Many moral theorists have relied on intuitions in both building up and challenging theories. (Kamm 2007, p. 425)

> Many contemporary ethicists like to treat moral intuitions as evidence, akin to experimental data that are to be explained by theories. (Thagard 2010, p. 202)

> Ethicists often appeal to moral intuitions in defending a theory. In this practice, the contents of intuitions are taken to support moral beliefs in a way that is often compared to the way the contents of perception support empirical beliefs. (Kauppinen 2014, p. 169)

> In the sciences, we decide between theories on the basis of observations, which have an important degree of objectivity. It appears that in moral reasoning, moral intuitions play the same role which observations do in science: we test general moral principles and moral theories by seeing how their consequences conform (or fail to conform) to our moral intuitions about particular cases. (Boyd 1988, p. 184-5)

> It is hard to imagine any way to develop a moral theory without relying on moral intuitions at all. How could you choose among consequentialism, Kantianism, contractarianism, and virtue theories without appealing to moral intuitions at some point in some way? (Sinnott-Armstrong et al. 2010. p. 246)

> In the end, all ethicists appeal to intuition. They can do no other. (Bedke 2008, p. 266)

I am going to argue that all these claims, as well as countless similar ones, are false: they all refer to a non-existent practice. In this chapter I will explain what exactly philosophers mean by relying on intuitions, appealing to intuitions, using intuitions as evidence etc. and identify a number of problems with this view, drawing on the work of several dissidents who have recently challenged it. I will also argue that there are several interpretations of "relying on intuitions in philosophy" which are most likely true – but they they are substantially different from what philosophers have in mind when they make assertions like the ones I have just listed.

2. Separateness of ethics

Should we think of the role of intuitions in moral philosophy as fundamentally different from that of other philosophical disciplines? Or perhaps there are reasons to think that *moral* intuitions, as opposed to other intuitions, are treated differently by philosophers? Some have made suggestions to that effect. For example, Brian Weatherson writes that "in epistemology, particularly in the theory of knowledge, and in parts of metaphysics, particularly in the theory of causation, it is almost universally assumed that intuition trumps theory", whereas "matters are quite different in ethics", (Weatherson 2003, p. 1). Something of a mirror image of this view has been endorsed by James Griffin, who writes that "in (...) other branches of philosophy, finding a conclusion intuitively

repugnant does not close an argument; it is a reason to start looking for a good argument" (Griffin 1988, p. 2).

John Mikhail argues there exists "Universal Moral Grammar", analogous to Universal Grammar first proposed by Noam Chomsky in 1960s (Mikhail 2013). According to Mikhail, studying moral intuitions might help us reveal the underlying structures of our innate moral knowledge, just like studying linguistic intuitions helps reveal the underlying structures of our innate linguistic knowledge. However it does not seem that a similar analogy can be drawn between linguistics and any of the other philosophical disciplines.

Furthermore, many believe that there is a widespread intuition-based method of moral philosophy – "the method of reflective equilibrium". As Norman Daniels points out, "despite the fact that the origins of reflective equilibrium (minus the name) lie in mid-twentieth century discussions about the justification of inductive logic, its principal development through the rest of the century lies primarily in ethics and political philosophy" (Daniels 2020). Today the method is rarely mentioned in the context of epistemology, metaphysics, philosophy of language, or philosophy of mind – which may suggest that even if these disciplines rely on intuition too, it must be a different kind of reliance. Finally, several philosophers have argued that there are two methodologies of ethics: one intuition-based, and one intuition-free (Brandt 1979, Unger 1996, McMahan 2013). However none of them seems to believe that this dualism can be extended to philosophy in general.

If the separateness thesis is true in some form, one should be careful not to jumble together ethics and other philosophical disciplines while examining claims about the role of intuitions. On the other hand, most proponents of the intuition-centred view do not differentiate between ethics and the rest of philosophy. To try to do justice to both groups, I have decided to adopt a compromise approach. For the most part, I am going to write about philosophy in general, using examples from across the disciplines – typically examples provided by the proponents of the intuition-centred view themselves. I am also going to give reasons for thinking that the separateness thesis is false: intuitions are consistently treated as irrelevant in ethics just like in any other philosophical discipline. However in chapter 4 I am going to examine the way in which philosophers reach counterintuitive conclusions *only in contemporary ethics*. I believe that examining analogous cases from epistemology, metaphysics, philosophy of mind etc. would yield similar results – namely that dismissing intuitions never involves appealing to other intuitions. But someone unconvinced by my arguments against separateness will not take my word for it. The main target of this thesis is therefore the view that intuitions are standardly relied on in contemporary ethics. My case against the broader view can be seen as somewhat weaker.

3. State vs content

Until recently the view that philosophers rely on intuitions did not even have a name, which is not very surprising given how universally accepted it was. Currently the most common term in the literature seems to be "Centrality" coined by Herman Cappelen. Here is how Cappelen defines it:

> **Centrality (of Intuitions in Contemporary Philosophy):** Contemporary analytic philosophers rely on intuitions as evidence (or as a source of evidence) for philosophical theories. (Cappelen 2012, p. 3)

What is the difference between "evidence" and "source of evidence"? Some have pointed out that "relying on an intuition that *p* as evidence" is ambiguous between "relying on *p*, which is intuitive, as evidence" and "relying on the fact that *p* is intuitive as evidence". In other words, what constitutes evidence can be intuitions in the propositional content sense or intuitions in the mental state sense (Molyneux 2014, p. 443, Deutsch 2015, pp. 35-9). Proponents of Centrality sometimes endorse the latter, and sometimes the former – but with an addition that intuition-states then serve as a source of evidence (Cappelen 2012, p. 13).

Both the propositional and the mental state reading can also be endorsed simultaneously. For example, George Bealer argues that there is an element of philosophical inquiry when the propositional content is used as evidence and an element when the fact of intuiting this content is used as evidence (Bealer 1998, p. 205). What I am going to object to is the intuiting element only, as I believe that philosophers often rely on propositions which merely *happen* to be intuitive. Consider the statement: "Contemporary analytic philosophers rely on propositions formulated by carbon-based life forms as evidence for their theories". It sounds odd, as it is *pragmatically inappropriate*: there is little point of bringing up the chemical composition of creatures who formulate propositions in the context of discussing philosophical evidence. However, pragmatics aside, the statement is not false. In my view, the statement "Contemporary analytic philosophers rely on intuitive propositions as evidence for their theories" has a similar status. It may be odd or unhelpful, but it is not false.

Cappelen's take on this issue is somewhat different – he does not believe that philosophers rely on intuitions in the propositional content sense. This is because of his scepticism about what intuitions, as philosophers use the term, are. I do not find his argument persuasive, but also, and more importantly, I do not find it *necessary*. The only reason why Bealer and others say things like

"contents of intuitions count as evidence" is that they believe that the state of being intuitive also counts as evidence. As I am going to argue, when the latter is rejected, uttering the former becomes pragmatically inappropriate in most contexts, and it does not matter whether it is true, false or neither – at least not to someone interested in the problem of philosophical evidence. I do, however, agree with Cappelen that those who prefer to speak of intuition-states as *a source of evidence* (like, for example, Nado 2017) are mistaken. And like Cappelen, for brevity's sake I am also going to refer to what they call "a source of evidence" simply as "evidence". My reasons to think that the distinction is not substantial, as well as reasons to reject Cappelen's scepticism, are going to be explained in detail later in this chapter.

4. Own content vs other content

To be more precise, I am only going to attack one version of the mental state interpretation. Note that "relying on the fact that *p* is intuitive as evidence" is itself ambiguous between "relying on the fact that *p* is intuitive as evidence for *p*" and "relying on the fact that *p* is intuitive as evidence for *q*". It is only the former interpretation that I am going to object to.

For example, Robert Nozick in his famous thought experiment asks whether you would plug into a machine that could give you any experience you wanted, indistinguishable from experiencing reality (Nozick 1974, p. 42). He argues that for many people their "first impulse" is to say no, even if they might later change their mind upon reflection (Nozick 1989, p. 105). I do not want to deny that Nozick is using an intuition as evidence against psychological hedonism (the view that all that motivates us is pleasure). Philosophers occasionally evaluate psychological claims and appealing to people's intuitions – in the form of "it is intuitive that *p*, therefore *q*" – may be a way to do it. Nor do I want to deny that Nozick also tries to provide evidence against *ethical* hedonism (the view that all that matters is pleasure). My quarrel is with the idea that Nozick is offering the intuition about the experience machine as evidence against ethical hedonism. On this reading the intuition is used as evidence *for its content*: one should not plug oneself into the machine because it is intuitive that one should not plug oneself into the machine. This is how most commentators think relying on intuitions as evidence works in this case (see Hewitt 2010, Weijers 2014, Rowland 2017). Later in the chapter I am going to examine further examples in more detail to show that this understanding is virtually universal.

Bernard Molyneux has put forward a definition of the intuition dogma that is free of the two ambiguities I have just discussed:

> [intuitions] are standardly *treated* as evidence of their contents, whether or not it is right to do so (Molyneux 2014, p. 441).

The contents are then, of course, used as evidence for and against philosophical theories. Molyneux calls this view "descriptive evidentialism" (as he contrasts it with *normative* evidentialism, according to which intuitions *are* evidence). Admittedly, the term has not gained much popularity in the literature. Nevertheless, as I find Molyneux's definition more accurate, I am going to prefer his term over the more widespread "Centrality". I will attempt to show that descriptive evidentialism (henceforth "DE") is the assumption behind most assertions about intuitions being relied on, accounted for, appealed to, deferred to, trusted, invoked, captured, matched, accommodated, systematised, explained, employed, used or treated as evidence in philosophy.

Some might be tempted to remove the word "standardly" from Molyneux's definition and argue that at least this weaker version of DE is correct: perhaps the practice of treating intuitions as evidence of their contents is merely occasional, niche, or unorthodox. I am going to challenge this claim too, and "DE", as I use it, is going to refer to both the weak and the strong, "standard practice" version. In my view, barring one fairly recent exception I will discuss in more detail in chapter 5, intuitions are simply *never* treated as evidence of their contents in philosophy. Granted, philosophy is vast and diverse, and one can only be familiar with a tiny fraction of what has been published. Some methods, however, seem too off the mark to be even considered rare. We can be justified in believing that divinations from the entrails of sacrificed animals are never treated as evidence of their contents in philosophy. I will argue that DE, including its weaker variety, does not fare any better.

5. The nature of intuitions

What exactly are intuitions, according to proponents of DE? Recently Nevin Climenhaga has offered the following definition:

> I take intuitions to be mental states that we find ourselves in when considering particular propositions. I take it that when one has an intuition that P:
>
> (i) it seems to one that P;
>
> (ii) this seeming is not the conscious result of an inference;
>
> (iii) this seeming is not the conscious result of an apparent memory that P, a sensorial experience as of P, or someone else's testimony that P. (Climenhaga 2018, p. 69-70)

This is a fairly general account. Some want to be more specific, and some more exclusive. While virtually everyone agrees that intuition is a propositional attitude – that is some sort of relation between an agent and a proposition – there is a deal of controversy over what kind of propositional attitude it is. Three main competing options are: a kind of belief, a kind of inclination, or a disposition, to believe and a *sui generis* attitude. None is without difficulties. Opponents of the first argue that certain probability puzzles, like the Monty Hall problem, show it is not only possible to have an intuition that *p* without believing that *p*, but also to have an intuition that *p* while believing that not-*p*, which does not bode well for the belief theory. Opponents of the second option often argue it fails to account for the occurrent and episodic nature of intuitions (Pust 2000, pp. 39-43). Opponents of the third deny that intuitions must always be occurrent – to think they are is to commit the "refrigerator-light fallacy", that is to "confuse that which is always the case when you are looking with that which is always the case" (Earlenbaugh & Molyneux 2009, p. 103).

Those who agree that intuitions are occurrent and episodic often add that they must also be spontaneous, or immediate – an intuitive episode cannot develop in a gradual way (Goldman & Pust 1998, p. 179). Sometimes they also argue that they must be accompanied by a special phenomenology: there is *something it is like* to have an intuition, intuitions *seem true* is a particular way (Bealer 1998, p. 207, Chudnoff 2013, pp. 32-40). Bealer argues that any intuition used as evidence in philosophy has a specific kind of content: it "presents itself as necessary; it seems that things could not have been otherwise" (Bealer 1999, p. 30). This is not true of any intuitive content. For example, in Newton's famous thought experiment we are asked to imagine a bucket partly filled with water, spinning in an otherwise empty space. It seems to us that water would creep up the side of the bucket, but not that this is *necessarily* the case. We are therefore dealing with what Bealer calls a "physical intuition", which is not something philosophers typically rely on (Bealer 1998, p. 205).

Intuitions are sometimes believed to be judgments generated by a special *faculty of intuition*, a sort of sixth sense. This view is often associated with the so-called ethical intuitionists, such as Henry Sidgwick, G. E. Moore or W. D. Ross. However there has been some controversies over how the faculty view should be interpreted and, consequently, whether different intuitionists actually subscribed to it (Stratton-Lake 2002, Crisp 2002). Another view associated with ethical intuitionism is that intuitive judgments are *self-evident*: they are justified simply by being understood, and no further justification for them can or needs to be offered. Unfortunately there seems to be little agreement on which particular judgments are self-evident, which means that neither this nor the faculty view can be fruitfully used as a criterion for distinguishing intuitions.

Some philosophers, like David Lewis and Peter van Inwagen, tend to be much more inclusive and allow inferential, and, for that matter, *any* beliefs to be classified as intuitions (Lewis 1983, p. x, Van Inwagen 1997, p. 309). In addition to restrictive accounts (like that of Bealer), moderately inclusive accounts (like that of Climenhaga) and broadly inclusive accounts (like that of Lewis) we can also distinguish idiosyncratic accounts that reject all criteria listed above and introduce other criteria instead. The most prominent account of that sort is arguably the one offered by John Rawls, who argues that while intuitions can be consciously inferred from other claims, they cannot be consciously inferred from *ethical principles* (Rawls 1951, p. 183).

There is a lot more that can be said about the nature of intuitions according to different philosophers, however further discussion would be largely pointless. This is because in my view DE is false *irrespective of which account is adopted*, including the most liberal ones: it is not the case that the fact that *p* is non-inferential is used as evidence for *p*, it is not the case that the fact that *p* is partly non-inferential is used as evidence for *p*, it is not the case that the fact that *p* is believed is used as evidence for *p*, and so forth. Proponents of DE often argue that if intuitions are understood narrowly then perhaps they are not used as evidence in philosophy, however on a less restrictive understanding they clearly are used as evidence (Chalmers 2014, Bengson 2014, Stich & Tobia 2016, p. 8). They accuse critics of DE like Cappelen of setting up a straw man: supposedly his way to question the practice of relying on intuitions in philosophy is to put a number of unreasonable qualifications on the nature of the intuitive. I think this is a misunderstanding: the reason why accusations like this are made is that DE is often conflated with something else – I will explain it in more detail later in this chapter.

I mentioned that Cappelen rejects not only the idea of relying on intuitions in the mental state sense, but also of relying on intuitions in the propositional content sense. This is because of his scepticism about the very existence of intuitions. Supposedly the way philosophers use "intuition" and cognate terms is not far from gibberish – sentences containing these terms often fail to express propositions that could be true or false. This is not to say that "intuition" in everyday English, or non-philosophical technical English, is an element of a semantically defective discourse. Sentences like "The new operating system lacks an intuitive interface" or "Her intuition told her something was wrong" can be perfectly meaningful in their respective contexts. However sentences like "contents of intuitions are used as evidence in philosophy" are problematic, as it is hard to tell what "intuitions" refer to. What makes Cappelen think that? He argues that philosophical intuition-discourse does not meet certain meaningfulness criteria:

> There is no agreed upon definition of 'intuition'. There are no agreed upon paradigms. There is minimal unity in usage between different schools and subdisciplines and there is no group of experts within the discipline who agree on how the term should be used. (Cappelen 2012, p. 52)

To me, however, this seems exaggerated. I do not wish to question the adequacy of Cappelen's criteria, but rather his claim that the criteria are not met. I think there is an agreed upon definition of "intuition" – for example, Climenhaga's moderately inclusive definition is something most proponents of DE would subscribe to. It might not be the most precise, however it still allows us to easily exclude a fair number of judgments from the realm of the intuitive. There are also agreed upon paradigms – I am going to discuss them later in this chapter. Of course occasionally an eccentric member of the philosophical community would opt for an entirely different definition or reject the paradigms, but this is no reason to think that the entire intuition-discourse is flawed in some fundamental way.

Moreover, Cappelen suspends his own scepticism in the second part of his book where he adopts a particular account of the intuitive and analyses a number of particular paradigm cases to show that DE is false. His choice is not arbitrary: he focuses on "features that, according to at least a fairly wide range of intuition-theorists, are characteristic of appeals to the intuitive" (Cappelen 2012, p. 111). A hard-line sceptic would argue that no such features exist. He also suspends his scepticism in the first part of the book to address what the calls "the argument from intuition-talk" – he argues that expressions like "intuitively" in philosophical texts refer to something tangible, however never to anything that could be best explained by DE. In the next chapter I am going to give additional reasons to think that this idea is correct. If we were instead to argue that whenever philosophers use intuition-talk, they gibber, the reply would lose much of its force.

6. The nature of evidence

So much for what exactly is used as evidence according to the view I will attempt to refute. We can now ask: what exactly does it mean to *treat something as evidence*, according to this view? In most cases proponents of DE do not specify how "evidence", let alone "treating as evidence", should be understood. Critics of DE tend to assume that it is meant to be true irrespective of which specific theory of evidence is adopted and that it is possible to prove it false in a similarly theory-neutral way (Cappelen 2012, pp. 11-12, Molyneux 2014, p. 443). On the other hand, Climenhaga suggests we should be more specific. He proposes to understand DE along Bayesian lines:

> E is evidence for T relative to background knowledge K iff P (TjE&K) > P(TjK) – that is, E raises the probability of T relative to K. A person takes E to be evidence for T or uses it as evidence relative to K iff his conditional credence in T given E&K is greater than his conditional credence in T given K. (Climenhaga 2018, p. 71)

This, however, strikes me as too author-oriented. First, philosophical writings are rarely framed as reports of their authors' psychology, let alone estimates of fictions such as authors' conditional credences in propositions. It is unclear how we can learn much about what philosophers could believe given that something is true, based on what they write.

Second, on rare occasions when philosophers do comment on how their evidence influence their mental states, they do not necessarily confirm Climenhaga's view. Take William Lane Craig, who puts forward several arguments for the existence of God, most notably the so-called Kalam cosmological argument. Craig confesses that even if he were fully convinced that all his arguments were unsound, it would not diminish his belief in God one iota "because of the self-authenticating witness of God's Spirit who lives within him" (Craig 2008, p. 46). Should we conclude that what Craig explicitly calls evidence for the existence of God is not used by him as evidence for the existence of God? Or perhaps that he must be wrong about his own beliefs?

Philosophers can be similarly attached not only to their religious beliefs, but to all sorts of philosophical beliefs. Think of Elizabeth Anscombe's remark about not wanting to argue with someone who thinks a judicial execution of an innocent person can be justified, as anyone who believes it "shows a corrupt mind" (Anscombe 1958, p. 17). It seems perfectly possible to come up with reasons against executing an innocent person and treat them just as Craig treats his reasons to think God exists: as something that does not strengthen one's own belief that *p*, and yet supports *p* in one's published work.

There are also philosophers who do not appear to find their own arguments compelling in any way. William Lycan writes that if God offered him to bet on a doctrine he "would kill and die for" in his publications, he would not take the bet, even if the stake were only $10 (Lycan 2013, p. 115). Keith DeRose writes that if aliens who knew solutions to philosophical problems threatened him to destroy the Earth and entire humankind if he did not answer their philosophical question correctly, he would be more likely to go with the profession's majority view rather than the view he defends "when discussing the matter in a philosophical setting" (DeRose 2017, p. 267-9). Apparently this kind of scepticism is not the outcome of disbelieving one's premises or taking one's own arguments to be invalid. Lycan and DeRose might of course be wrong about what they would do in such

outlandish circumstances. My point is not, however, that they must be right, but rather that there exists something they use as evidence when they do philosophy which is independent of how it influences their own beliefs.

In reply Climenhaga might argue that people like Craig, Lycan and DeRose are in fact outliers as most philosophers believe what they preach, and they believe it on the basis of their own evidence. This would mean that his definition can at least be used as a sort of rule of thumb for determining what is treated as evidence in philosophy. But even this is problematic. Climenhaga mentions a distinction between private evidence and public evidence: the latter consists of reasons to accept a claim *offered in a public discussion* (2018, p. 98). For some reason, however, he is not troubled by the fact that his approach blurs the line between the two. I think this is a mistake. Philosophy is, after all, a public endeavour. A philosopher might believe that *p* for a number of reasons, and she might publish an argument which relies on *p* as one of the premises. This, however, does not mean that all her reasons to accept *p* are automatically used as philosophical evidence. To count as such, they must be *appealed to* in what is published. It might be the case that some philosophers, perhaps even numerous philosophers, believe certain things just because they find them intuitive. It can also be possible to find some indication that they believe certain things just because they find them intuitive in their published work. But unless they offer the fact they (or someone else) find them intuitive as a reason to accept them, DE is not true.

One response to this problem might be to modify Climenhaga's definition by replacing the author with the reader: perhaps something is treated as evidence for a claim by a philosopher so long as it raises *the reader's* credence in the claim. This way private evidence could be kept out of the equation. The reader-centred approach also seems to make more sense of the fact that a philosophical argument is essentially a dialectical device: its primary point is to persuade whoever it is presented to, rather than to represent its author's internal thinking process.

But this proposal has serious flaws. First, just like philosophers' beliefs can remain intact by what they treat as evidence, their readers' beliefs can remain intact by what they are presented with as evidence. This might be due to irrationality, or for other reasons. For instance, I do not think that Zeno's paradoxes of motion make me any more likely to accept that motion does not exist. I believe that I can detect flaws that these paradoxes are based on, but, as I am not entirely sure whether I am right, they should have some influence on my view on the existence of motion. Moreover, when I first encountered the paradoxes I could not tell what was wrong with them, however I did not find Zeno's conclusion any more plausible. At least as far as I am concerned – but I suspect my case is

not very odd – these arguments seem completely ineffectual. And yet clearly something is being used as evidence here.

Secondly, as Climenhaga points out, evidence on his view is *context-relative*: whether something counts as evidence always depends on one's background knowledge. This means that readers with different background knowledge cannot always rationally increase their credence in a claim by learning the same thing, and it is unclear which reader we should focus on. One might be tempted to overcome these difficulties by specifying we are only concerned with some sort of ideal reader with certain background knowledge, certain cognitive abilities, certain level of rationality etc. But this would not take us very far – if something is being treated as evidence for *p* when it raises the ideal reader's credence in *p*, then how can we know whether something raises the ideal reader's credence in *p*? The answer must be either circular or unknowable.

Climenhaga's proposal and other possible Bayesian accounts can be characterised as instances of a *doxastic* view, according to which treating something as evidence is understood in terms of a relation between beliefs. The general idea behind this view can be expressed in the following way: *p* is treated as evidence for *q* if someone's belief that *q* is in some way based in their belief that *p*, where "based in" is understood broadly as causing, reinforcing, increasing the likelihood of etc. Note that most problems with Bayesian accounts that I have just described are also problems with doxastic accounts in general. This means we should probably abandon the doxastic picture of DE altogether: there seems to be no viable way of determining whether something is treated as evidence in philosophy in terms of how, if believed, it influences other beliefs.

To be fair to proponents of DE, not all of them are happy with the doxastic picture. Elijah Chudnoff suggests that it would be more fruitful to understand treating as evidence in terms of a relation between an *experience* and a belief. He thinks that intuition is a lot like perception – in fact it is "a form of intellectual perception" (Chudnoff 2013, p. 1) – and to explain his idea it is useful to make an analogy with how we justify our perceptual beliefs. What evidence do we have for them? For example, what evidence do I have that there is a computer screen in front of me right now? The obvious answer is that I see the computer screen in front of me. But saying that seeing the screen justifies, or is evidence for, believing that there is a screen can be interpreted in several different ways. One of them would be doxastic: my belief that I see the screen justifies my belief that there is a screen in front of me. However epistemologists have identified certain difficulties with this view (Lyons 2016). Some of them argue that the best way to overcome these difficulties is to assume that it is my perceptual experience, *the seeing of the screen itself*, that directly justifies my perceptual belief. How exactly is this possible? Chudnoff's answer is that it happens in virtue of the

experience's phenomenology: it is the way that it seems true that is connected to facts about what it represents. And, according to Chudnoff, what is true of perceptual experience is also true of intuition experience. But the experientialist proposal runs into the same problems as the doxastic one. Justifying a philosophical belief with an intuition-experience is not equal to using the experience as evidence in philosophy – for the former to become the latter, the justification needs to be somehow made *public*.

One might also altogether abandon the idea that being evidence is a relation between mental states. Perhaps evidence consists in mind-independent facts, or states of affairs. Here is how Jack Lyons outlines the idea:

> To say that *e* is evidence for *h* is not to say that anyone is prima facie justified in believing *h*; it is not even to say that anyone who believes *e* has any justification for believing *h*, for one might fail to appreciate *e*'s evidential significance regarding *h*. However, *e*'s being [factual evidence] for *h* does imply that someone could become justified in believing *h* on the basis of *e*. It implies that *e* is the sort of thing that could justify one in believing *h*, even if only when supplemented with the right, true, background beliefs. (Lyons 2016, p. 1055)

If DE is a thesis about factual evidence, then what philosophers use as evidence that *p* is the mind-independent fact that someone has an intuition that *p*. How plausible is it? In a sense, the proposal is even more problematic than the previous two. Not only does it leave a gap between having justification and offering justification, but also between knowing facts and having justification. The idea of justification that stems from "appreciating the evidential significance of a fact" seems much harder to flesh out than, for example, the doxastic idea of justification that stems from one credence influencing another. The proposal also seems to introduce the dubious idea of *idle evidence* that fails to justify anything as, for example, it is unknown – and yet still counts as evidence. I am not suggesting that these additional problems cannot be overcome. However even if they can, the basic difficulty remains unsolved: being justified in believing something by intuition is not the same as offering this justification for others to accept.

But if the doxastic, experiential and factual accounts are rejected, what are we left with? My suggestion is to pay more attention to *logical* relations between propositions. After all, if there is one thing that all philosophers do to defend their views, it is *making arguments*. Why not simply take premises of an argument as something treated as evidence, and its conclusion as something it is meant to be evidence for? In other words, treating as evidence can be understood as synonymous with *inferring*.

Inferences can, of course, be valid or invalid – we do not need to assume that philosophy is free of logical errors. They can also be of different types: deductive, inductive or abductive. But, in any case, they always link propositions, not mental states. Focusing on arguments themselves rather than on mental states of people who deal with the arguments seems to capture the phenomenon of philosophical evidence in a simpler and more straightforward way.

That said, it is important to stress that the inferential account is not *irreconcilable* with doxasticism, experientialism, or factualism. Perhaps what constitutes evidence *on a basic level* is a belief, or an experience. Perhaps it is something mind-independent, like a state of affairs. Whatever it is, it can be translated into an inference, and this inference is eventually expressed in natural language. The doxastic, experiential and factual accounts of evidence in DE should therefore be dismissed only to the extent they refer to evidence that is not translatable into an inference that can be identified in a philosophical text. This restraint is dictated simply by the public nature of philosophy.

One can now ask: how do we go about testing DE, thus understood, as a hypothesis about philosophical practice? The most obvious solution would be to pay attention to linguistic means used to express the inferences in a text. That is we should look for expressions like "so", therefore", "hence", "thus", "it follows that", "if – then", "for", "as", "because", "indicates that", "suggests that", "makes it plausible that", "due to", "is the reason why", "for that reason", "is a reason to think that", "by virtue of", "as a result of", "accounts for", "explains", "on the basis of", "thanks to", and synonymous. If DE is true, on the one side of such connective we should be able to find the fact about some proposition's intuitiveness, which would be expressed by phrases like "intuitively", "it is intuitive that", "there is an intuition that", "it seems that", "it appears that", "it strikes me that", "it is non-inferentially believed that" and so forth. On the other side of the inference-indicator we should be able to find the proposition itself. For example, something like "it seems that *p*, therefore *p*", or "the fact that *p* is intuitive suggests that *p*" in a text would clearly support DE.

It might be objected that intuitions are sometimes used as *indirect evidence*, which could not be reflected by inferences of the kind I have just described. For example, Chudnoff argues that intuition-states "immediately justify believing some of their contents, namely those associated with presentational phenomenology", but also "mediately justify believing other of their contents, namely those not associated with presentational phenomenology but appropriately supplemented by background information" (Chudnoff 2021, p. 210). I am going to argue that while Chudnoff's specific account of the latter kind of justification in philosophy is problematic, philosophers sometimes do rely on intuitions as indirect evidence. However this concession does not undermine

the adequacy of my criteria of testing DE, or my claim that DE is false. As I am going to explain, what is typically meant by "relying on intuitions" does not include the kind of relying on intuitions that actually takes place.

Another objection might be that it is not impossible to rely on intuitions as evidence without mentioning it in the text. Perhaps the practice of relying on intuitions is so transparent and universally accepted that philosophers do not need to make it explicit. For example, Bealer writes that "is is truistic that intuitions are *used* as evidence (or reasons) in our standard justificatory practices" (Bealer 1999, p. 30). If he is right, we probably should not expect philosophers to state the obvious. I think this objection needs to be taken seriously – later in this chapter I am going to explain how this "tacit" version of DE should be understood, and how it should be tested.

7. Whose intuitions?

Obviously, different people may have different intuitions, which can make one wonder: *whose* intuitions are proponents of DE talking about? Joshua Alexander and Jonathan Weinberg distinguish three answers to this question:

> First, it might be supposed that when a philosopher relies on intuitions as evidence, she is relying only on her own personal intuitions as evidence. Let's call this view, *intuition solipsism*. Second, she might be relying on her own intuitions because she takes those intuitions to be representative of the intuitions of the class of professional philosophers. Let's call this view, *intuition elitism*. Third, she might be relying on her own intuitions because she takes those intuitions to be representative of the intuitions of a broader class that includes non-philosophers – commonly referred to as "the folk." Let's call this view, *intuition populism*. (Alexander & Weinberg 2007, p. 57)

Alexander and Weinberg find intuition solipsism to be the least plausible option. First, philosophers typically use impersonal forms like "it is intuitive that", or plural forms like "our intuition is that" to refer to their evidence. Secondly, philosophy is a dialectical enterprise – it involves communicating with others and, typically, trying to persuade them. It is hard to imagine how philosophy could remain dialectical if philosophers are only concerned with their own private intuitions.

This leaves us with intuition elitism and intuition populism. One reason to reject the latter is based on the idea that philosophers are not interested in how the folk understand phenomena like knowledge, justice, reference, causation, truth etc. Rather, they are interested in investigating, or perhaps creating, their own technical concepts of knowledge, justice etc. – and folk intuitions are of

no use for this purpose. Another reason would be to suppose that even if philosophers are interested in folk concepts, their intuitions are still better suited for investigating these concepts due to philosophers' expertise. Alexander and Weinberg are sceptical about both arguments and opt for the populist reading of DE. For my part, I will discuss the debate over relying exclusively on expert intuitions in philosophy in more detail later in chapter 3. Here I only want to point out that I believe DE is false irrespective of which of the three answers we adopt, and my arguments against DE are mostly neutral in this respect. There is one exception: advocates of the elitist view can offer a reply to what I call the argument from counterintuitive conclusions that is not accessible to the advocates of the populist view. But, as I am going to argue in chapter 4, the reply does not work for any of my case studies, and the reasons why it fails can be generalised to other cases.

8. "The method of cases"

Proponents of DE typically argue that the practice of relying on intuitions is best exemplified by what they call "the method of cases". Common instances include Searle's Chinese Room, Putnam's Twin Earth, Chalmers's zombies, Nozick's utility monster, Burge's arthritis-in-the-thigh, Gettier cases, Frankfurt cases, trolley cases, Thomson's violinist, Lehrer's Mr Truetemp, Foot's transplant surgeon, Jackson's Mary the colour scientist or Kripke's Gödel the thief.

But what is it that they all have in common? When explanation of any kind is given, virtually everyone agrees that any instance of its use consists of three elements: there is the case itself, there is one particular judgment that the case is supposed to "elicit" or "trigger", and there is a theory, or a generalisation, that the judgment is meant to be evidence for, or against. The description of the case is often, but not always, characterised as *a thought experiment*. The judgment is usually, but not always, characterised as *an intuition*. Here I set aside the intuition-free accounts of the method (such as Machery 2017) and only focus on the more common, intuition-oriented ones (such as Malmgren 2011 or Pust 2019).

Let us take a closer look at one of the most prominent examples: the so-called trolley problem, first introduced by Philippa Foot (Foot 1967), and later developed by Judith Jarvis Thomson in her two seminal articles (Thomson 1976, Thomson 1985). I am going to focus on the two versions of the case that for some reason have received most attention. The first is what I will call the bystander case. It is a modification of the original scenario described by Foot, in which a tram driver is about to hit and kill five people on the main track unless he turns the tram onto a sidetrack and kills one person. In Thomson's new version it is not the driver, but a bystander that faces the dilemma:

> you have been strolling by the trolley track, and you can see the situation at a glance: The driver saw the five on the track ahead, he stamped on the brakes, the brakes failed, so he fainted. What to do? Well, here is the switch, which you can throw, thereby turning the trolley yourself. Of course you will kill one if you do. (Thomson 1985, p. 1387)

The other is what I will call the footbridge case:

> you are standing on a footbridge over the trolley track. You can see a trolley hurtling down the track, out of control. You turn around to see where the trolley is headed, and there are five workmen on the track where it exits from under the footbridge. What to do? Being an expert on trolleys, you know of one certain way to stop an out-of-control trolley: Drop a really heavy weight in its path. But where to find one? It just so happens that standing next to you on the footbridge is a fat man, a really fat man. He is leaning over the railing, watching the trolley; all you have to do is to give him a little shove, and over the railing he will go, onto the track in the path of the trolley. (Thomson 1985, p. 1409)

The first scenario is supposed to elicit the judgment that it is morally permissible to throw the switch and the second scenario is supposed to elicit the judgment that it is morally impermissible to push the fat man off the footbridge. How about generalisations that these judgments are supposed to undermine, or support? Some have suggested that the difference between the two somehow corresponds to the difference between rights-based and utilitarian ethics. For example, Joshua Greene argues that the apparent clash is “Kant versus Mill, all in one neat little puzzle” (Greene 2013, p. 116). However this has little to do with the points Thomson is trying to make. In her article she simply takes it for granted that utilitarianism is a flawed moral theory: people have moral rights, and “rights trump utilities” (p. 1404). The bystander judgment might be in line with utilitarianism and the footbridge judgment might not, however Thomson is only interested in explaining the difference in terms of how different utility-trumping rights are violated, not violated or waived in both cases.

Immediately after having introduced the bystander scenario she makes it clear that the bystander judgment is meant to serve as a counterexample to “Killing one is worse than letting five die”, that is the principle defended by Foot in her discussion of the original version of the problem. If the bystander can throw the switch, then we have a situation when killing one is not worse. The footbridge scenario is in turn meant to provide a counterexample to “it is not morally required of us that we let a burden descend out of the blue onto five when we can make it instead descend onto one”. If the footbridge judgment is true, then we have a situation when this is exactly what is morally required of us.

Another example that is routinely offered as a clear and obvious case of relying on intuitions in philosophy is Edmund Gettier's "Is justified true belief knowledge?". In his paper Gettier takes on the theory according to which knowledge is justified true belief. He comes up with two scenarios that are meant to undermine this claim. Here is the more popular one:

> Suppose that Smith and Jones have applied for a certain job. And suppose that Smith has strong evidence for the following conjunctive proposition: (d) Jones is the man who will get the job, and Jones has ten coins in his pocket. Smith's evidence for (d) might be that the president of the company assured him that Jones would in the end be selected, and that he, Smith, had counted the coins in Jones's pocket ten minutes ago. Proposition (d) entails: (e) The man who will get the job has ten coins in his pocket. Let us suppose that Smith sees the entailment from (d) to (e), and accepts (e) on the grounds of (d), for which he has strong evidence. In this case, Smith is clearly justified in believing that (e) is true. But imagine, further, that unknown to Smith, he himself, not Jones, will get the job. And, also, unknown to Smith, he himself has ten coins in his pocket. (Gettier 1963, p. 122)

The story elicits the judgment that Smith does not know that (e), which contradicts the claim that knowledge is justified true belief, as Smith's belief that (e) is both true and justified.

Proponents of DE usually ague it primarily applies to "contemporary philosophy", or "analytic philosophy", however many of them point out that the method of cases has been in use since antiquity. The favourite example seems to be Plato's discussion of justice in Book 1 of *The Republic*. This case is somewhat more problematic to interpret for the same reason any Plato's dialogue is problematic to interpret: the relation between the views presented by different characters and the author's views is not always obvious. Moreover, the characters in *The Republic* are not discussing justice as such, but rather the poet Simonides's beliefs about justice. Here I will assume that the standard interpretation, according to which Socrates's criticism of Simonides's definition of justice, "truthfulness without qualification, and the giving back of whatever one may have taken from someone else", expresses Plato's view. Here is the famous counterexample:

> I think everyone would agree that if one were to take weapons from a friend who is a man of sound mind, and if he were to go mad and demand them back, one ought not to return them. The one giving them back would not be 'just' to do so, and again one should not be willing to tell the whole truth to somebody in that state (Plato/Emlyn-Jones & Preddy 2013, p. 19, 331c)

Any account of justice needs to takes this fact into consideration, which means that there must be something wrong with the definition.

In each case, we can identify the case description, the judgment about the case and the generalisation undermined by the judgment. Proponents of DE argue that what makes the judgment special is the fact that it is intuitive. According to them if the trolley judgments, the Gettier judgment and Socrates's judgment were not intuitive, the whole exercise would be pointless: there would be little evidence against Foot's thesis about killing, little evidence against the classical theory of knowledge and little evidence against Simonides's theory of justice, respectively. We take those judgments to be true because they just *seem true*, and then we reject the generalisations as inconsistent with the judgments.

9. The justification interpretation

What can be wrong with the methodological picture I have just outlined? One objection raised by critics of DE like Max Deutsch is that "philosophers argue for their judgments about thought experiments and cases" (Deutsch 2015, p. xvi). This means that we are expected to accept these judgments on the basis of *arguments*, not on the basis of the judgments' intuitiveness. I largely agree with Deutsch on this point, however I think his choice of words might be somewhat misleading. After all, the objective of *arguing for* a claim is typically to convince someone that the claim is true, however here we are dealing with claims that are probably already taken to be true by the interlocutors. Cappelen points out that judgments about cases often constitute "assumptions that in a typical non-philosophical context would be accepted by the conversation partners without a demand for further justification" (Cappelen 2012, p. 189), and since philosophy is about questioning everything, philosophers often try to find justification that is not demanded in a typical non-philosophical context. So instead of saying that philosophers *argue for* judgments about cases, I think it would be more accurate to say that they *provide justification* for them, or that they *explain what makes them true*, or they *back them up with evidence* (which has nothing to do with their intuitiveness).

For example, Thomson backs up her judgments that it is permissible to throw the switch and that it is not permissible to push the fat man off the bridge with the following principle:

> it is not morally required of us that we let a burden descend out of the blue onto five when we can make it instead descend onto one if we can make it descend onto the one by means which do not themselves constitute infringements of [stringent] rights of the one (Thomson 1985, p. 1409)

Similarly, Gettier backs up the judgment that Smith does not know with:

> (e) is true in virtue of the number of coins in Smith's pocket, while Smith does not know how many coins are in Smith's pocket, and bases his belief in (e) on a count of the coins in Jones's pocket, whom he falsely believes to be the man who will get the job. (Gettier 1963, p. 122)

This is combined with an unstated (but clearly identifiable) general premise, something along the lines of "if one bases one's belief that *p* in something disconnected from what makes *p* true, then one's belief that *p* is not knowledge". This way we are offered a straightforward deductive argument for the claim that Smith does not know.

Finally, Plato seems to back up the judgment that returning weapons would not be just by stating that "friends owe it to friends to do them something good and not something harmful" (Plato/Emlyn-Jones & Preddy 2013, p. 21, 332a) – this is Cephalus's reply to Socrates, referring to Simonides's beliefs, which reveals that either Simonides is inconsistent, or returning the weapons cannot be classified as "giving back of whatever one may have taken".

If the justification interpretation is correct, it seems that nothing turns on whether judgments about cases are intuitive: all that matters is whether their justification is sound. Proponents of DE often overlook the justification of judgments about cases, or sometimes even explicitly deny it is present in the text (for example, see Gutting 1998, p. vii). This seems to lead them astray: they mistakenly conclude it is the intuitiveness of the judgments that is intended to support them.

10. Overlooking justification

To see how serious a problem it may be, consider Pust's account of the method of cases. To illustrate how it works, he brings up four examples: the Gettier case, the transplant surgeon case, the Chinese nation case and the flagpole case. He argues that in each of them the judgment is used against a particular theory: the justified true belief theory of knowledge, act utilitarianism, the functionalist theory of mind and the deductive-nomological theory of explanation, respectively (Pust 2019). I have already discussed how Gettier supports the judgment that Smith does not know with evidence. The other three examples are not any different. To avoid making this section unreasonably long, I will not delve into somewhat complex details of the arguments, and mostly restrict myself to identifying the justification for each of the three judgments. What follows requires more familiarity with relevant theories than my previous discussion of Thomson, Gettier and Plato.

First, we have the judgment that it is not morally permissible for a surgeon to kill one healthy patient and use his organs to save five other patients. The scenario first appears in Thomson's 1976 *Killing, letting die, and the trolley problem*.

For some reason Pust believes that the judgment is meant to undermine *act utilitarianism.* This is not correct: while it is true that the case is used to illustrate a problem with consequentialist ethics, Thomson is not interested in assessing consequentialism in any form. Rather, she contrasts the judgment about the surgeon with the judgment about Foot's original trolley case ("it permissible for the driver to divert the tram") to investigate the morality of killing and letting die. In any case, regardless of what exactly Thomson needs the judgment for, it should be clear that she tries to offer a positive case that goes far beyond simply making the judgment. It is hard not to notice that immediately after introducing the scenario, Thomson explains why on Foot's theory the surgeon should not proceed:

> We must accept that our 'negative duties', such as the duty to refrain from killing, are more stringent than our 'positive duties', such as the duty to save lives. If David [the surgeon] does nothing, he violates a positive duty to save five lives; if he cuts up the healthy specimen, he violates a negative duty to refrain from killing one. Now the negative duty to refrain from killing one is not merely more stringent than the positive duty to save one, it is more stringent even than the positive duty to save five. (Thomson 1976, p. 206)

Thomson then goes on to argue that Foot's theory is mistaken: the actual reason is that "the healthy specimen has more claim on those [body] parts than any of the five has" (ibid., p. 213). A large part of the article is devoted to explaining why Thomson's justification is superior to Foot's justification. There is no indication that Thomson thinks we can know that it is not permissible to kill the healthy specimen to save the five on independent grounds (for example, because our intuition tells us so), and then tries to infer further claims from this fact.

Let us now turn to Ned Block's Chinese nation argument, which is a particularly striking example. Block not only provides justification for his judgment that the Chinese nation, organised in a certain way, lacks qualia, he explicitly says he is not relying on an intuition that the Chinese nation lacks qualia and expects the reader to accept this judgment solely on the basis of his argument (Block 1978, pp. 281-2). He goes on to argue that while intuitions about qualia in general are highly unreliable, there are reasons to take this particular one seriously: the Chinese nation is designed to mimic a system that we know possesses qualia, and this fact explains the behaviour of the system better than the existence of qualia. This is the first part of the justification. Block admits that the point is far from decisive, so he proceeds to offer additional reasons. He argues that functional equivalence entails neither psychological equivalence nor neurophysiological equivalence, which makes it reasonable to suppose that what is functionally equivalent in this case lacks qualia, unless

there is a good argument for the claim that mental properties are functional properties. Block thinks that the only argument for this claim is that "functional identities can be shown to be true on the basis of analyses of the meanings of mental terminology" (ibid., p. 296), so he spends the next several pages attacking this argument. All these considerations support Block's judgment that the Chinese nation, organised in a way specified in the scenario, lacks qualia.

Finally, there is the flagpole case. Pust refers to a seminal paper by Sylvain Bromberger, which uses the example about the Empire State Building, better known in its later derivation involving a flagpole (see Levin&Levin 1977). It is worth quoting the vignette, accompanied with the judgment, in full:

> There is a point on Fifth Avenue, *M* feet away from the base of the Empire State Building, at which a ray of light coming from the tip of the building makes an angle of θ degrees with a line to the base of the building. From the laws of geometric optics, together with the "antecedent" condition that the distance is *M* feet, the angle θ degrees, it is possible to deduce that the Empire State Building has a height of *H* feet. Any high-school student could set up the deduction given actual numerical values. By doing so, he would not, however, have *explained* why the Empire State Building has a height of *H* feet, nor would he have *answered* the question "Why does the Empire State Building have a height of *H* feet?" nor would an exposition of the deduction be the explanation of or answer to (either implicitly or explicitly) why the Empire State Building has a height of *H* feet. (Bromberger 1966, pp. 92-3)

According to Pust what matters for Bromberger is primarily the fact that it is intuitive that the student would not have explained the height of the Empire State Building by setting up the deduction: the intuitiveness makes the judgment true (or at least likely to be true), so we must reject any theory that implies that the judgment is false. However Bromberger himself never suggests anything like it. Instead he *justifies* his judgment by arguing that "there must be laws according to which the Empire State Building will have the height it has even in total darkness" (ibid., p. 106), and its height cannot be explained without appealing to these laws. The bulk of the paper discusses the nature of the relation between laws and explanation, and how the relation supports the judgment in question, along several other judgments about what counts as an explanation.

Even if we put aside that Block openly dismisses any DE-friendly interpretation of what he is doing, there is something odd about the fact that Pust ignores these crucial passages in his discussion of the four examples. I think it neatly illustrates a distortion that DE often leads to – it makes philosophers overlook how judgments about cases *are justified* and focus on how they *seem*.

11. The abductive interpretation

One might think that if only proponents of DE stopped overlooking justification of judgments about cases, it would immediately become clear to them that intuitions are not treated as evidence in philosophy. Unfortunately, it is not that simple. There is a DE-friendly way of accounting for passages I have just discussed. It can be argued that judgments like "it would be wrong to return the weapons to one's friend who has gone insane" are not inferred from principles like "friends owe it to friends to do them something good" – it is the other way around. We are dealing with a sort of *inference to the best explanation*, or *abduction*, from the former to the latter (I am going to use the two terms interchangeably, which is not universally accepted – see Mackonis 2013). On this account, judgments about cases serve as independently attested data to be explained by theories. For example, when someone argues that human activity is the best explanation for crop circles, they take it for granted that crop circles exist – apparently because they have been observed. Similarly, Plato argues that "friends owe it to friends to do them something good" is the best explanation of why it would be wrong to return the weapons, taking it for granted that it would be wrong to return the weapons – apparently because it has been *intuited*. In both cases we have an independent source of knowledge of the facts we are attempting to explain: observation and intuition, respectively.

But this response runs into serious difficulties. First, offering the best explanation for data typically involves acknowledging, more or less explicitly, that there exist other explanations and demonstrating they are inferior, according to certain criteria. For example, Peter Lipton writes that "better explanations explain more types of phenomena, explain them with greater precision, provide more information about underlying mechanisms, unify apparently disparate phenomena, or simplify our overall picture of the world." (Lipton 2001, p. 106) To stick with the crop circles example: proposing human activity as the best explanation for their existence typically involves acknowledging that extraterrestial intervention has been proposed as an alternative explanation. This, however, does not resemble what philosophers do while discussing the paradigm cases. For example, Plato does not mention, or even hint at, any alternative explanations of why it is wrong to return the weapons to a friend and does not argue that they are worse in any of the respects mentioned by Lipton. This strongly suggests that Plato is not engaging in abductive reasoning.

Secondly, it is worth examining the wording of the relevant passages. As Deutsch admits, one needs to be careful here, as expressions like "explains", "accounts for", "is the reason for" or "because" can be used to represent both a deductive and abductive inference (Deutsch 2015, p. 96-7). I agree with Deutsch that the inference-language can be ambiguous, however some cases are still fairly clear-cut. For example, there seems to be little room for interpretation of how Plato uses the word

“for” / γάρ (“*for* he believes that friends owe it to friends...”/ τοῖς γὰρ φίλοις οἴεται ὀφείλειν τοὺς φίλους...). This is not how one would normally present this claim were it meant to serve as the best explanation of why we should not return the weapons. I am not suggesting that a word like “for” or “because” would be completely out of place in a presentation of an abductive inference, however it should not appear as its main indicator. Consider: “Crop circles exist because they were created by humans”. This sentence might look suitable in a concluding section of a discussion of what best explains crop circles, but as a standalone statement it simply would not work. For the same reason Plato’s sentence does not work as an abduction-indicator.

Third, philosophers sometimes come across conflicted judgments about cases, and the way they deal with the conflict shows they do not understand them as independently attested data to be explained by theories. For example, throughout her paper Thomson reports presenting her friends with different versions of the trolley scenario and asking them about their opinions. In most cases there is a consensus. Sometimes, however, her judgment differs. There is a version of the bystander scenario in which in order to throw the switch it is necessary to cross a patch of land that belongs to the person on the sidetrack, or to use his nail file, in both cases without the owner’s permission. To Thomson’s interlocutors diverting the tram in this situation seems permissible, but Thomson herself “does not find it obvious”. In another scenario, the person on the sidetrack, which has been unused for years, is a convalescent at a local hospital, having a picnic lunch. He was invited there by a city mayor, who had promised him no trams would ever be diverted onto the sidetrack. Unexpectedly, a tram is about to hit and kill five people on the main track, unless someone turns it towards the convalescent, and the only person who can do so happens to be the mayor himself. To Thomson’s “great surprise”, her interlocutors thought it would be permissible for the mayor to throw the switch in this situation as well.

Here are some possible ways of dealing with the judgment discrepancy, assuming the DE-friendly abductive interpretation is correct: one could conclude it is impossible to proceed as there is no clear intuition-data to explain; one could try to come up with different generalisations for different sets of judgments, one could try to find out which judgment is more widespread or more strongly intuitive; one could try to argue that someone’s faculty of intuition – if there is such a thing – was impaired or malfunctioning in some way. As it turns out, Thomson does none of these things. Instead, she looks into *reasons* to accept and reject judgments about cases. In the patch of land/nail file case, she argues that her interlocutors must be correct as “the rights which the bystander would have to infringe here are minor, trivial, non-stringent-property rights of no great importance” (ibid., p. 1411). In the city mayor case, she is not likely to change her mind straight away. She believes her

interlocutors assume that breaking one's promise does not infringe a stringent right, or at least a right not stringent enough to override the exemption allowing to sacrifice one in order to save five. Thomson remains unconvinced: it seems clear that in order to resolve the disagreement it would be necessary to examine reasons *behind reasons* to accept the judgment, that is reasons to think that breaking one's promise is too trivial to override the exemption. In both cases, Thomson believes judgements about moral permissibility of particular actions should be accepted or rejected on the basis of an *argument*, not on the basis of whether they are intuitive to anyone. The intuitiveness of judgments seems completely irrelevant.

Fourth, philosophers sometimes make comments about what on the abductive interpretation serves as the best explanation that the advocates of this interpretation must find baffling. For example, Thomson says she does not "find it clear why there should be an exemption for, and only for, making a burden which is descending onto five descend, instead, onto one" (ibid., p. 1408). Note that on the abductive interpretation this comment does not make much sense. Thomson should find it perfectly clear why there should be an exemption: the exemption thesis accounts for a number of judgments about different versions of the scenario. However she does not appear to think that the judgments are something that can justify, or support the exemption thesis. Rather, it is the other way around. We are then left with the exemption thesis that is far from obvious or self-evident, which means it needs to be supported by some further facts. Thomson says one such fact is that we are dealing with "something that is *already* a threat to more, and thus something that will do harm *whatever* [the bystander] does" (ibid.), but this can only serve as a *partial* justification. She feels she does not have enough evidence to justify the exemption thesis – hence her perplexity.

Fifth, philosophers sometimes *change* their judgments about cases over time and the way they do it does not bode well for the abductive interpretation. For example, in 2000s Thomson had come to the conclusion that after all it was *not* permissible to throw the switch and divert the tram in the bystander case. How was it possible? On the DE-friendly abductive view, Thomson's intuition about the case must have changed, or she must have decided something had been wrong with her ability to intuit the correct answer, or perhaps she must have learnt that people's intuitions about the scenario were different than she had previously thought – in any case, there must have been some sort of turnaround, failure or misunderstanding concerning someone's intuitions, which serve as the independent source of data to be explained by a theory. The problem is this is nowhere near how Thomson actually explains her change of mind. She says she was persuaded by Alexander Friedman, who argued that since she had first presented the problem, nobody – herself included – had been able to offer a satisfactory account of what makes it permissible to throw the switch.

According to Friedman this is because there is no such account to discover. On the other hand, we have a good reason to believe that it would be wrong to divert the tram onto the sidetrack: "it is intuitively plausible that negative duties really are weightier than positive duties." (Thomson 2008, p. 363)

Here the objection might be that the word "intuitively" indicates that we are dealing with a situation in which it is impossible to account for all intuitions and one intuition ("negative duties are weightier") simply trumps another ("it is fine to throw the switch"), but overall it is still true that intuitions are treated as starting premises in abductive arguments. But this response is problematic for a number of reasons. First, why did it take Thomson several decades to realise that "negative duties are weightier" is intuitive and therefore has to be treated as some sort of explanandum? Why did it not occur to her in 1976 or 1985? This seems highly implausible. Another possibility would be that Thomson did not find the proposition intuitive in the past, but this conjecture seems even more far-out: surely, if it were the case, she would have at least flagged it up in her article. Secondly, if Thomson or Friedman are trying to somehow weigh two intuitions against each other, why are they not invoking any criteria for solving this kind of conflict? Why exactly is one intuition supposed to override the other? Is it because it is more intuitive, or for some other reason?

There is much more to be said about the idea of sacrificing intuitions for the sake of preserving other intuitions, and I will return to this problem in chapter 3. As for using the word "intuition" and its cognates in one's first order philosophical practice supports – in chapter 2 I am going to argue that this kind of terminology never indicates anything close to DE.

12. The noninferential interpretation

Another DE-friendly interpretation of considerations that I call evidence for judgments about cases has recently been proposed by Elijah Chudnoff. According to it, these considerations are neither inferred from the judgments, nor the judgments are inferred from them: there is simply no inference-relation between the two. Rather, the considerations *enable* the judgments. Here is how Chudnoff understands the difference:

> If you infer c from $p_1...p_n$, then your justification for believing c is constituted by your justification for believing $p_1…p_n$. Say your justification for believing in the principle of mathematical induction is constituted by the testimony of a textbook. Then in the inference case your justification for believing the formula is partly constituted by testimony. If consideration of $p_1…p_n$ enables your intuition that c, then your justification for believing c need not be

> constituted by your justification for believing $p_1 \ldots p_n$. Rather, it is constituted by your intuition and whatever background information it draws on. Say your justification for believing that 4 + 2 is 6 is constituted by the testimony of a textbook. You learned this in school and just haven't thought about it since. Nonetheless, in the intuition case your justification for the formula need not be partly constituted by testimony. It is important not to assume that considerations used to enable an intuition are thereby incorporated into the background information drawn on in the intuition. (Chudnoff 2021, p. 147)

To illustrate: Gettier's "Smith does not know that (e)", call it GJ, can be interpreted as enabled by "(e) is true in virtue of the number of coins in Smith's pocket, while Smith does not know how many coins are in Smith's pocket, and bases his belief in (e) on a count of the coins in Jones's pocket, whom he falsely believes to be the man who will get the job", call it GC. In this case GC would not be evidence for GJ, but rather something that makes GJ intuitive, and the fact that GJ is intuitive would be Gettier's evidence for GJ. There would also be some "background information" behind the intuitiveness of GJ, and it would constitute part of the evidence.

According to Chudnoff, an analogy can be drawn between enabling intuition and enabling perception. Consider the phenomenon of multistable perception: certain images can depict different things, depending on which way they are looked at. For example, in a popular image known as "My wife and my mother-in-law" one can see a young woman facing away or a left profile of an old woman – but not both at the same time. We can imagine, argues Chudnoff, someone who can only see the old woman in the picture, and someone else telling him that the old woman's nose is the young woman's jawline, the old woman's mouth is the young woman's necklace etc. This consideration would make the first person see the young woman, but it would not constitute evidence that there is a young woman in the picture. The evidence would be the very experience of seeing the young woman, together with whatever background information it makes use of. Similarly, the role of considerations like GC could be to merely make a proposition like GJ seem true, without justifying it.

However there are strong reasons to think that Chudnoff's interpretation is not correct. First and foremost, GC simply does not seem to enable the experience of finding GJ intuitive, and the same can be said about other case judgments and their respective considerations. Note that proponents of DE typically ignore considerations like GC when they discuss judgments about cases (see Pust 2019 or Stich & Tobia 2016). On Chudnoff's account, this should lead to some sort of fatal miscommunication between proponents of DE and their readers: certain judgments are constantly pronounced to be epistemically special in virtue of being intuitive, but the reader cannot find them

intuitive, as there is nothing in the text to enable their intuitiveness. But no such miscommunication happens: nobody seems to accuse philosophers like Pust of arbitrarily calling certain judgments "intuitions" without offering any justification.

Secondly, if the non-inferential interpretation of the relation between considerations like GC and judgments like GJ is correct, why do philosophers routinely use "inferential" language to describe it? I have pointed out that words like "for", "because", "as" etc. in the original texts refer to the said relation. Note how unnatural it would be to say "there is a young woman in the picture *because* the old woman's nose is the young woman's jawline, etc.". If Chudnoff's analogy between perception and intuition is valid, it should also be unnatural for Gettier to say "*for* (e) is true in virtue of..." – and yet it is precisely how he formulates his sentence.

Having said that, I believe Chudnoff is on to something when he argues that philosophers do rely on intuitions to make things "more vivid" – he is only wrong to think that this practice has something to do with DE. Later in the chapter I am going to explain how I think intuitions are used as tools of discovery and tools of clarification – something that often is, but should not be conflated with what is typically meant by "relying on intuitions".

13. Uncontroversial abduction

I have argued that one way to defend DE would be to try to portray certain deductive arguments as abductive arguments. This, however, does not mean there are no genuinely abductive arguments in philosophy. Moreover, these genuinely abductive arguments often start with claims that can be characterised as intuitions. It may be tempting to appeal to this fact as evidence for DE, but I think it would be a mistake.

Let me illustrate this point. David Boonin challenges the account of the right to life put forward by Don Marquis. Both agree that in order to find out what the right to life is, we should first select several obvious cases of creatures who possess this right. Marquis proposes a number of such cases: an infant, a suicidal teenager, a temporarily comatose adult and a healthy, adult human being like you and me. Boonin accepts this proposal (Boonin 2006, p. 57). According to Marquis, the best explanation of why all four have the right to life is something along the lines of:

> If an individual P has a future-like-ours F and if either (a) P now desires that F be preserved, or (b) P will later desire to continue having the experiences contained in F (if P is not killed), then P is an individual with the same right to life as you or I. (ibid., p. 63)

Boonin believes that the best explanation is slightly different: "If an individual P has a future-like-ours F and if P has a present, dispositional and ideal desire that F be preserved, then P is an individual with the same right to life as you or I." Boonin then argues that his account of the right to life is superior to that of Marquis's for three reasons. First, it is more parsimonious: instead of two different morally relevant factors, it offers one. Secondly, its explanatory power is greater: the wrongness of acts that have nothing to do with killing can also be explained in terms of thwarting present, dispositional and ideal desires. Third, it turns out to have greater scope if we add one more case to the list: that of a depressed, suicidal person who due to a neurological malfunction will never be able to recover. On Boonin's account, it would be wrong to kill this person. Marquis seems to agree that it would be wrong, but this does not follow from his principle.

To some this may look like a perfect example of relying on intuitions in ethics: it simply *seems true* to both Boonin and Marquis that infants, suicidal teenagers, temporarily comatose people, and healthy adults have the right to life. Otherwise, there is nothing in their texts to support the claim. They then take their four intuitions and try to come up with a general principle that best captures the intuitions' content. How is this not a case of DE?

14. Common ground

I have argued that judgments about cases are typically *backed up with evidence.* Suppose I am wrong and there is nothing to support the judgments in the text. Would that make DE plausible? I do not think it would. It would only mean that judgments about cases are unsupported and other claims are inferred from them. This tells us nothing about whether the intuitiveness of unsupported claims plays any kind of justificatory role.

It is hardly surprising that philosophical arguments, or *any* arguments, for that matter, rest on unsupported premises. It simply follows from the fact that arguments cannot be infinitely long. Trying to support one's unsupported premises means one is only going to end up with another set of unsupported premises. However that fact that all arguments rest on unsupported premises does not mean that all arguments rest on intuitions, in the DE-sense. For example, I can start an argument with an unsupported claim that the distance between Tehran and Isfahan is shorter than the distance between Tehran and Shiraz. Does it mean I am using the fact that this claim is in some sense intuitive to support its content? Of course not. Most likely I am simply assuming this is something my readers already know, so I do not need to waste their time explaining why it is the case. Or, should they not know it, that the claim's truth and evidence in favour of it is quite uncontroversial

and easy to look up, I can therefore expect the readers to take my word for it. Simply put, I am placing "the distance between Tehran and Isfahan is shorter" *in the common ground*.

Let us now ask: why not think about various philosophical starting premises in the same dialectical way – namely as something that does not need to be argued for in a particular text? Why not think it is the quality of being already accepted by the readers that makes various judgments suitable to start philosophical arguments with? Someone might reply that if we identify "intuition" with "something assumed to be already accepted by the readers", it would follow that intuitions are used as evidence. However, even putting aside the eccentricity of this usage, DE would still be false: on this account the fact that someone assumes *p* to be widely accepted is clearly not meant to be evidence for *p*.

The common ground interpretation has at least one clear advantage over DE: it explains the suspicious lack of explicit claims in the form of "*p* is intuitive, therefore *p*" or similar in the texts in question. As Deutsch points out, if DE were true we would expect perhaps not all, but at least some philosophers to conform to this pattern – but in fact none of them do (Deutsch 2015, p. 97). Of course Deutsch's argument does not undermine the "tacit agreement" version of DE, and the difference between it and the common ground interpretation might be somewhat elusive. David Chalmers, who defends a form of DE, suggests to understand it in the following way:

> Propositions in the common ground typically have a broadly inferential dialectical justification: it is just that this justification is in the background, stemming from how the proposition entered the common ground in the first place. Often the justification will be a testimonial or perceptual justification, deriving from previous communications or from external sources. As before, these dialectical justifications need not be explicitly articulated by the parties to a conversation; they merely need to be mutually recognized. By contrast, with intuitions as I am characterizing them, there need be no broadly inferential justification that the parties recognize; there will only be a broadly noninferential justification, perhaps associated with the obviousness of the claim in question. (Chamlers 2014, p. 538)

I agree with Chalmers that mutual recognition of justification typically characterises propositions in the common ground, however his account seems too restrictive. Suppose a philosopher puts forward an argument whose starting premise is *p*. She assumes that all her readers accept *p*, but she is not sure *why* they accept it. This situation hardly vindicates DE, however it is reasonable to say that *p* is in the common ground. Or suppose that a philosopher puts forward an argument whose starting premise is *p*, but she thinks different readers are going to accept *p* for different reasons. Here, again, even though there is no one particular justification recognised by all parties, the common ground interpretation seems correct while DE clearly does not. Finally, suppose that a philosopher puts

forward an argument whose starting premise is *p* and assumes all her readers accept *p*, but she cannot think of any reasons for *p*. This situation still does not confirm DE – in fact, it is not even consistent with DE, which states that a philosopher offers a particular reasons for *p*, namely that *p* is intuitive. In short, "common ground" is better understood as "assumed to be accepted by all parties for any reason, or even without an identifiable reason", rather than "assumed to be accepted by all parties for the same reason (other than being intuitive)". The tacit version of DE would in turn imply that something is assumed to be accepted by all parties on the basis of being intuitive – on this point Chalmers seems to agree.

It must be stressed that placing *p* in the common ground does not equal believing that *p*. One often starts with what one's opponent's already believe, without necessarily believing it oneself – familiar phrases like "for the sake of argument", "I'll grant you that", "let's assume that" etc. are often used in this context. The proponent of the argument may even disbelieve her own starting premises and be open about it – although keeping one's attitude towards *p* to oneself is also perfectly consistent with putting *p* in the common ground. Simply put, what is in a philosopher's common ground should be treated as independent of what she believes and whether she reveals what her beliefs are.

How can one decide between the common ground hypothesis and the tacit DE hypothesis in a given case? We need to check whether *p*, which serves as a starting premise in an argument, is challenged *in a different text* – by a different or perhaps even the same author. If DE is true, we would expect the text to mention the consensus that the intuitiveness of *p* counts as evidence for *p.* If DE is false, we would expect this idea to be ignored.

As I mentioned, I do not believe that the common ground interpretation is correct with regards to the judgments I have discussed – I think Thomson, Gettier and Plato provide justification for their judgments. Perhaps I am wrong about this, or perhaps the same is not true about other paradigm judgments listed above. In any case, proponents of DE always face a double challenge: first, they need to show that judgments about cases constitute argumentative starting points, and, secondly, they need to show how the fact that judgments about cases are starting points supports DE. I do not think this challenge can ever be met, regardless of which alleged example of relying on intuitions one takes up. Moreover, what I have just said about the so-called method of cases can be applied to any assertion about using intuitions as evidence in philosophy. Whenever DE commits one to identifying a proposition as an intuition that is being treated as evidence, two questions can be asked: is this proposition backed up with any evidence (other than the fact that the proposition is intuitive) in the text and if not, is it part of the common ground? DE implies that the answer to both questions is no. In my view, the answer to one them is always yes.

15. "*Prima facie*"

Proponents of DE might complain that I have just presented them with a false dilemma: it is possible to accept the justification interpretation and still believe that intuitions are used as evidence for their contents. Perhaps intuitions are taken to be some sort of *defeasible* evidence, which can get confirmed or undermined – maybe even overridden – by further evidence. For example, it can be argued that Gettier treats the intuitiveness of "Smith does not know" as initial evidence that Smith does not know, and this initial evidence is then further confirmed by the fact that basing one's belief that *p* in something disconnected from what makes *p* true means that one's belief that *p* is not knowledge. Or perhaps Thomson treats the intuitiveness of "it is not permissible to use the nail file to throw the switch" as initial evidence that it is not permissible to use the nail file to throw the switch, but then this initial evidence is overridden by the fact that using the nail file without the owner's permission does not violate a stringent right of a person. Some philosophers like to talk about "*prima facie* reasons", "*prima facie* objections", "*prima facie* problems", "*prima facie* doubts" or "*prima facie* counterexamples" – perhaps what they mean is this kind of defeasible intuitive evidence. I will call this view the *prima facie* version of the justification interpretation.

Does this account hold water? I do not think it does. Several reasons to reject the abductive interpretation and the DE-friendly version of the abductive interpretation are also reasons to reject the *prima facie* version of the justification interpretation. First, there is the lack of explicit inferences from "*p* is intuitive" to "*p*". For example, Gettier does not say that Smith does not know because it is intuitive that Smith does not know *and* because Smith bases his belief on a count of the coins in Jones's pocket; he says Smith does not know because he bases his belief on a count of the coins in Jones's pocket, full stop. If Gettier is appealing to the intuitive, why does he stay silent about this? And why does virtually everyone else stays silent?

Some might think that even though Gettier is not explicitly appealing to two different sources of evidence for the claim that Smith does not know, he still hints at them by using particular words and expressions. Ethan Landes defends this position in his recent paper:

> Consider the passage in which Gettier first gives his verdict: 'But it is equally clear that Smith does not know that [E] is true; for [6, 7, and 8]' (1963, 122). There are signs of Gettier taking both options in this passage.'Equally clear' suggests Gettier is appealing to obvious external justification, while 'for' suggests Gettier takes 9 as following from 6, 7, and 8. (Landes 2020, p. 10)

"9" stands for "Smith does not know E", "6" for "E is true in virtue of the number of coins in Smith's pocket", "7" for "Smith does not know how many coins are in Smith's pocket", and "8" for "Smith bases his belief in E on a count of the coins in Jones' pocket, whom Smith falsely believes to be the man who will get the job" (ibid., p. 8). By "obvious external justification" Landes means something like "justification by its own intuitiveness, not mentioned in the text". I do not think, however, that this reading of Gettier's sentence is even remotely plausible. Consider the statement: "It is clear he does not have a PhD in philosophy, for he has never heard about Kant or Hegel." How reasonable would it be to interpret it as "He does not have a PhD in philosophy for it is intuitive that he does not have a PhD in philosophy *and also* because he has never heard about Kant or Hegel"? Surely "clear" in my sentence indicates that I take my reason for believing that someone does not have a PhD in philosophy to be a strong reason, and consequently I think my conclusion is well-established. It does not refer to the fact that I take the alleged intuitiveness of my conclusion to support its content. The same goes for Gettier's sentence, and the point can be generalised to other judgments about cases. The wording of relevant passages simply does not favour the *prima facie* interpretation.

16. Discovery vs justification

At this point one might ask: if this is not what philosophers mean by "*prima facie*", what do they mean? For example, why does Block repeatedly mention "*prima facie* doubts" concerning functionalism? If he does not think that the intuitiveness of those doubts matter in terms of justification, why is he even talking about them? I doubt whether there is one uniform way of using the expression "*prima facie*" in philosophy, however I think there is a plausible interpretation of how it is often used, which is both incompatible with DE and allows a role for intuitions to play in philosophical enquiry.

An analogy may be helpful here. Imagine a detective to whom a strong intuition occurs: it seems to her that one of the suspects has committed the crime, but she has no idea why. She decides to follow the intuition and pursue a certain line on inquiry, in the course of which she is able to collect evidence that reveals the suspect to be the culprit: fingerprints, DNA samples, CCTV recordings, witness testimony etc. The evidence is then presented at a trial. The detective does not treat her own intuition as worthless: she thinks it indicates that there is evidence to be found somewhere, and not much evidence to be found elsewhere. She does, however, treat it as worthless *in court*: arguing that someone had an intuition that someone else was guilty cannot help convict anyone, and she is

perfectly aware of this fact. In this metaphor the detective's intuition is analogous to our intuitions about philosophical cases; fingerprints, DNA samples etc. are analogous to whatever philosophers justify those judgments with, and the trial is analogous to a typical philosophical debate. Philosophers take what seems true to us and try to find out what, if anything, backs it up. Whatever they think backs up is treated by them as evidence. Whatever they back up with evidence is the *prima facie* claim.

One might argue that if the detective thinks her intuition indicates there is (court-compliant) evidence to be found somewhere and not much elsewhere, she is clearly using her intuition *as evidence for its content*: the fact that something seems true to her indicates that it is in fact true, or likely to be true. I do not disagree with this point, however it needs to be stressed that we are talking about *non-public evidence* here. Just like it would be bizarre to appeal to this kind of non-public evidence in court, it would be bizarre to appeal to this kind of non-public evidence in a philosophical debate. Intuitions are simply not used to publicly support or undermine philosophical views, even though they can, in a sense, lead one to discover what is used to support and undermine philosophical views. Philosophers often assume that when it seems to us that something counts as justice, knowledge, reference etc., we are on to something – that is, there probably are good reasons to think it actually counts as justice, knowledge, reference etc. They then try to discover what those reasons are, and publish their findings.

Moreover, my analogy can easily be modified to eliminate treating intuitions even as non-public evidence while still engaging in essentially the same practice. Suppose that the detective does not not really trust her intuition, but, for whatever reason, she still decides to follow it, which leads her to discover evidence. The same can be true of a philosopher: she can give priority to intuitive judgments in the process of examining reasons behind them without ever treating the judgments' intuitiveness as evidence of their content. The reason for prioritising these judgments can be simply that they are more *interesting* than random judgments which nobody finds plausible. There is a significant overlap between what is intuitive and what is believed, and we are naturally more curious about what we believe: we want to know why we believe it and whether we are justified in believing it.

It is sometimes argued that DE is supported by the fact that respectable philosophical theories of justice, knowledge, reference etc. generally accommodate our intuitions about what counts as justice, knowledge, reference etc. To anyone who rejects DE, the objection goes, this must look like a surprising coincidence (Climenhaga 2018, pp. 79-80). On my view, however, there is no coincidence. Philosophers often pay more attention to our intuitions, however this does not mean

they treat intuitions as evidence – at least not as *public evidence*. This explains why, for example, Thomson reports asking her friends about different trolley scenarios. Perhaps she assumes her friends must be on to something when they make their verdicts, perhaps she merely finds the verdicts more attractive to explore. In any case, she is not *justifying* claims with the fact that her friends make them, or find them intuitive, which is what DE implies. Or consider how Thomson describes her change of mind about the bystander case in her 2008 article: she writes that many philosophers for many years have focused on judgments like "it is permissible to throw the switch" and "it is impermissible to push the fat man off the bridge" and strove to find good justification for them, but failed (Thomson 2008, p. 363). Because of this failure we should turn to nonintuitive judgments, such as "it is impermissible to throw the switch" and see if they can be justified. Whether something counts as a good justification has nothing to do with the fact that it is intuitive, however it is still true that philosophers often prioritise intuitive judgments in their investigations.

A similar point has been made by R. M. Hare. He argues that moral philosophers often appeal to what he calls "the opinions of the ordinary man". Plato's "one should not return the weapons" is one of his examples. According to one interpretation, these opinions are used as data to be explained by moral theories. But if this is true, it follows that philosophers are "merely being conservative or conventionally-minded or just stupid" (Hare 1972, p. 124-5). In any case, they are engaging in a terrible kind of reasoning, as intuitions are clearly a very bad guide to moral truth. Fortunately, writes Hare, there is another, more plausible way of understanding the practice:

> in spite of the fact that the opinions of the ordinary man have in themselves no probative force in moral philosophy, a due respect for them may lead us to understand its problems better. They do not supply an argument, but they make us look for one. (ibid., p. 134)

Since Hare wrote it in 1970, much seems to have changed for worse. The argument from received opinion used to be one of several interpretative possibilities, today it has become the prevalent view. This might look somewhat surprising if we compare contemporary philosophy of philosophy with contemporary philosophy of science, which uses a well-established distinction between the context of discovery and the context of justification. The former has to do with, roughly, actual thinking processes behind the creation of new scientific ideas and theories, and the latter with what is used to evaluate those ideas and theories in the scientific community. In an anecdote often invoked to illustrate the distinction Friedrich August Kekulé is led to discover the ring structure of benzene by dreaming about a snake seizing its own tail (Kekulé 1890/1958, p. 22). Even though Kekulé's dream played an important role in the discovery, it would be strange to suggest that dreams can be treated as evidence in chemistry. Perhaps Kekulé thought that the content of his dreams carried

some evidential weight, perhaps he did not: as far as scientific justification is concerned, this is beside the point. The hypothesis about the ring structure of benzene had to be tested by standards that were independent of contents of anyone's dreams.

A number of ways of understanding the discovery vs justification distinction have been proposed since Hans Reichenbach introduced it in 1938: it can refer to two processes distinct in time, to the *process* of discovery and *methods* of justification, to something that can be analysed empirically and something that can be analysed logically, or to something still different (see Hoyningen-Huene 2006). The version that I propose to apply in metaphilosophy centres on theory validation: according to it, evidence in the context of philosophical justification is simply something suitable to support or undermine a theory *in a philosophical debate*. There might also exist evidence in the context of philosophical discovery: something that can be relied on in a creative process, but not something that can be used to justify a theory in a philosophical community. And DE, as I understand it, always refers to treating intuitions as evidence *in the context of justification only*.

Distinguishing between relying on something as evidence in the context of discovery and relying on something merely as a working hypothesis – without treating it as evidence in any sense – is an intricate psychological matter. For example, I am not sure whether the way that Kekulé relied on his dream resembled the way that the detective relies on her intuition in my first thought experiment, or perhaps the way she relies on intuition in the second one. I have similar doubts with regard to the way many philosophers seem to rely on intuition. As I do not need to solve this issue to make my case against DE, for simplicity's sake I am going to refer to both kinds of practice as "using as evidence in the context of discovery".

17. Clarification and persuasion

Just like Thomson's trolley judgments and Gettier's judgment about Smith, Plato's "weapons" judgment is probably not intuitive *by accident*. According to his argument, fulfilling certain obligations, such as one's obligation to do good to one's friends, can require not giving back what one owes – which conflicts with the definition of justice put forward by Simonides. To clarify, Plato introduces his thought experiment. As it *seems obvious* that returning the weapons would be wrong, we can easily understand how the premises support the conclusion. We can imagine a situation in which it is not obvious that someone should not give back what she owes, and yet it still follows from the fact that she should to good to her friend. Had Plato decided to use such judgment, his argument would have become more difficult to comprehend, but its substance would not have

changed. We can say that Plato is relying on an intuition *as a clarification device*, which is very different from using an intuition as evidence.

A related, but distinct function of appeal to intuitions has to do with *persuasion*. In addition to making himself clear, Plato probably tries to make his readers believe that justice is not what Simonides's definition suggests, and the intuitiveness of "one should not return the weapons" helps him achieve this goal. It is well known that impeccable arguments are neither necessary nor sufficient for successful persuasion. One can accept the premises of an argument as undeniably true and the logic of it as perfectly valid and yet still refuse to accept the conclusion. For this reason, it may be wise to avoid conclusions that seem false, and opt for ones that seem true. This, of course, is not always possible, but when it is possible philosophers often take advantage of the opportunity, which gives us yet another type of philosophical practice that can be confused with DE.

Using intuitions as clarification devices and using them as persuasion devices often go hand in hand, but there is no necessary connection between the two. Consider the Monty Hall problem mentioned earlier. A game show host offers you a choice between three gates. Behind one of them there is a prize, two others are empty. You pick your gate, then the host opens one of the empty gates and asks if you would like to change your original choice. To most of us it seems false that the probability of winning increases after switching. We think it is just obvious that it does not matter whether we switch or not, the probability is 1/2 either way. There are many ways of explaining of why this is not the case. For example, one can utilise formal probability calculus in one way or another. Another solution would be to slightly modify the original scenario. Suppose that instead of three gates there are one hundred of them. Everything else stays the same: there is only one prize behind one gate, and the host knows which one it is. You pick your gate, the host then opens ninety-eight empty ones and asks you if you would like to change your original choice. In this case most people immediately understand that switching increases their chances of winning: clearly the probability is 1/100 if you stick, and 99/100 if you switch. After all, if you are lucky with your first guess (1/100 chance that you are), the alternative gate would be a randomly selected empty one. And if you are unlucky (99/100 chance that you are), the alternative gate would be the winning one. But if this is so, then in the three gate version of the game the probability of winning must rise from 1/3 to 2/3 after switching.

Jason Rosenhouse, who has spent years teaching probability theory using the Monty Hall example, writes that "students who are totally unpersuaded by elaborate probability calculations or arguments based on Bayes' theorem typically cry uncle at this point" (Rosenhouse 2009, p. 39). Elaborate calculations do not necessarily fail to clarify, however they do fail to persuade. For this reason, a

mathematician whose aim is to actually change people's minds about the chances of winning after switching may *rely on an intuition* by preferring the hundred gate scenario over other ways of explaining the problem. But it would be absurd to think that the (correct) intuition that it is advantageous to switch in the hundred gate version is used as evidence that it is indeed advantageous to switch, or that the (mistaken) intuition that it is not advantageous to switch in the three gate version is used as evidence that it is not advantageous to switch. These intuitions tell us nothing about whether switching is a good idea, and they are treated accordingly by mathematicians. In my view, philosophers are not any different in this respect: they often use intuitions to persuade without using them as evidence.

18. Four ambiguities

So far I have argued that the claim that philosophers rely on intuitions is ambiguous in at least four ways. First, there is the evidence versus clarification/persuasion device ambiguity. Then within the evidence interpretation there is the propositional content versus mental state ambiguity. Then within the mental state interpretation there is the evidence for its content versus evidence not for its content ambiguity. And finally, within the evidence for its content interpretation there is the context of discovery vs context of justification ambiguity. The last reading is what I call DE and what I argue against in this article.

It is undeniable that in some sense philosophers do rely on intuitions, however what is typically meant by "philosophers rely on intuitions" is DE, and DE is false. It might be objected that my approach is too stringent: perhaps the commitment to DE is not as widespread as I suggest it is. Let us explore this possibility. First, why should we reject the clarification/persuasion device reading? The main reason seems to be that "philosophers rely on intuitions" is often used interchangeably with "philosophers rely on intuitions *as evidence*", and it would make little sense to talk about *evidence* in this context: helping someone understand or trying to persuade someone that something is the case is different from giving evidence for why something is the case. Other common synonymous expressions are "philosophers account for intuitions with their theories" and "philosophers construct their theories by appealing to intuitions" – and they seem equally incompatible with the clarification/persuasion device interpretation, according to which intuitions are clearly not any kind of building blocks or raw material of theories. Moreover, it is often claimed that intuitions are indispensable in philosophical theorising, that it is impossible for philosophers not to rely on intuitions in one way or another, etc. However under the clarification/persuasion

device interpretation there is nothing indispensable about intuitions. It may be helpful to appeal to them while presenting a theory, but nothing beyond that: giving up on such appeals does not change the substance of argumentation.

The problem with the propositional content reading seems fairly straightforward: if "intuitions" were to mean exclusively "propositional contents of intuitions", then what would be the point of singling out this kind of propositional content? I agree that propositions that merely *happen* to be intuitive are often used as evidence, but I also think that propositions that happen not to be intuitive are often used as evidence, in very much the same fashion: both can serve as starting or intermediate premises of philosophical arguments. Recall the example I used at the beginning of this article: "propositions formulated by carbon-based life forms are used as evidence in philosophy". This statement is not only literally true, it may be true about all philosophical evidence. And yet it sounds odd – this is because being formulated by carbon-based life forms is a property that stems from a historical contingency, not from anything *methodologically salient*. After all, in principle non-carbon-based life forms or sophisticated machines seem perfectly capable of formulating the exact same propositions.

Perhaps it might be objected that if "evidence" is limited only to starting premises, and if a very broad account of "intuitions" is adopted, then all propositions used as evidence in philosophy would count as intuitive, their intuitiveness would be methodologically salient and yet it would not be treated as evidence for those propositions. As I mentioned, some philosophers, like van Inwagen and Lewis, argue that intuitions can be identified simply with beliefs or opinions. On this account, it would be true that philosophers generally try to start their arguments with intuitions – that is with what they think is already accepted by their readers. Surely there is little point in offering an argument which starts with premises that the addressees of the argument are going to reject straight away. Is it, however, really what those who argue that philosophers rely on intuitions have in mind? This is highly implausible. First, when they specify what they mean by "intuition", they practically always opt for a narrower account – typically one that at least involves non-inferentiality. On the narrower account many starting premises of philosophical arguments are not intuitions. Secondly, those who argue that philosophers rely on intuitions typically also argue that there is something uniquely philosophical about this practice, and clearly there is nothing uniquely philosophical about trying to start arguments with premises already accepted by its intended recipients. It is a common feature of arguments as such, not just philosophical arguments, and it would not be reasonable to suppose that this fact is widely overlooked.

Let us now turn to the "evidence not for its content" interpretation. The main reason to disqualify it is the typical choice of examples that illustrate the thesis: we are constantly reminded that the practice of relying on intuitions is best exemplified by trolley cases, Plato's discussion of justice etc. This clearly suggests that intuitions are meant to be used as evidence for their contents. If the fact that "one should not return the weapons" is intuitive can help refute Simonides's theory of justice, it is only because it is used as evidence that one should not return the weapons, and the fact that one should not return the weapons is incompatible with the claim that justice is "truthfulness without qualification, and the giving back of whatever one may have taken from someone else". The same goes for all other examples. Moreover, a number of philosophers make this point explicit. For example, Christopher Daly writes that "those who appeal to intuitions take an intuition that *p* to provide *prima facie* evidence that *p*" (Daly 2015, p. 11). Similar claims have been made by Norbert Paulo (2020, p. 334), Brian Weatherson (2003, pp. 19-20), Alvin Goldman and Joel Pust (1998, p. 181), or James Andow (2017, p. 184).

Finally, there is the evidence in context of discovery reading. One problem with it is that there is only so much we can learn about this kind of evidence from studying philosophical material. When Kekulé first published his findings about the structure of benzene, he did not mention dreaming about a snake seizing its own tail, as it was – and still is – considered inappropriate to include detailed information concerning one's own creative process in scientific publications. Philosophy is not much different in this respect: even if it is more acceptable to include such information, the information is often not there. This means that what is eventually published can be the outcome of many different ways of thinking. It is not impossible to learn something about how a given argument came about, but this often requires reaching beyond strictly philosophical publications to sources such as interviews, letters, diaries, memoirs, private conversations etc. However those who argue that philosophers rely on intuitions hardly ever refer to such sources. The entire evidence that Plato relies on an intuition is to be found in *The Republic*, the entire evidence that Thomson relies on intuitions is to be found in her papers on the trolley problem, and so on.

It might be objected that I am now hoist by my own petard as I myself argue that philosophers use intuitions as evidence in the context of discovery without appealing to extraphilosophical sources I have just mentioned. There is, however, an important difference between what I do and what I argue is hard to explain on the context of discovery interpretation of "philosophers rely on intuitions". I think some *limited* evidence for my claim can be found in philosophical publications. For example, I have pointed out that Thomson discusses reasons to think that it is impermissible to push the fat man off the bridge, but never mentions any reasons to think it is permissible to do so. I think this

counts as evidence that in her thinking process she has not given each option a fair hearing, but rather focused only on reasons behind the intuitive one. I have also argued that way philosophers use the expression "*prima facie*" in their first-order philosophical practice – as opposed to their metaphilosophical claims – does not indicate engaging in anything resembling DE, while it may well indicate engaging in relying on intuitions in the context of discovery. Overall I think that *solely on the basis of what can be found in philosophical sources* we can conclude that the context of discovery hypothesis is always a better explanation of what philosophers do than DE. This, however, is far from offering a full-blown defence of the context of discovery hypothesis, which would require a careful examination of extraphilosophical material. In contrast, those who argue that philosophers rely on intuitions never seem to think that appealing to such material would be suitable, which suggests they are talking about using evidence in the context of justification.

Moreover, the way the intuition thesis is typically worded leads to the same conclusion. We hear that intuitions in philosophy are like observations in science, that theories are judged to be acceptable if they capture intuitions, that refuting a theory amounts to showing it has counterintuitive implications, etc. None of these expressions make much sense on the context of discovery reading. Perhaps it could be objected that I am interpreting at least some of the expressions uncharitably – for example, "rejecting a theory because of its counterintuitive implications" might simply be a shorter and less precise way of saying "rejecting a theory because of reasons discovered by examining its counterintuitive implications". This, however, seems too much of a stretch. Analytic philosophers pride themselves on being exceptionally meticulous. Sometimes their devotion to rigour is even seen as a flaw: apparently it makes academic texts lengthy, dry, tedious and generally unreadable. How plausible is it that in one particular case philosophers universally decide to prioritise conciseness over precision? It is true that claims about relying on intuitions are sometimes little more than passing comments, but even then making it clear that one is not referring to DE does not require a lot of effort. Given the confusion that results from using less precise language, there is simply too much to lose and too little to gain.

Having said that, I concede that on rare occasions claims like "philosophers rely on intuitions" and similar do not necessarily reveal the commitment to DE. I have already cited Hare whose discussion of the problem makes this point clear. Another example, perhaps a slightly less straightforward one, is a point made by Thomas Nagel:

> Given a knockdown argument for an intuitively unacceptable conclusion, one should assume there is probably something wrong with the argument that one cannot detect – though it is also possible that the source of the intuition has been misidentified. If arguments or systematic

> theoretical considerations lead to results that seem intuitively not to make sense, or if a neat solution to a problem does not remove the conviction that the problem is still there, or if a demonstration that some question is unreal leaves us still wanting to ask it, then something is wrong with the argument and more work needs to be done. Often the problem has to reformulated because an adequate answer to the original formulation fails to make the sense of the problem disappear. (Nagel 2002, p. x-xi)

I believe this passage not only can, but most likely should be interpreted in a DE-unfriendly way. According to Nagel it is not the fact that one has an intuition that speaks directly against the argument, but rather the fact that one has an intuition indicates there must be something wrong with the argument, even though one is currently unable to *detect* what it is. All this is compatible with the denial of DE: taking on the argument in a philosophical publication would require identifying and describing the mistake, merely asserting that something is counterintuitive and therefore must be false would be dismissed as a flawed kind of reasoning, not worthy of consideration. Moreover, if the same mistake is not identified by means of intuition, it changes nothing as far as justification goes. One might object that Nagel is putting too much faith in intuition: perhaps many arguments for counterintuitive views are perfectly sound and what he recommends is a wild-goose chase. But whether it is or not is an epistemological question that has no bearing on DE. When he says that philosophers should trust their intuition, and that many of them – himself included – do trust it, he does not endorse DE in any way.

Both Nagel and Hare seem to understand "relying on intuitions" as "relying on intuition-states as evidence of their contents in the context of discovery". Occasionally one can come across other DE-unfriendly readings. For example, Timothy Williamson dismisses the idea of not relying on intuitions as a "non-starter" by pointing out that all reasoning, philosophical and non-philosophical alike, must begin with unsupported premises (Williamson 2018, p. 63). Perhaps he simply identifies "intuitions" with "starting premises". Perhaps what he has in mind is slightly different: "intuitions" are "widely shared beliefs", whose contents are suitable to serve as starting premises. In any case, his reading of "philosophers rely on intuitions" is a truism nobody objects to. Nevertheless, interpretations like that of Williamson, Hare or Nagel seem exceptional. As I have tried to show, in most cases we can find a more or less explicit commitment to DE.

19. Trading on ambiguities

Williamson might not endorse DE himself, however proponents of DE often do something similar: they appeal to evidence for DE-unfriendly readings of "philosophers rely on intuitions" in order to justify DE. In this section I will give one example of equivocating on each of the four ambiguities I have just described.

Tomasz Wysocki takes on Deutsch's argument against the idea that Gettier relies on intuitions to refute the justified true belief theory of knowledge. As I mentioned earlier in this chapter, Deutsch insists that Gettier *justifies* the claim that Smith does not know and that it is the justification for the claim, not the intuitiveness of it, that plays an important role in Gettier's argument. Wysocki argues that if this is the case, we would expect people to believe that Smith does not know solely on the basis of Gettier's reasons for this claim, regardless of whether they find it intuitive. To test this hypothesis, Wysocki conducted an empirical study in which participants who had reported not sharing "the Gettier intuition" were presented with something approximating Gettier's reasons to believe that Smith does not know. It turned out that the reasons did not do the job – the participants remained unpersuaded. Wysocki concludes that this result is strange if we assume that Gettier did not want intuitions to "point to the right answers". (Wysocki 2017, p. 497)

What exactly does Wysocki mean by "pointing to the right answers"? Deutsch makes it clear that he only opposes the idea that Gettier uses intuitions in the state sense as evidence for their contents, and Wysocki makes it clear that he opposes Deutsch. The distinction between the context of discovery and the context of justification is not made explicit in Deutsch's work, however, for reasons outlined above, it is reasonable to assume that what Deutsch has in mind is not the former – which leaves us with DE. Let us now ask: what does Wysocki's evidence show? He argues that despite what Deutsch suggests, Gettier must have relied on intuitions to *persuade* his readers. However, we need to be careful about what exactly Gettier wanted to persuade them *of*. As I mentioned in section 11, Deutsch's choice of words might be rather unfortunate: he says that Gettier "argues for" the claim that Smith does not know, but it does not make much sense to think he wants to change their minds about that. Rather, he tries to *explain what makes it true* that Smith does not know – something he assumes his readers already believe. If Gettier is trying to persuade his readers of anything, it surely is that knowledge is not justified true belief. It seems that Deutsch's terminology has led Wysocki astray.

Suppose, however, that Wysocki's experiment addressed the question of whether Gettier relied on intuitions to achieve his *actual* goal, and suppose that people lacking the intuition that Smith does

not know were not likely to be persuaded that knowledge is not justified true belief by Gettier's argument. Would this result vindicate DE in any way? The answer is: no, it would not. The result would show, at best, that Gettier relied on an intuition *as a persuasion device,* which is not an unreasonable assumption to make. After all, Gettier could have made essentially the same point without bringing up stories about Smith's job application or Brown in Barcelona, or any other stories – he could have stated simply that the justified true belief theory of knowledge is incompatible with the true claim that having a true belief justified by something that does not make it true is not knowing. This, however, would not sound very persuasive. Introducing a scenario in which it is immediately clear, to many people at least, that someone does not know despite having a justified true belief, helps get the point across. But the proposition "Smith does not know" is not indispensable to the argument, it merely serves as a convenient illustration. In short: neither Wysocki's actual experiment, nor the hypothetical experiment I have described shows that Gettier relies on intuitions *as evidence*, let alone evidence in the DE sense.

Let us turn to the second ambiguity. An example of someone who seems guilty of trading on it is Kevin Tobia, who argues that intuitions are used as evidence in philosophy if they are understood in the right way:

> Intuitions can be thought of as special kinds of philosophical *assumptions*, ones to which we are invited to assent that are suitable for argument (for example, providing evidence without further justification) and that are not purely inferentially formed. (Tobia 2015, p. 576)

Tobia gives several examples of such intuitions, one of them being trolley case judgments. He thinks that these judgments "support Thomson's principles", such as "killing one is worse than letting five die" (ibid., p. 585). As I have shown, these are in fact Foot's principles that Thomson rejects, but let us put this aside. Tobia's choice of examples, as well as how he thinks they work, indicate that he is wedded to DE. A full-on endorsement of DE can also be found in a book chapter Tobia co-authored with Stephen Stich:

> A philosopher describes a situation, sometimes real but more often imaginary, and asks whether some of the people or objects or events in the situation described have some philosophically interesting property or relation (…) When things go well, both the philosopher and her audience will agree on an answer, with little or no conscious reflection, and they will take the answer to be *obvious*. The answer will then be used as evidence for or against some philosophical thesis. The mental states that underlie episodes of this sort are paradigm cases of philosophical intuitions. (Stich & Tobia 2016, p. 6)

They go on to present Plato's "one should not return the weapons" as a paradigm case of this practice. This leaves little room for interpretation of what they mean by using intuitions in philosophy.

We can now ask: how does Tobia justify DE? We should note that his proposal is at least partly tautological: on the one hand, he wants to show that intuitions are used as evidence in philosophy, on the other he *defines* intuitions as what is used as evidence in philosophy. It seems that the only thing that rescues his argument from full circularity is the idea of intuitions being *non-inferential.* Propositions that play an important role in philosophical arguments, argues Tobia, often share an interesting feature: they are not inferred, or at the very least least not fully inferred, from other propositions. What exactly does he mean? Unfortunately, he offers no stipulation beyond the claim they "cannot be the mere product of some simple inference" (ibid., p. 582), which is rather vague. One can distinguish between *textual non-inferentiality* – the idea that a proposition is not inferred from other propositions in a given text – and *psychological non-interentiality* – the idea that someone comes to believe a proposition without (consciously) inferring it from other propositions. If Tobia is talking about the former, then his claim seems false. As I have argued, it is not the case that Thomson starts with judgments like "it is permissible to throw the switch" and proceeds with inference to the best explanation – what many take to be her best explanation of independently attested data are actually Thomson's reasons to accept judgments like "it is permissible to throw the switch". I think this can be generalised to Tobia's other examples. However even if I am wrong and these judgments are textually non-inferential, it would only follow they constitute part of the common ground. Either way, Tobia provides us with no evidence for DE.

Perhaps what Tobia has in mind is psychological non-inferentiality. In this case, he is not, strictly speaking, wrong. For example, we can grant that many people presented with different trolley scenarios often form judgments about them without being able to tell why they find the judgments true. However this does not mean that the fact that someone comes to believe a proposition in a certain way plays any role in Thomson's argument, or any other trolley-related argument. All that matters is the content of the proposition, nothing turns on its psychological relations. In other words: what Tobia demonstrates is at best that contents of intuitive propositions are used as evidence in philosophy. This should not be controversial, however according to DE it is the state of being intuitive that is used as evidence, which is something Tobia fails to offer any reasons for.

Let us move on to the evidence for its content vs. evidence for something else ambiguity. At the beginning of this chapter I mentioned Nozick's experience machine thought experiment, which is

often presented as an excellent example of DE. For example, in his article about the scenario, Dan Weijers writes:

> Regardless of the actual causes of intuitions, when the vast majority of philosophers share an intuition, or when a philosopher holds one so strongly that she assumes it is widespread, that intuition is often used as a premise in philosophical arguments. (Weijers 2014, p. 516)

According to Weijers, this is evident in Nozick's case: the intuition that one should not plug oneself into the machine is used as a premise in the argument against "all internalist mental state theories of well-being", which, for our purposes, can be identified with what I earlier described as ethical hedonism. What Weijers says in the passage quoted above might be interpreted in a DE-unfriendly way – namely as simply stating the fact that intuitive judgments are often placed in the common ground. However Weijers believes that the philosophical way to evaluate Nozick's case against hedonism is not to examine reasons for plugging oneself into the machine. Rather, it is to check how many people share this intuition and whether the intuition is produced by an unreliable cognitive process. Someone who rejects DE would argue that these psychological matters are philosophically irrelevant.

How accurate is Weijers's representation of Nozick's argument? We need to be careful answering this question, as Nozick is trying to kill two birds with one stone: he argues against both ethical and psychological hedonism without clearly distinguishing between the two. Undeniably Nozick believes that one should not plug oneself into the machine, and that this judgment is incompatible with ethical hedonism. However we should notice that immediately after introducing the scenario he specifies three reasons to accept the judgment: "we want to *do* certain things, and not just have the experience of doing them", "we want to *be* a certain way, to be a certain sort of person", and "plugging into an experience machine limits us to a man-made reality, to a world no deeper or more important than that which people can construct" (Nozick 1974, p. 43). The expression "we want to" might be slightly misleading. Nozick is not merely stating that, as an empirical fact, we find something valuable – he also believes that it is *objectively* valuable:

> Notice that I am not saying simply that since we desire connection to actuality the experience machine is defective because it does not give us whatever we desire—though the example is useful to show we *do* desire some things in addition to experiences—for that would make "getting whatever you desire" the primary standard. Rather, I am saying that the connection to

> actuality is important whether or not we desire it—that is *why* we desire it—and the experience machine is inadequate because it doesn't give us *that*. (Nozick 1989, pp. 106-7)

Nozick's evidence against ethical hedonism has nothing to do with the fact that the proposition "one should not plug oneself into the machine" is intuitive, and everything to do with the fact that connection to actuality etc. is important. The experience machine story is merely a convenient illustration of how experiencing pleasure does not always go hand in hand with being connected to actuality etc.

Just like with trolley judgments or Gettier judgments, proponents of DE might be tempted to explain away Nozick's reasons against plugging oneself into the machine as abductively inferred from intuitive data. But just like with trolley judgments or Gettier judgments, this interpretation does not make much sense. First and foremost, why would Nozick say that "the connection to actuality is important *whether or not we desire it*", if the claim that the connection to actuality is important is merely his best explanation of our intuitions about the experience machine? Secondly, why would he ask questions like "*why* do we want to do the activities rather than merely to experience them?" (1974, p. 43) On the abductive interpretation, this question has already been answered: wanting to actually do the activities explains our intuitions. Third, why would he use words like "because" the way he uses them? For example, why would he say that the experience machine is inadequate *because* it does not give us what is important? This would be a strange way of presenting an inference from "the machine is inadequate" to "the connection to actuality is important".

One might now ask: if intuitions are irrelevant in Nozick's argument, why is Nozick so concerned about them? For example, why does he urge that "readers who hold they *would* plug in to the machine should notice whether their first impulse was *not* to do so, followed later by the thought that since only experiences could matter, the machine would be all right after all" (1989, p. 105)? To answer, we need to draw a line between Nozick's argument against ethical hedonism and Nozick's argument against psychological hedonism. As for the former, intuitions are not, strictly speaking, irrelevant: they are likely used as clarification and persuasion devices. Surely if the idea of plugging into the machine were universally appealing, Nozick would not have used the scenario to make his point – it would not be useful for persuading or clarifying how what is valuable can be dissociated from what is pleasurable. As for the latter, intuitions are not irrelevant either: they are used as evidence, just *not as evidence for their contents*. According to Nozick, the fact that we have the intuition that we should not plug ourselves into the machine shows that we value something beyond our experiences, however it does not show that we should not plug ourselves into the machine. To

sum up: Weijers is not wrong to think that Nozick appeals to intuitions in some sense, however he is wrong to think that Nozick's appealing to intuitions fits into the DE picture.

Finally, I turn to the context of discovery vs. context of justification ambiguity. Climenhaga's article seems to exemplify this equivocation quite aptly. The paper's stated target is the thesis that philosophers do not use intuitions as evidence – as understood by Cappelen, Deutsch, Molyneux and others (Climenhaga 2018, p. 73). As I have argued, what Cappelen, Deustch or Molyneux reject is first and foremost DE, and DE is a claim about *public* evidence. However Climenhaga is explicitly appealing to the fact that intuitions are used as non-public evidence to make his case:

> Even if philosophers are aware that they use intuitions as evidence, this still does not imply that they will cite their intuitions as evidence in their published work. (Knowingly) using something as evidence oneself is distinct from offering something as evidence in a public discussion. It is true that the former will often make the latter more likely. But in some cases features of the dialectical context will make one unlikely to offer one's private evidence as public evidence. (…)
>
> First, it is dialectically unhelpful to cite as evidence propositions that you know your interlocutors will not accept. So I may not cite P as evidence if I know my interlocutors do not find it intuitive, not because I do not take P to be evidence for my theory, but because it will be dialectically ineffective. Similarly, I may use my intuition that P as evidence for P even while knowing that my interlocutors will not be moved by the fact that I have that intuition. (Climenhaga 2018, p. 98-9)

Climenhaga may well be right that philosophers often believe certain things because they find them intuitive. It may also be true that this fact plays some role in the process of discovering what is then offered as evidence in philosophical publications. This, however, has little to do with DE. What is overwhelmingly understood as "using intuitions as evidence in philosophy" is then left unsupported by Climenhaga's argument.

20. Source of evidence

I have mentioned philosophers who would be happy to endorse DE if only "evidence" in my definition were replaced with "source of evidence" – call it DE(s). They might complain I have painted them and proponents of DE with the same brush by assuming that such replacement would not be substantial. For example, Jennifer Nado argues that Cappelen and Deutsch ignore the

possibility of relying on intuition in philosophy that is similar to relying on perception in everyday conversations. Here is how she illustrates the analogy:

> Suppose I am walking down the hallway with a colleague, and I make the following remark: 'Professor Smith must be in his office – the door is open'. Perception is clearly involved in my belief formation here, but it is quite plausible that I am not 'treating perception as evidence' in Deutsch's sense. The best representation of the argument I have given would contain a premise of the form *the door is open*, but it would not also contain a premise of the form *it visually appears to me that the door is open*, from which *the door is open* is inferred. It likely has not even *occurred* to me that I am currently undergoing a perceptual state; thoughts about one's mental states are simply not that common. The content of my perception is treated as evidence, in Deutsch's sense; but no proposition about observation facts is so treated. (Nado 2017, pp. 392-3)

Similarly, argues Nado, Gettier's argument against the justified true belief theory of knowledge does not have a premise in the form of "it is intuitive that Smith does not know", but Gettier is still treating the intuition that Smith does know as *a source of evidence* for the premise in the form of "Smith does not know".

Is this a valid analogy? Imagine that Nado's colleague challenges her premise by asking "But how do you know that the door is open?". The question may sound odd, but surely her answer would be "Well, I can see it!". I think it undermines Nado's claim that there is no inference from "I can see that the door is open" to "the door is open" in her argument. It might be helpful to distinguish between the argument in the narrow sense, which includes only what is explicitly stated, or only what is occurrent to its proponent, and the argument in the broad sense, which includes the proponent's epistemic basis for what is explicit or occurrent – provided they are willing to share it as a public reason when queried about the argument. The last qualification is important, as not every epistemic basis is always suitable for argumentation. Suppose another reason why Nado believes the door is open is that it had been predicted by a soothsayer the previous day. As her colleague is famously sceptical about soothsaying, Nado is not likely to use this reason to convince her.

Arguments, as I have tried to show, are never dialectically neutral. They are always addressed *to someone*. The author of the argument may not be aware of having a target audience. The audience may not be explicitly mentioned. It may also be purely hypothetical. But it must have a set of beliefs, which determine what is suitable to be used as a starting premise. This is why Nado's broad argument can take several different forms, depending on who she is talking to.

For these reasons I do not differentiate between DE and DE(s) in this thesis, except for this section. It seems to me that Nado's "source of evidence" under scrutiny collapses into "evidence". Suppose, however, that I am wrong and that the relation between the perceptual source of evidence and evidence is never inferential in any sense. Even in this case Nado's analogy seems to break down. Imagine asking Gettier: "How do you know that Smith does not know that the man who will get the job has ten coins in his pocket?" On Nado's view, his answer should have been "I intuit that Smith does not know". But why think that? As I explained in previous sections, Gettier openly answers the question in his article: "for (e) is true in virtue of the number of coins in Smith's pocket, while Smith does not know how many coins are in Smith's pocket, and bases his belief in (e) on a count of the coins in Jones's pocket, whom he falsely believes to be the man who will get the job" (Gettier 1963, p. 122). To defend her analogy, Nado would need to show that Gettier's answer is not actually an answer, and that the real answer is completely different, or at the very least that Gettier gives two different answers, and only one of them is made explicit. This strikes me as extremely implausible. We have no reasons to suspect not only that Gettier himself, but that any philosopher who has ever commented on Gettier's article would have found such appeal to intuition appropriate (see Deutsch 2016). Alternatively, Nado could argue that Gettier's answer would not have been "I intuit that Smith does not know". But then it is unclear what the elements of her analogy are.

Another philosopher who seems to at times endorse a version of DE(s) is Chudnoff. He agrees with Nado that Gettier's argument does not contain a premise in the form of "it is intuitive that Smith does not know", but the believes that the intuitiveness of Gettier's judgment still plays some evidential role. To explain, he introduces the distinction between The Premise View, which he rejects, and The Basis View, which he accepts. According to the former "many philosophical arguments treat the fact that certain contents are intuitive as premises", according to the latter "many philosophical arguments treat certain contents as premises because of the fact that those contents are intuitive." (Chudnoff 2021, p. 174).

I have argued against Chudnoff's idea of enabling intuitions in philosophical writings. To the extent that this idea overlaps with the Basis View, I think the Basis View is false. However what Chudnoff says about the view also seems to echo what I say about relying on intuition as devices of clarification, and relying on intuitions as evidence in the context of discovery. In this sense, the Basis View can turn out to be true, and so can DE(s). My point is therefore not that DE(s) must be false on any interpretation, but rather that any of its potentially true interpretations are far removed from what philosophers have in mind when they say that intuitions are used as evidence.

21. The nature of thought experiments

I have argued that DE can be tested independently of considerations about what "the method of cases" is and how it works. This point is important as it is often not easy to determine whether a particular instance of philosophical practice counts as deploying the method. For one, its proponents cannot agree on whether thought experimentation is a necessary requirement, a sufficient requirement, both, or none. This largely stems from the disagreement over what counts as a thought experiment. For example, according to Pust descriptions of actual situations often "elicit intuitions", and yet they are not thought experiments, which is why the practice of relying on intuitions in philosophy can be thought experiment-free. On the other hand, one can occasionally come across examples of referring to such descriptions as "a kind of natural thought experiments" (Kolodny 2017, p. 101), or "real thought experiments" (Jeske 2018, p. 21), which suggests that perhaps whether something can be classified as a thought experiment is more a matter of how it is used rather than what it describes.

The grey area is not limited to actual situations. For example, Michael T. Stuart, Yiftach Fehige and James Robert Brown have recently suggested that works of art such as Sophocles's plays, Stanley Kubrick's films or even Jackson Pollock's paintings can be "fruitfully characterised" as thought experiments. They also believe that thought experiments permeate our everyday thinking, which includes "planning out a busy day [...]; figuring out how best to get from one place to another, deciding what to eat, etc" (Stuart et al. 2018, p. 2). One might wonder what does not count as a thought experiment on this view.

There are also, of course, less inclusive theories. Several philosophers have suggested that thought experiments are *narratives* of some sort (an overview of these claims can be found in Souder 2003, pp. 208-9). If this is accurate, we could rule out at least such non-propositional phenomena as abstract paintings. But is it? Counterexamples are not too difficult to find. Take Hume's missing shade of blue: we are asked to imagine a man who has never seen one particular shade of blue. Other shades are placed before him, organised from the deepest to the lightest, with a blank spot in the place of the unseen one (Hume 1748/1999, pp. 98-9). This does not look like a narrative. A narrative requires a series of events, linked to each other in some meaningful way, but what Hume is describing here is an isolated, static situation. The same can be said about several other typical thought experiments, like Black's two spheres, Chalmers's zombies, Laplace's demon, Avicenna's floating man or Aquinas's cannibal.

John Norton agrees that not all thought experiments are narratives, but argues there is another thing they all have in common: they are essentially arguments "disguised in some vivid picturesque or narrative form" (Norton 2004, p. 1139). This idea would allow us to exclude not only non-propositional phenomena, but possibly also a fair number of images or stories that could not be adequately translated into any kind of premise-conclusion structure. However Norton's theory has been attacked by Tamar Szabó Gendler, who argues that it is not always possible to replace a thought experiment with an argument without losing at least some of its "demonstrative force" (Gendler 2010). In any case, Norton and Gendler seem to agree that thought experiments must be, in some way, vivid, they only disagree about the epistemic role that their vividness plays. But how should we understand this characteristic? The most plausible option seems to appeal to the idea of *mental image*: conducting a thought experiment involves some sort of "seeing in the mind's eye". However it is very far from clear what mental imagery is and how it works. For example, is it a form of subjective experience? Is it a form of mental representation? What gives rise to it? How does it differ from perceptual experience? How much of it is subject to voluntary control, and in what way? (see Thomas 2021) Depending on how these questions are answered, different phenomena can count or not count as thought experiments.

Another fairly restrictive – or perhaps only seemingly restrictive – theory has been proposed by Rachel Cooper, who argues that thought experimentation is about applying a certain degree of *intellectual rigour* to counterfactuals:

> When a thought experimenter is faced with a "what if" question, she attempts to answer it in a rigorous fashion. She follows through all the relevant implications of altering one part of her worldview and attempts to construct a coherent model of the situation she is imagining. The rigor with which thought experimenters attempt to answer "what if" questions is what differentiates thought experiments from daydreams and much fiction. In a daydream I might lazily imagine being prime minister – there I am bossing everyone about, issuing edicts that extend university vacations, and so on. In a thought experiment such slapdash imaginings are not permitted. If I conduct a thought experiment in which I dictate that university vacations should be extended, then I am obligated to at least sketch a coherent model of the situation – the courses must be correspondingly shorter, degrees must be longer, funding per student greater, and so on. (Cooper 2005, p. 337)

It is unclear to me, however, how much fiction would not count as thought experimentation on this account. What if one morning a young sales representative woke up and realised he had been transformed into a giant insect? Kafka sketches a coherent model of the situation: the character's

sister removes furniture from his room, as it is no longer of any use to him, and he needs more crawling space. His father starts looking for a job, his mother has to sell her jewellery and dismiss their maid, and so on. It seems easy to describe virtually any fictional story this way.

The expression "*what if* question" might be misleading here: in one sense, thought experimenters ask this question, but in another, they do not. For example, Saul Kripke invites us to imagine a man named Schmidt who discovers the proof of the incompleteness of arithmetic. Then, after Schmidt's mysterious death, a man named Gödel gets hold of the manuscript, presents Schmidt's proof as his own and becomes famous (Kripke 1980, pp. 83-4). The question that Kripke asks is not exactly "What if it were the case?", as this question has an infinite number of answers, nearly all of which are irrelevant to the point Kripke is trying to make. For example, one might respond that Gödel is not a decent person, or that he might have had something to do with Schmidt's death. Rather, Kripke's question is "When people who associate the name "Gödel" solely with "the man who proved the incompleteness of arithmetic" say "Gödel", do they refer to Gödel or to Schmidt?". Answering this question is the only point of the scenario. And this is what makes the difference between Kafka's story and Kripke's story: it is not about posing a "what if" question, or trying to construct a coherent model of a world. It is about asking a very *specific* question, relevant to a very specific problem.

But what does it mean for a question to be specific enough in this context? We may grant that the "what if" in Kafka's *Metamorphosis* is too open-ended to consider it a thought experiment, but what about narratives whose point is more clearly identifiable? Take something like Aesop's *The Fox and the Crow.* There is an obvious moral to the story: "beware of your flatterers". However there still seems to be an important difference between the point of Kripke's story and that of Aesop's. Perhaps it has to do with the latter's *didactic* nature: its purpose is more to teach a lesson and less to help understand a problem. Maybe this kind of moralising is not welcome in a thought experiment? On the other hand, Edward Davenport argues that it is "common sense" to think of Aesop's fables, along with many other classical works of fiction, as thought experiments, as they all "dramatize certain hypotheses about society and enable us to see the logical conceptual implications of these hypotheses" (1983, p. 284).

Yet another way of characterising thought experiments would be to liken them to empirical experiments: in both cases there would be an independent and dependent variable and an attempt to isolate them – the point of which would be see how the two interact (Miščević 2021, pp. 8-9, Baggini 2006, p. ix). Many thought experiments seem to conform to this pattern quite well. For example, David Boonin is interested in the relation between being a member of homo sapiens and

having a right to life. He asks us to imagine a group of people, all with an undeniable right to life, and then introduces a twist: one of them turns out to be an alien who resembles a human being very closely in terms of his appearance, behaviour, consciousness, mental life etc. The only difference is his DNA. We would not, argues Boonin, renounce his right not to be killed in the face of this revelation, which means that granting the right does not depend on species membership (Boonin 2003, pp. 22-23). The way that Boonin creates two experimental setups – one where an independent variable (being homo sapiens) is present, and one where it is absent, other things being equal – clearly resembles what scientists do in their empirical testing, for instance in a randomised control trial. The same might be said about Kripke's Gödel scenario: here the two variables that the author tries to control would be "being referred to with a name" and "satisfying a description associated with this name".

On the other hand, some thought experiments do not fit into the picture. Take Plato's cave: it seems impossible to analyse it in terms of independent and dependent variables. It might be replied that it is an allegory, and allegories are not, strictly speaking, thought experiments, even though they are often characterised as such. But this raises the question: why are the two so often conflated? Moreover, there are number of thought experiments that are clearly not allegories and yet still defy the independent and dependent variables characterisation: think of Rawls's veil of ignorance or scenarios illustrating different puzzles and paradoxes, like Buridan's ass, the ship of Theseus, Parfit's amoeba-people, the liar paradox, the Russell-Zermelo paradox, or Newcomb's paradox.

All in all, the debate on thought experimentation is moot. Not only is there a lot of disagreement over its nature, but also over which elements of philosophical practice fall under the scope of the concept. Philosophical writing is replete with dubious cases. For example, are remarks like "I think my opponents would agree" thought experiments? Is everything that follows "Suppose that..." a thought experiment? Every example of hypothetical reasoning? Every narrative? Every piece of fiction? Everything that triggers a mental image? There is no definite answer. This creates a problem for someone who wants to test DE via "the method of cases": as it is far from clear what counts as a thought experiment, it is also far from clear what counts as an example of using the method.

22. More problems with "the method of cases"

Furthermore, even if we can be sure we are dealing with a thought experiment, it is often difficult to identify the intuition "elicited" by it. A good illustration of this problem is Newcomb's paradox. In

its original published formulation Robert Nozick imagines a creature capable of predicting your behaviour with great accuracy. It offers you a choice between taking what is inside box B and taking what is inside boxes A and B. Box A contains 1000 dollars. Box B contains either 1 million dollars or nothing. The creature had predicted your choice, and then either put or did not put 1 million into box B, depending on whether it had predicted you would take one box or two boxes, respectively. Now you have to make a choice.

Nozick writes that after putting the problem to many of his friends and students, roughly half of them thought it was "perfectly clear and obvious" that they should take one box, and the other half thought the same about taking two boxes (Nozick 1969, p. 117). My own feeling is rather different: both options seem right to me at the same time, even though I believe only one of them can be right. In any case, it looks like we are not dealing with an example of the method of cases here, as it is impossible to pinpoint the intuition that the case is supposed to elicit.

However some proponents of DE disagree. Pust argues that "two boxes" is the presumed intuitive answer (Pust 2000, p. 8-9). Needless to say, I deeply disagree with this interpretation – I do not think anyone treats either of the responses as a datum to be explained by an adequate theory. My point here, however, is that it can be very problematic to try to test DE by first finding a thought experiment, then identifying the intuition it is supposed to elicit, and then checking whether this intuition is treated the way DE predicts it is. Many thought experiments do not seem to fit into the "eliciting an intuition" picture, at least not in any obvious way – this includes virtually all those labelled as paradoxes, and also some of the most celebrated philosophical scenarios, like Descartes's evil demon or the aforementioned Plato's cave. Perhaps they should be excluded, but we do not seem to have a reliable criterion for excluding them. This problem seems even worse for cases that do not involve thought experimentation. How are we to distinguish between intuition-eliciting descriptions and non intuition-eliciting ones?

There is another reason why "the method of cases" might not be the full story of how intuitions are used as evidence in philosophy, assuming DE is true. As some proponents of DE argue, while the method is only concerned with intuitions about particular cases, philosophers also rely on intuitions about more general or abstract claims. George Bealer argues that philosophers can appeal to intuitions like "if *p* then not not *p*" (Bealer 1998, p. 205). Ernest Sosa expresses a similar view when he says that philosophers appeal to intuitions concerning "not only hypothetical cases, but also principles in their own right" (Sosa 2009, p. 10). Michael Strevens argues that "some intuitions are not in any significant sense case judgments: the intuition that time flows or the intuition that I am conscious or that my body is extended in space, for example", and yet these intuitions "might surely play another

kind of role in grounding philosophical knowledge" (Strevens 2019, p. 3). Pust lists a number of examples, including the intuitiveness of consequentialism, "suitably formulated" (Pust 2000, p. 11-12). He believes that while intuitions about particular cases are usually treated as "better evidence in virtue of their greater determinateness and clarity", both kinds are taken into account in a philosophical enquiry.

Some have argued that "the method of cases" does not pick out anything methodologically distinct and the word "method" is a misnomer in this context (Cappelen 2012, pp. 190-1, Cappelen & Deutsch 2018). I am largely sympathetic to this view, with one caveat: if "the method of cases" is identified with the method of using *counterexamples*, understood simply as questioning general claims by putting forward more particular claims that conflict with them, then undeniably there is such a thing as the method of cases. This is not, however, what philosophers writing about this method seem to have in mind. For example, in metaphysics radioactive decay can be used as a counterexample to the claim that every event has a cause, or in philosophy of mind the phenomenon of blindsight can be used as a counterexample to the claim that all perception must be conscious. I doubt whether such intuition-free and thought experiment-free cases would be considered by many to be the examples of utilising the method. Moreover, on my account there is nothing distinctively philosophical about it: counterexamples are constantly used in everyday conversations or in science, and their role outside philosophy is no less prominent, which is not how philosophers typically understand it. For example, Malmgren writes that the method is "what demarcates philosophy from other academic disciplines, specifically from (other) sciences" (Malmgren 2011, p. 263).

My case against DE, however, does not hinge on my views about whether "the method of cases" exists or what role it plays. I am merely pointing out that looking into examples of using the method to test DE amounts to a fairly limited approach: one is practically restricted to the list of two dozen or so cases that have explicitly been labelled as such by proponents of DE. However, it is claimed that the practice of relying on intuitions is extremely widespread, which suggests we are likely to find examples of DE in a randomly selected philosophical text. In the remaining part of this thesis I will then go off the beaten track and focus on how intuitions are treated elsewhere.

CHAPTER 2: Defending the orthodoxy

1. The argument from inevitability

Claims about using intuitions as evidence in philosophy are rarely defended – typically they are presented as obvious and undeniable. This, however, does not mean that arguments for these claims do not exist. I have been able to tease out seven. Arguably the most popular is the one according to which there is no meaningful alternative to relying on intuitions in philosophy. It also seems to be the weakest. As I tried to show in the previous chapter, what is typically meant by "relying on intuitions", "using intuitions as evidence" etc. is DE, and denying DE amounts to maintaining that philosophers do not infer any *p* from "*p* is intuitive" in the context of justification. But, of course, they still make all sorts of other inferences, and all of them constitute an alternative to relying on intuitions, as it is commonly understood.

The remaining five arguments – I am going to call them "from intuition-talk", "from endorsement", "from non-coincidence", "from error theories", "from counterexample diversity" and "from intuitionism"– require a more detailed replies.

2. The argument from intuition-talk

A large portion of Cappelen's book is devoted to what he calls "the argument from intuition-talk". The idea behind it that philosophers extensively use words like "intuition", "intuitively" etc., which gives us a reason to think they rely on intuitions, in the DE-sense. Cappelen believes that this is unfounded. He argues that words like "intuitively" sometimes constitute an unnecessary embellishment which can be safely removed without changing the meaning of the text, sometimes they refer to "judgments or understandings that are (or can be) reached with relatively little reflection or reasoning", and sometimes to "a conclusion reached prior to or independently of an investigation of the question under discussion" (Cappelen 2012, p. 2012).

I find Cappelen's arguments entirely convincing, however his list does not seem exhaustive. For example, Cian Dorr points out that it is hard to avoid using intuition-talk in philosophy "without seeming to bully one's readers" (Dorr 2010). Qualifiers like "intuitively" or "it seems that" are part of the stylistic etiquette of the profession – but, of course, this fact lends no support to DE. It may

seem that what Dorr has in mind is similar to Cappelen's idea of "simple removal". However Cappelen insists that removing "intuitively" of the kind he describes makes the text clearer and more rigorous, and it is dubious whether removing "intuitively" which makes the text gentler and more polite would have this effect. Being blunt often diverts the addressees' attention from the substance of the utterance and focuses it on the speaker's attitude. For this reason, saying "intuitively, *p*", "it seems that *p*" etc. instead of simply "*p*" can somewhat paradoxically make it clearer that what one means is simply *p*.

Philosophers also occasionally use the word "intuition" to mean simply "opinion". Take this passage from a recent book on higher education by Jason Brennan and Phillip Magness:

> So, grades *could* mean any number of things. This gives rise to a normative question: What *ought* grades signify? Frankly, we don't have any strong intuitions. Of the nine or more possible meanings delineated in the prior list, it's not obvious which meaning grades ought to have. You could probably make a case for each of them. (Brennan & Magness 2019, p. 120)

Here "intuitions" definitely does not denote any kind of noninferential attitudes, attitudes accompanied by a special phenomenology, attitudes with modal content, or any other properties picked out by more restrictive theories of the intuitive. And it should be clear that using the word "intuition" in this sense has nothing to do with engaging in anything that can resemble DE.

The passage quoted above is an example of using intuition-talk generously, which appears to be a recent phenomenon. It is hard to deny that "intuition" has become something of a new philosophical buzzword. James Andow conducted a study which revealed that "intuition" and its cognates can be found in 53.6% of philosophy articles indexed by JSTOR in 2000s. The figure drops to 50.5 in 1990s, 47.5 in 1980s, 44.1 in 1970s, 34.9 in 1960s, 32.7 in 1950s, all the way to 21.7 in 1900s. A similarly rapid increase can be observed in many other disciplines, however the extent of using the term outside philosophy is lower. For example, the figure for 2000s was 39.2 for linguistics, 30.8 for law, 24.0 for anthropology, 15.5 for mathematics, and 6.4 for astronomy (Andow 2015, p. 197). One can sometimes get the impression that sprinkling one's text with intuition-terminology is a simple way of making it sound more philosophical. Words like "opinion", "belief", "claim", "thesis", "idea", "view", "insight", "proposition", "contention", "conclusion", "remark", "suggestion" etc. all seem to be giving way to "intuition". I find this tendency rather unfortunate: the term is becoming increasingly vague and often misleading. However, regardless of what one makes of the current usage of the term in academic literature, we have no reason to think that it can be explained by the fact that philosophers rely on intuitions in any DE-friendly sense.

3. The argument from endorsement

A great number of philosophers, including many eminent ones, are convinced that they frequently appeal to intuition. Some have suggested that this widespread endorsement of the intuition-based view gives us a reason to accept it (Knobe & Nichols 2017). In the previous chapter I argued that what is meant by "appealing to intuitions" is typically DE, and that DE is false. This raises the question: how is it possible that so many philosophers are fundamentally mistaken abut what their do?

Two varieties of the argument from endorsement can be distinguished. According to one, DE is endorsed with respect to one's own philosophical practice. According to the other, DE is endorsed as a more general about what philosophers do, not necessarily including the endorser herself. The first might appear more difficult to refute. However, we should note that it is not uncommon for philosophers to make questionable claims about their own methods. Take one of the most well-known examples of methodological self-reflection of the 20th century philosophy – the penultimate paragraph of Wittgenstein's *Tractatus Logico-Philosophicus*:

> My propositions serve as elucidations in the following way: anyone who understands me eventually recognizes them as nonsensical, when he has used them—as steps—to climb up beyond them. (He must, so to speak, throw away the ladder after he has climbed up it.) He must transcend these propositions, and then he will see the world aright. (Wittgenstein 1921/2001, p. 89)

A number of commentators have argued that this characterisation simply cannot be accurate: if Wittgenstein's propositions, properly understood, are nonsensical, how can they elucidate anything? (Horwich 2012, pp. 90-95) Of course, there might still be a way of making sense of these remarks – however to take it as a given that Wittgenstein must be correct would be to treat him as a prophet, not as a philosopher.

Or consider a more recent example: neurophilosophy. According to Patricia Churchland it can be defined as a research programme which "explores the impact of discoveries in neuroscience on a range of traditional philosophical questions about the nature of the mind" (Churchland 2017, p. 72). Churchland is convinced that her own work fits the description perfectly. For example, she argues that dualism about the mind is undermined by discoveries about different kinds of brain damages, effects of various drugs on brain functioning, results of studies on the so-called split brain patients,

or the fact that our brains evolved via natural selection (ibid., pp. 74-88). However Churchland's critics are not always happy with this characterisation. This is because they tend to think that neuroscientific data is neutral between physicalism and dualism – or at least dualism in some form (Goff 2017, pp. 2-11). If this is true, Churchland's project would be better described as more traditional metaphysics peppered with largely irrelevant neuroscientific details. This example shows that it is sometimes impossible to disentangle questioning one's first-order philosophical views from questioning their methodological self-understanding. And since the former is considered perfectly legitimate, so should be the latter.

In the previous chapter, I offered reasons to reject DE. I also argued that philosophers often equivocate between different senses of "relying on intuitions" – to justify the false sense, they invoke various true senses of the expression. This, I think, explains why DE is so popular and so widely endorsed both as a claim about one's own and others' philosophical practice. We do not need to assume that those who endorse DE are irrational – rather, they make an understandable mistake.

4. The argument from non-coincidence

In the previous chapter, I mentioned Climenhaga's "no coincidence" argument for DE: what philosophers believe are instances of knowledge, reference, consciousness, causation, justice etc. generally accord with our intuitions about what counts as knowledge, reference, consciousness, causation, justice etc. This fact, the argument goes, is better explained by DE than by coincidence. I pointed out that one weakness of this argument is that it blurs the line between public and private evidence: what philosophers offer as evidence can be divorced from what they believe, and what they believe is often not included in what they offer as evidence. But the argument can be easily fixed to overcome this weakness: instead of saying that there is a significant overlap between philosophers' beliefs and our intuitions, we can say that there is a significant overlap between what follows from respectable philosophical theories and our intuitions, and that DE explains this fact better than coincidence.

In my view, however, neither DE nor coincidence is the right answer. The best explanation comes from the combination of two facts. First, intuitions about what counts as knowledge, reference, justice etc. are often placed in the common ground. Philosophers start their arguments with claims like "infants have a right to life" as these claims are suitable *for dialectical reasons*. This has nothing to with arguing that infants have a right to life because it is intuitive that they do.

Second, when philosophers do not use intuitions as argumentative starting points, they can rely on them as evidence *in the context of discovery* – that is, they can pay more attention to intuitive judgments and try to find good reasons to accept them. For example, it is not unlikely that Thomson had paid more attention to the judgment "It is morally impermissible to push the fat man off the footbridge" than to various nonintuitive judgments about the same case, such as "it is morally permissible to push the fat man off the footbridge" or "it is morally permissible to push the fat man off the footbridge only if his name is Kevin". As she thought she was able to secure good evidence in favour of the intuitive judgment, she did not bother to examine the nonintuitive ones in a similar fashion. This, however, does not mean that she tried to adhere to a standard according to which a theory that implies it is impermissible to push is superior to a theory that implies it is permissible to push, other things being equal – which is what proponents of DE would have us believe.

It might be objected that my "context of discovery" reply faces a dilemma: either there usually is good evidence to support intuitive judgments, or there is not. If there is not, we would expect philosophers to turn to nonintuitive judgments after failing to find good evidence behind the intuitive ones – it does not seem plausible that they would simply give up and abandon their projects whenever unable to account for what seems true. In this case, however, most respectable theories of knowledge, reference etc. would not accord with our intuitions about what counts as knowledge, reference etc. There would be plenty of theories that often lead to counterintuitive verdicts – however this, at least according to Climenhaga, is not the case. We are then left with the other horn of the dilemma: there usually is good evidence to support intuitive judgments. In this case, however, it seems strange that the judgments are not *used as* evidence by philosophers. Why would they refuse to rely on a readily available source of evidence for their theories?

One answer would be that while there might be a lot of good evidence for counterintuitive claims, it is simply more difficult to discover it. An analogy with finding evidence in science may be useful here. Some philosophers of science use the distinction between new evidence and old evidence: sometimes discovering evidence for theory T can be a matter of discovering a new piece of information that is in line with T, but sometimes it can be a matter of discovering that a long-known piece of information is implied by T. Clark Glymour argues that the latter kind abounds in science:

> Scientists commonly argue for their theories from evidence known long before the theories were introduced. Copernicus argued for his theory using observations made over the course of millenia, not on the basis of any startling new predictions derived from the theory, and presumably it was on the basis of such arguments that he won the adherence of his early disciples. Newton argued for universal gravitation using Kepler's second and third laws,

> established before the *Principia* was published. The argument that Einstein gave in 1915 for his gravitational field equations was that they explained the anomalous advance of the perihelion of Mercury, established more than half a century earlier. (Glymour 1980, p. 86)

Note that revisiting old evidence in science often leads to the abandonment of a more commonsensical theory in favour of a less commonsensical one: geocentrism is replaced with heliocentrism, Newtonian physics with the relativity theory, and so forth. It takes time and effort to realise that what we know about the world supports a rather strange picture of the world. This is partly because evidence is often ambiguous: it is far from clear which theory it points to or which theories it is compatible with. For example, after the precession of the perihelion of Mercury was observed, physicists did not suddenly abandon Newtonian physics, even though they quickly noticed that the planet's behaviour was at odds with what the theory predicted. Some of them postulated a new planet called Vulcan between Mercury and the Sun, others simply hoped that the anomaly would eventually be explained one way or another without the need to reject Newton's laws of motion.

Philosophers, just like scientists, or possibly even more so, appeal to old evidence. We might come across the odd theory developed in response to new experimental data coming in (Doris 2002 or Machery 2017 might serve as an example), however most philosophical theories are not born this way. Think of Plato pointing to the fact that nobody has ever seen two things perfectly equal to each other to support his theory of forms, Berkeley pointing to the fact that we cannot perceive size, shape or motion without perceiving colour at the same time to support his immaterialism, J. L. Mackie pointing to the fact that different cultures disagree about polygamy to support his error theory of morality – examples are plentiful and easy to find. Also just like in science, it might be easier to appeal to old evidence to justify a philosophical theory that does not run counter to our intuitions. For example, Climenhaga may be right that most, if not all, non-sceptical epistemological theories account for the fact that Smith does not know that the man who will get the job has ten coins in his pocket, but this might be because supporting this claim with evidence is generally easier that supporting its negation. Perhaps, as epistemology matures, future theories will imply that Smith does in fact know, just like modern physics implies that, for example, simultaneity is observer-relative, which is highly counterintuitive.

A more radical response to the dilemma is also possible. One may argue that it rests on a problematic assumption, what Williamson calls Evidence Neutrality:

> As far as possible, we want evidence to play the role of a neutral arbiter between rival theories. Although the complete elimination of accidental mistakes and confusions is virtually impossible, we might hope that whether a proposition constitutes evidence is *in principle* uncontentiously decidable, in the sense that a community of inquirers can always in principle achieve common knowledge as to whether any given proposition constitutes evidence for the inquiry. (Williamson 2022, p. 212)

Williamson argues that Evidence Neutrality is generally false. If this is correct with respect to evidence offered for case judgments, then relying on intuitions as evidence in the context of discovery can work as a sort of self-fulfilling prophecy: philosophers find what they take to be good evidence for intuitive judgments only because they have assumed there is good evidence to be found in the first place. This could explain the fact that philosophical theories often account for intuitions while never being supported with intuitions in the DE sense.

Finally, the first horn of the dilemma is not impossible to take. It is questionable whether philosophical theories generally account for our intuitions: there are plenty that do not, and this fact is often underappreciated in metaphilosophical debates. In the next chapter I am going to list a number of such theories, and analyse some of them in chapter 4.

5. The argument from error theories

Climenhaga admits that that there is no *full* coincidence between what philosophical theories imply and what our intuitions tell us – occasionally philosophers dismiss intuitions. But this is no evidence against DE, because whenever they do, they try to explain why intuition leads people astray in a particular case. Were intuitions not to be used as evidence, philosophers would not bother to offer such explanations. Far from not undermining DE, the way that intuitions are dismissed actually supports it. To illustrate this point, Climenhaga comes up with the following thought experiment:

> Suppose we have two philosophers, Deon and Connie. Deon takes his intuition-state E to support theory T. Connie objects by saying that there's a better explanation of E than T – namely, error theory T*. T* implies that E was to be expected whether or not T, so that E does not substantially confirm T. For example, consider the footbridge variant of the trolley thought experiments, where a trolley is hurtling down a track and about to run over and kill five people (Thomson 1985). (...) Here we can let E be the intuition that it is wrong to push the man. We could imagine T to be some elaborate deontological theory, but for simplicity's sake let's

> consider the limiting case in which T is simply the proposition that it is wrong to push the man. Deon proposes that E supports T because T is the simplest explanation of E: Deon has the intuition that it is wrong to push the man because it is wrong to push the man. Connie, a consequentialist, proposes T*: in most situations like the above in salient respects, pushing the man would have overall worse consequences than not pushing him. She then claims that, although T would lead us to expect E, so would T*, because our intuitions about these kinds of cases respond mainly to coarse-grained features of the cases (e.g., that the case involves pushing a man to his death). Thus, E is not as good evidence for T as Deon thought, because it is equally well explained by T*. (Climenhaga 2017, pp. 85-6)

According to Climenhaga, Connie offers what he calls an "error theory" for Deon's intuition to invalidate it as a source of evidence for its content, and what she does only makes sense if we accept that Deon is relying on his intuition. But how far does this scenario take us? To answer, we need to determine how *realistic* it is. Surely, if Deon publicly offers his intuition that it is wrong to push the man off the bridge as evidence that it is wrong to push the man off the bridge, then he engages in DE. However Deon is an imaginary philosopher, not an actual one, and DE is a thesis about actual philosophy. As I argued in the previous chapter, in the real world Thomson does not use the footbridge intuition as evidence for anything – at least not in the "context of justification" sense, required by DE. Moreover, it is highly implausible that anyone who has ever discussed the example tried to treat the claim that it is wrong to push the fat man as some sort of intuition-attested datum to be explained, or explained away, by a theory. By the same token, in the real world Deon would not even be able to publish his defence of T, as he has nothing of philosophical value to support it, and Connie would have nothing to respond to.

Furthermore, Connie's defence of consequentialism is manifestly *circular*. Her "error theory" hinges on the claim that the worse action is the one what has worse overall consequences – without this assumption, there is no error, and she has nothing to undermine the view that pushing the man off the bridge is wrong. This illustrates a wider problem: if DE is true and if philosophers debunk intuitions to argue for their theories, then all they do is arguing in a circle. But this is implausible. There are many things one can accuse academic philosophers of, but being unthoughtful is not one of them. And it would be highly unthoughtful to constantly argue that a theory is correct because it accounts for correct intuitions, and what makes these intuitions correct is the fact they conform to a correct theory.

In any case, to make his point, Climenhaga would need to provide us with some non-hypothetical examples. And to be fair to him, he offers three such cases. First, some epistemological pragamaticists argue that sometimes what seems to us to be a case of knowledge is not a case of

knowledge, or the other way around, because we are confusing utterances that are true with utterances that are pragmatically appropriate. Secondly, Paul Grice does something very similar with our intuition that morally neutral actions cannot be voluntary or involuntary. Finally, Derek Parfit questions what he thinks is a false judgment that desires provide us with reasons for action. One of the several arguments he puts forward is that when certain state of affairs are desired, we often have desire-independent reasons to bring them about, and that is how desires get associated with reasons to act in our minds (ibid., pp. 84-5). However, as Climenhaga himself admits, all these arguments can be plausibly interpreted in a DE-unfriendly way. For example, Parfit's argument may be interpreted as simply exposing flaws in his opponents' reasoning – not as suggesting that his opponents' intuition-states cannot justify their own contents as they come from dubious sources. In the next chapter I am going to analyse this example in more detail to show that Parfit's method cannot plausibly be interpreted in any DE-compatible way.

6. The argument from counterexample diversity

Suppose that one philosopher finds two intuitive counterexamples to a theory, but they describe two very similar situations. Then suppose another philosopher finds two intuitive counterexamples to the same theory, but they describe two situations very unlike each other. The first philosopher would be expected to feel less confident in rejecting the theory than the second one, and DE best accounts for this difference in confidence. This is because several similar intuitions can be explained away with the same error theory, but to explain away several diverse intuitions one needs several different error theories, and multiplying error theories to defend one's view seems *ad hoc* – "the more plausible explanation is that the intuitions are correct" (ibid., p. 96). This, in a nutshell, is Climenhaga's third argument.

His illustration of this point is W. D. Ross's criticism of consequentialism. According to Climenhaga, Ross feels very confident in rejecting consequentialism as his anti-consequentialist intuitions are quite diverse:

> In arguing against consequentialism, Ross (1930: ch. II) presents three counterexamples to the claim that it is always wrong for someone 'to do an act which would produce consequences less good than those which would be produced by some other act in his power'. The first counterexample involves choosing between fulfilling a promise to A and bringing about slightly more good to B, to whom one has made no promise. In the second case, one is choosing between benefiting A by fulfilling one's promise to him or by doing some other act that would

> benefit A slightly more. In the third case, one is again choosing between helping A or helping B slightly more, but now 'A is a very good and B a very bad man'. In each case Ross thinks it intuitively clear that we ought to choose the first option, contra consequentialism. (ibid., p. 92)

But this seems to be a caricature of Ross's argument. First, by offering his three counterexamples, Ross is not arguing against consequentialism. He is arguing against the claim that "there is [a] self-evident connexion between the attributes 'right' and 'optimific'" (Ross 1930/2007, p. 35) – that is, that it is possible to immediately and *a priori* apprehend that "right" is coextensive with "brings about the best possible consequences". This claim might be used to defend consequentialism, but is by no means entailed by it. Many consequentialists would be happy to admit there is no such connection.

Of course, if we identify "self-evident" with "intuitive", the claim that Ross is challenging is a psychological claim about how our intuition works, and he seems to be doing so by appealing to facts about what we find intuitive. Therefore Climenhaga may be wrong about what Ross is trying to attack, but he is not wrong about *how* he tries to attack it – namely by relying on intuitions. However note that on this account Ross is not relying on intuitions as evidence *of their own contents*. He is not saying that it is intuitive that fulfilling a promise to A is better than benefitting B slightly more, therefore fulfilling a promise to A is better than benefiting B slightly more. He is saying that it is intuitive that fulfilling a promise to A is better than benefiting B slightly more, therefore it is not true that we intuitively take "right" to mean "brings about the best possible consequences".

Moreover, even this is not entirely correct. Ross actually *justifies* each of his three statements, and he does so without ever appealing to the fact they are intuitive. First, he points out that a promise "constitutes a serious moral limitation to our freedom of action", then, in the second case, that A should be given priority as there is a "*prima facie* duty to do him the particular service I have promised to do him", and finally, in the third case, that there is a "*prima facie* duty of justice, i.e. of producing a distribution of goods in proportion to merit" (ibid., p. 35). It is these more abstract claims about *prima facie* duties that can be characterised as intuitions being treated as evidence, if anything can. But this is a side remark. The main point is that Ross is not treating any intuitions as evidence of their contents. Hence it is impossible to prove him wrong by offering "error theories" for his intuitions – whether his argument works does not depend on whether the intuitions he appeals to are correct.

On Climenhaga's view, the standard method of criticising theories like consequentialism is to point to their counterintuitive implications: it seems to us that in a given situation breaking a promise

would be wrong, but consequentialism implies it would not be wrong, therefore consequentialism is not correct. The more intuitions a theory clashes with and the more diverse these intuitions are, the more problematic the theory becomes. However this view is simply untenable in the face of well-known facts about the nature of philosophical inquiry. Note that consequentialists themselves are often happy to list counterintuitive implications of their theory. Take Peter Singer, who argues that the form of consequentialism he endorses implies a wide range of claims that many people find strange, if not outrageous: this includes claims about infanticide (Kuhse & Singer 1985), the use of non-human animals (Singer 2002), or international aid (Singer 2009). Is Singer arguing against himself? Of course not – he thinks it is philosophically irrelevant whether his conclusions merely *seem false* to anyone. Does anyone hold it against Singer that his conclusions just seem false? Outside philosophical literature – perhaps yes, but not within it. More specifically, no philosopher appears to be more confident in rejecting Singer's theory because its counterintuitive implications are numerous and diverse, and less confident in rejecting a theory whose counterintuitive implications are limited to just one type of cases. Surely, there are plenty of philosophers who deeply disagree with Singer about all the aforementioned claims, but they always give *reasons* to reject these claims. And the same goes for all philosophical claims, irrespective of whether they conflict with our intuitions or not. In the next two chapters I will explain in more detail why there is no special philosophical standard for justifying the counterintuitive, and why this fact constitutes strong evidence against DE.

How do we then explain the fact that philosophers seek numerous and diverse counterexamples against theories they attack? The answer is trivial: it is always better to have more rather than less reasons to reject something. Philosophers, like all human beings, are fallible, and their counterexamples might always be flawed. The chances that one counterexample is flawed are greater than the chances that many counterexamples are. But this has nothing to do with relying on intuitions as evidence of their contents.

7. The argument from intuitionism

Another objection to my critique of DE might look like this: there is a rich and venerable tradition of intuitionism, and many philosophers have labelled themselves "intuitionists". The very existence of the tradition, the objection goes, undermines my thesis: surely intuitionists must be providing some strong reasons to think relying on intuitions is central to philosophy, or they must be relying on intuitions themselves – or possibly both.

To address the objection, we need to first clarify what the doctrine of intuitionism is. In contemporary philosophy there are two major theories that bear the name: ethical intuitionism and intuitionism in the philosophy of mathematics – I will start with the former. Ethical intuitionism has two core components. First, there is a metaphysical thesis that moral properties are non-natural properties, and that whether something has a moral property is a matter of objective fact. Second, there is an epistemological thesis that we can know moral facts *by intuition*. Just like in science sense experience gives us access to objective, opinion-independent facts about the natural world, in ethics intuition gives us access to objective, opinion-independent facts about morality. There is no one shared understanding of "intuition" among intuitionists. Some of them presuppose the existence of a special faculty, some sort of sixth sense. Some argue that to know a proposition by intuition means to know it *a priori*, that is simply by understanding it, which does not require any extra cognitive faculties. Some believe knowledge by intuition must be immediate, or accompanied by a special phenomenology. Some are not very specific (see Stratton-Lake 2020). It seems that intuitionism as a whole does not offer solutions to most disagreements over the nature of intuition I discussed in the previous chapter.

Intuitionists typically call moral judgments knowable by intuition "self-evident", but few would argue that all true moral judgments have this status. For example, judgments like "The US should not have invaded Afghanistan in 2001" or "human embryonic stem cell research is morally acceptable in all circumstances" might be true, but are not self-evident, or at least not self-evident to most of us. This is because, first, in order to assess them one needs to be familiar with a multitude of non-moral, empirical facts, and these facts are hardly knowable by intuition. Secondly, the judgments seem *derived* from certain basic principles, and the act of deriving them does not have to be intuitive in any sense. To search for self-evident moral truths, we need to focus on something more abstract and more fundamental. For example, W. D. Ross argues it is self-evident that "if there are things that are intrinsically good, it is *prima facie* a duty to bring them into existence rather than not to do so, and to bring as much of them into existence as possible" (Ross 1930/2007, p. 24). It is also self-evident that things like virtue, knowledge and pleasure are intrinsically good (ibid., pp. 134-141). However Henry Sidgwick, another famous intuitionist, disagrees: according to him it there is only one intrinsically good thing, namely pleasure (Sidgwick 1874/1962, pp. 400-407). This is just one of many examples of how intuitionists' views about what intuition tells us are conflicted.

So much for what intuitionism says. We can now ask: is it irreconcilable with the rejection of DE? The answer is: it is plainly not. Intuitionism is a doctrine about the nature of moral reality and moral knowledge, and DE is a doctrine about philosophical practice. One can hold that philosophers rely

on intuitions, in the DE sense, and yet moral facts are natural, or relative to human opinion, or not knowable by intuition. This might imply that it is not very wise for philosophers to rely on intuitions, at least in ethics, but the position itself is clearly not self-contradictory. One can also hold that philosophers never rely on intuitions, in the DE sense, and yet intuitionism is true. This, again, might mean that philosophers should revise their methodology, but nothing more. Moreover, intuitions that proponents of DE usually talk about are very different from those discussed by the intuitionists. The former mostly include particular case intuitions, like trolley judgments, Gettier judgments etc., while the latter concern very abstract and general principles. For that reason, it is far from clear what conclusions about the adequacy of philosophical methodology one should draw on either of these views. But, in any case, the truth of intuitionism has no bearing whatsoever on DE.

It might be replied that while intuitionism itself may indeed be compatible with the denial of DE, it can lead to practices incompatible with it. Perhaps when intuitionists engage in normative or applied ethics, they try to put their metaethical views into practice, which results in treating certain propositions as supported by their own intuitiveness. How plausible is this view? There is no better way of testing it than analysing a concrete work. Here I will focus on a recent book by Michael Huemer, *Dialogues on ethical vegetarianism* (2019). I have selected it for three reasons. First, Huemer is a leading contemporary intuitionist. Second, the book is a straightforward example of an argument in practical ethics. Third, it is an accessible, introductory level text, which makes the argument fairly easy to follow. I am going to argue that Huemer does not rely on any intuitions in his book and suggest that reasons why he does not do so are also reasons why intuitionists in general never rely on intuitions – at least not in any DE-friendly sense.

The book has a form of a dialogue between two students: the meat-eating M and the vegan V, apparently Huemer's alter ego. Both use a lot intuition-talk. M frequently appeals to what he calls "intuitions", "self-evident claims", "brute facts", or "basic axioms", such as "it's morally okay to inflict severe pain on those who are much less intelligent, for the sake of small benefits to those who are more intelligent" (ibid., p. 6), "the threshold for having moral status is above the intelligence level of cows, chickens, or pigs" (p. 8), or "human interests are a *million* times more important than animal interests" (p. 12). In response, V offers reasons to reject these views. He points out it would not be permissible for Albert Einstein, or superintelligent aliens, to torture us for fun (p. 7). He argues that animal pain is bad just like human pain is bad, which means animals must have some kind of moral status (p. 9). He asks M to imagine two people, one smarter than the other, suffering from a headache of the same intensity, and argues that both have the same interest in getting rid of the headache (p. 13). Eventually M always gives in and abandons or modifies his

"intuitive" statement. Note that even though M is portrayed as someone desperately clinging on to whatever can help him rationalise his habit of eating meat, he nevertheless does not find DE an attractive method of defending his claims. First, he thinks it is perfectly legitimate to challenge them by offering reasons that are unrelated to facts about the claims' intuitiveness. Moreover, he does not even try to adopt a more moderate, "*prima facie*" variety of DE, according to which his claims should be evaluated on the basis of both whether they are intuitive and other considerations. For instance, he never says anything like "fine, V, you have your reasons, but I still have my intuitions, and my intuitions count for something – you now need to show me how your reasons outweigh my intuitions". Instead, all that matters to M are V's reasons.

Proponents of DE may reply that this is only because M's intuitions are not *genuine* intuitions. V points this out when he says that M is "saying what is convenient for [him], and declaring that to be intuitive" (p. 8) and that M's statements are "arbitrary at best, not intuitive at all" (p. 21). However in one case V admits that M's claim is "actually intuitive" (p. 47). The claim is that animal pain matters less because it is generally less conscious than human pain. Is it a suitable example to defend DE with? Arguably not. Surely whether the claim is true depends on all sorts of data concerning how consciousness can come in degrees, how human consciousness is of higher level than animal consciousness and how all this relates to the problem of experiencing pain. Whether one subscribes to intuitionism or not, it would be bizarre to try to support the claim with its own intuitiveness. V and M recognise this when they appeal to phenomena such as forgetting about one's pain while being immersed in an activity. On the other hand, they never try to appeal to the fact that the claim just seems true, or to the fact that it is spontaneous, or to the fact that it is nor consciously inferred from other claims.

At one point M seems to finally resort to DE:

> M: I don't know what's wrong with it, but the idea that animal agriculture is worse than the problem of war, or poverty, or disease, just *sounds* to me so extreme that it makes me want to say there must be something wrong with your argument.
> V: And you think that's enough to reject the argument?
> M: I do. I learned that from G. E. Moore: if you have an argument for a conclusion that seems crazy, you should reject it, even if you can't say exactly what's wrong with it. (p. 53)

Without further analysing the meaning of "sounds extreme" and "seems crazy", we can assume that using the fact that a proposition has this property as a reason against the proposition constitutes evidence for DE. However we must notice that V quickly proceeds to dismantle the argument. He

argues that in ethics just like in mathematics certain things can seem absurd and yet be true. In both cases we can understand where our intuitions come from and why they are misleading. M eventually accepts V's explanation and they move on to another topic. This hardly counts as engaging in DE on Huemer's part. If anything, Huemer argues that DE is a bankrupt methodology.

Finally, proponents of DE might argue that while V takes *some* intuitions to be unreliable, he must be appealing to other intuitions that he finds reliable. For example, he says that his argument "rests on intuitive, very widely shared moral beliefs, like "it's wrong to inflict a lot of suffering for no good reason" and "it's wrong to pay people for immoral behavior"" (p. 81). Elsewhere he specifies one of his premises as "suffering is bad" (p. 56). Surely at least some of these claims – the objection goes – must be supported with its own intuitiveness. But are they? The closest V gets to something that might resemble DE is the sentence I have just cited: he says that M should take meat eating to be wrong because V's argument rests on "intuitive beliefs". However when we place the sentence in a wider context, any DE-friendly interpretation falls apart. The point that V is making is that it is possible to know that buying meat is wrong without having to first decide which ethical theory is correct. One may subscribe to subjectivism, naturalism, dualism, consequentialism, deontology etc. – in each case V's conclusion follows.

If one wants to appeal to a broad audience with varied ethical and metaethical views, pointing out that one's starting premises are supported with the fact they are intuitive is not going to cut it – irrespective of what the premises are and what exactly one means by "intuitive". While ethicists tend to believe that DE is a widespread methodology, there is a lot of disagreement over whether it *should* be widespread, and whether there is a viable alternative (I discuss these views in detail in the next chapter). Huemer is perfectly aware of these disagreements. For example, in his *Ethical intuitionism* he writes that "the ethical naturalist does not recognize intuition as a legitimate source of knowledge" (2008, p. 230). This clearly does not favour the DE-friendly interpretation of Huemer's words. On the other hand, identifying what one's audience firmly believes and placing in the common ground is an obvious and straightforward way of overcoming metaethical differences. It seems much more plausible that this is what Huemer's "intuitive" refers to – especially given that it is immediately followed by "very widely shared". What Huemer is trying to say is that anybody who accepts, *no matter on what basis*, that suffering is bad, inflicting a lot of suffering for trivial reasons is wrong etc. should be vegan – not that claims like "suffering is bad" should be accepted because they are psychologically non-inferential, seem true in a distinct way etc. Once again, we find no evidence of DE in the text.

To this proponents of DE might reply: fair enough, what Huemer does in his book has little to do with DE. But this is only because he is addressing his argument to readers with all sorts of metaethical views, as well as to those whose metaethical views are not specified. He could, however, have said that his argument presupposed intuitionism and pointed any unconvinced readers to his elaborate defence of this position. This way at least some of the claims V calls "intuitions" would not be merely part of the common ground – there would be a presumption they are somehow supported by the fact they are intuitive. For example, in *Ethical intuitionism* Huemer lists "suffering is bad" as an example of a self-evident proposition, knowable by intuition (ibid., p. 231). The same proposition also appears to be one of the starting points of his case for veganism. Therefore, by introducing a simple modification, we could produce a neat example of using an intuition as evidence for its content. The scenario is of course counterfactual, but not unrealistic, which suggests that DE is a viable option at least in a certain type of philosophical literature.

There are, however, two problems with this response. First, Huemer may say that "suffering is bad" is self-evident, but he does not *argue* it is self-evident in his book. He puts forward a case for the thesis that there are *some* self-evident moral claims and gives some examples of such claims, but he does not explain why he chose these particular examples. Therefore someone who fully accepts Huemer's argument for intuitionism may still not have sufficient grounds to believe that "suffering is bad" is supported by "it is intuitive that suffering is bad". I will explore this problem in more detail later in this section, for now I will assume it can be solved. In this case, another problem arises: the starting point in Huemer's overall case for veganism would not be "it is intuitive that suffering is bad", but rather whatever he begins his metaethical argument with. For example, one of the starting points seems to be "hardly any of the vast vocabulary of an ordinary person is acquired through anyone's expressly *teaching* him the words, either by defining them or by giving him lists of examples" (ibid., p. 210) – which, of course, is not supported by the fact it seems true, but rather is placed in the common ground. This difference is important. Note that proponents of DE do not understand it as first arguing that certain intuitions under certain circumstances can be treated as evidence for their contents, and then treating them as evidence for their contents. DE is simply not a thesis about relying on intuitions in this intermediate way. Hence even if Huemer explained why some of his starting premises should be supported by the fact they are intuitive by pointing the reader to his metaethical basis, this would still not vindicate DE.

Proponents of DE might also try a slightly different thought experiment: Huemer could have addressed his *Dialogues* only to intuitionists, without endorsing any specific arguments for

intuitionism. This way "suffering is bad" would have been neither supported by further metaethical considerations, nor merely part of the common ground. There would be a consensus, tacit or explicit, that it is supported by "it is intuitive that suffering is bad". But this response does not work either. This is because it is still not clear that "suffering is bad", or whatever else Huemer offers as a self-evident truth, would be recognised as such by all intuitionists.

To see why, we must note that the claim is quite vague and can be interpreted in several different ways. One ambiguity is mentioned by Huemer himself: suffering can be *instrumentally* bad or *intrinsically* bad. The instrumental badness of suffering can be easily questioned – a dentist appointment is the usual counterexample. Painful treatment of a tooth can be good, however faced with a choice between painful and painless treatment, all else being equal, the patient would surely choose the latter, which shows that suffering is still intrinsically bad in this case. But is it *always* intrinsically bad? Imagine asking a masochist whether they would prefer to experience the pleasure they take from suffering pain, but without suffering any pain. The question does not seem meaningful. Another example of suffering that is not intrinsically bad might be patients anaesthesised with morphine who report still feeling pain, but being indifferent about it (Grahek 2007, p. 33).

"Suffering is bad" is also ambiguous in a different way. It can refer to the idea that suffering is bad *for the one who suffers*, or to the idea that "suffering is a bad thing, period, and not just for the sufferer" (Nagel 1986, p. 161). We can call it relative badness and absolute badness. What Huemer has in mind seems to be the self-evident *absolute* badness of suffering. Perhaps he could argue that the cases of masochism or morphine anaesthesia only show that suffering is not always bad *relatively*. However some philosophers have questioned the very possibility of absolute badness (and goodness) – one of them is Christine Korsgaard, whose argument I am going to discuss in chapter 4.

Finally, "suffering" in "suffering is bad" can also mean a number of things. For example, Korsgaard distinguishes between suffering as a sensation and suffering as a reflexive reaction (2018, pp. 160-1). Again, someone might argue that it is the badness of the latter that is self-evident, and all the counterexamples mentioned above concern the former. However it has clearly been challenged in both senses. For example, those who defend the view of well-being as self-realisation tend to think that self-realisation is impossible without suffering, understood as a reaction to things one is averse to – like owning up to one's mistakes (Clark 2021, p. 121). In short: regardless of which sense of "suffering is bad" we take, we are going to find philosophers arguing against it. And none of them seems to be weighing the fact that the claim is intuitive against their reasons to reject it. Rather, they also tend to ignore this fact in their arguments.

This shows that Huemer could not expect all his fellow intuitionists to take the badness of suffering as self-evident truth. If he believes it is something that can be known by intuition, he would need to argue that this is the case. And the same applies to any other putative self-evident principle. As there is no consensus over what counts as self-evident, there can be no tacit consensus over certain propositions being supported by their own intuitiveness, even within an exclusively intuitionist community. We can of course imagine an author who does not realise there is no tacit consensus of this kind, but this seems far-fetched. Academic philosophers typically interact with each other, their students, and other readers of their work on a regular basis. They tend to have a decent grasp of what their readers are likely to be committed to, and should they make any questionable assumptions about it, they are usually corrected by their colleagues and reviewers before their texts are published. We can also imagine explicitly addressing the argument to a very narrow intuitionist audience, one that would univocally assume that "suffering is bad" is supported by "it is intuitive that suffering is bad". But this sounds even more fanciful. It is normally expected of philosophers, especially in applied ethics, to reach as wide an audience as possible. The idea of a case for veganism that would only target members of a small, arbitrarily delineated subset of a particular metaethical school seems highly unpublishable.

I have argued that Huemer not only does not rely on any intuitions, in the DE sense, in his book on veganism, but also that it would probably not have been possible for him to publish the book had he tried to rely on intuitions in this sense. Someone might object that this is, after all, only one book: perhaps other intuitionists in other texts engage in something that resembles DE. However my considerations about DE not being a viable option even in principle can be generalised to all intuitionist publications. Regardless of what topic on takes up, the vagueness of intuitionism combined with the fact that few people endorse it makes DE highly implausible.
So far I have only discussed ethical intuitionism. How about intuitionism in the philosophy of mathematics? It is worth noting that besides the name the two have little in common. While ethical intuitionism is a form of realism, mathematical intuitionism is a form anti-realism: according to it, there is no mind-independent, extra-linguistic reality that mathematical statements could correspond to. Both maintain that intuition plays an important epistemic role, however they do not necessarily use the same notion of "intuition". L. E. J. Brouwer, the father of modern mathematical intuitionism, argues that its core is based on Kant's view about the relation between intuition, arithmetic and the representation of time. In Brouwer's own words, intuitionism "considers the falling apart of moments of life into qualitatively different parts, to be reunited only while remaining separated by time as the fundamental phenomenon of the human intellect, passing by

abstracting from its emotional content into the fundamental phenomenon of mathematical thinking, the intuition of the bare two-oneness" (Brouwer 1975, p. 85). Mathematical statements can only be be true or false due to this intuition: it is responsible for a subjective mental construction that constitutes a proof of a statement, and conditions under which a statement is proved determine its meaning.

There are some controversies over what exactly Kant meant by "intuition" (see Thompson 1972) and whether Brouwer interpreted Kant correctly. It is quite possible that we are dealing with a technical term that is not fully captured by what I discussed in the previous chapter. However further exegesis is unnecessary, as by this point it should be clear that just like ethical intuitionism, mathematical intuitionism does not imply DE. The former is a thesis about mathematical reality and mathematical knowledge, the latter is a thesis about philosophical methodology. What about the possible influence of intuitionism on philosophical practice? It is hard to imagine any intuitionism-inspired philosophical practice that would approximate DE. Intuitionism famously gave rise to a distinct logic that rejects the law of excluded middle (according to intuitionists, "*p* or not-*p*" is only true if either there is a proof of *p* or there is a proof that there is no proof of *p*, but this is not the case for a number of propositions). However even if we allow that practising intuitionist logic counts as practising philosophy, it does not follow that it involves publicly appealing to intuitions as evidence for their contents. The practice may *presuppose* that intuitions are evidence in this sense, but this is quite different from what I call tacit DE. Just like in the case of ethical intuitionism, this kind of presupposition would amount to treating intuitions as *intermediate* evidence: the intuitionist's background theory would include all considerations in favour of the claim that certain intuitions in certain circumstances support their contents – which is not the kind of evidence that DE refers to. This point, it seems to me, can be generalised to all possible consequences of embracing mathematical intuitionism: it might lead to relying to intuitions in some way, however it cannot lead to DE.

CHAPTER 3. The argument from counterintuitive conclusions

1. The argument

I have argued that so far in the literature two methods of testing DE have emerged: via the so-called method of cases and via intuition-talk. I also tried to show that while they provide strong evidence against DE, both have their limitations. My goal in the remaining part of the thesis is to overcome some of these limitations by offering a third way: what can be called "the argument from counterintuitive conclusions". The argument was first outlined by Molyneux several years ago (2014, pp. 454-457), but did not gain much traction since then.

It begins with a simple observation that philosophers dismiss intuitions on a regular basis. Think of Patricia and Paul Churchland arguing that beliefs do not exist, Galen Strawson arguing that everything is conscious, Keith Frankish arguing that nothing is conscious, Timothy Williamson arguing there is a sharp line between being thin and not being thin, Carl Hempel arguing that observing a red pencil confirms that all ravens are black, Willard van Orman Quine arguing that any statement, even "2+2=4", is subject to empirical revision, Paul Feyerabend arguing there is no scientific method, Donald Davidson arguing that animals cannot think, Daniel Dennett arguing that qualia do not exist, Peter Unger arguing that ordinary things, like rocks and chairs, do not exist, David Lewis arguing that every possible state of affairs is as real as the one we find ourselves in, Michael Dummett arguing that backward causation is possible, Harry Frankfurt arguing that moral responsibility does not depend on the ability to do otherwise, Jonathan Dancy arguing that there there is no role for moral principles to play in morality, Nancy Cartwright arguing that the fundamental laws of physics can explain a lot only because they are false, or Graham Priest arguing that a proposition and its negation can be true at the same time. Outlandish claims like that are nothing new – philosophy seems to have always been full of them. Think of Epicurus arguing that death cannot be harmful, Zeno arguing that motion is impossible, Pyrrho arguing that knowledge is impossible, Hobbes arguing that being subject to arbitrary will of a tyrant does not diminish one's liberty, Berkeley arguing that there is no matter, Kant arguing that it would be wrong to lie to the

murderer who asks about the whereabouts of his prospective victim, Hume arguing that we have no good reason to think the sun will rise tomorrow, D'Holbach arguing that free will does not exist, and many more.

If philosophers rely on intuitions as evidence, why are their claims so often so counterintuitive? There is a simple answer to this question: philosophers do not rely on intuitions as evidence, so they are free to dismiss them whenever they please. I believe this answer is roughly correct. However proponents of DE might object that I am attacking a straw man: few of them, if any, believe that intuitions can never be done away with in philosophy. DE, the reply goes, only commits one to the view that intuitions are generally treated as evidence, which does not mean they cannot be dismissed under certain circumstances. DE is therefore able to explain the fact that philosophers dismiss intuitions just fine.

One can think of an analogy between using intuitions in philosophy and using observations in science, often made by proponents of DE. It is somewhat naive to understand the scientific method as collecting empirical data and then coming up with theories that best account for it. As philosophers of science have long pointed out, not only scientific theories are influenced by data, but also data can be influenced by theories in a number of ways. It can also be far from obvious which theory best fits the data and why, or which data is relevant for which theory. It is not even clear what empirical data is. However few philosophers of science think that rejecting naive empiricism should lead us to reject the idea that observations are used as evidence in science. Rather, what we need is a more nuanced and sophisticated account of how observations are used as evidence.

Another analogy sometimes made by proponents of DE is one between using intuitions in philosophy and using intuitions in linguistics. Linguists, the story goes, typically appeal to what seems grammatically correct to native speakers of a language to determine what is in fact grammatically correct. Philosophers do the same with what seems to count as knowledge, justice, reference, causation, consciousness etc. However we must notice that even if this linguistics methodology picture is roughly accurate (for why it might not be, see Scholz 2021), linguists appear to occasionally dismiss people's intuitions about grammaticality. Consider the sentence "The car the man the dog bit drove crashed" (Fodor & Garrett 1967, p. 291). It does not seem correct to many native speakers of English, and yet linguists find it perfectly correct. The common explanation is that while people's grammaticality intuitions can and should be used as evidence, they cannot be taken at face value, as certain sentences may be difficult to process, speakers may be prone to performance errors etc. The job of a linguist is to filter out whatever distorts the intuition-data before using it to justify theories.

Similarly, we should not throw the baby out with the bathwater by abandoning the idea that philosophers rely on intuitions just because they sometimes dismiss intuitions. Rather, we need a more nuanced account of how both relying on intuitions and dismissing intuitions is possible at the same time. In this chapter I am going to describe eight such accounts. First, philosophers may dismiss intuitions that arise from a cognitive bias. Second, they may dismiss intuitions when they do not engage in conceptual analysis. Third, whenever it is necessary to reach reflective equilibrium. Fourth, whenever it is necessary to preserve theoretical virtues. Fifth, whenever it is required by the logic of the argument. Sixth, when intuitions do not concern abstract principles. Seventh, when intuitions are not based on expertise. Eighth, when the argument belongs to the intuition-free kind of philosophy. Each of these hypotheses can, alone or in combination with others, serve as a defence of DE against the challenge posed by the counterintuitive conclusions. In the next chapter I am going to test concrete examples of philosophical practice against these hypotheses to see if any of them has any merit.

2. Cognitive bias

Recall Climenhaga's argument from error theories discussed in the previous chapter. According to Climenhaga, intuitions incompatible with consequentialism are generally treated as evidence against consequentialism, in the DE-sense. However consequentialists may try to dismiss these intuitions by arguing they are in some sense biased. For example, they may try to explain away the appeal of the footbridge judgment ("it is impermissible to push the fat man off the footbridge to stop the tram") by arguing that we intuitively disapprove of any action involving pushing someone to their death, as pushing someone to their death *typically* produces worse overall consequences than the alternative. The association is a useful heuristic, however, as any heuristic, it goes haywire when applied to untypical situations – and the footbridge scenario is untypical, because it detaches pushing someone to their death from producing worse overall consequences.
Proponents of DE might be tempted to think that this is a common strategy: first biased intuitions are filtered out, either explicitly or implicitly, so that the remaining intuitions can be relied on as evidence. This way DE can be reconciled with the fact that philosophers often reach counterintuitive conclusions.

2.1 What is a bias?

The idea of wrong intuitions as products of fallible heuristics has gained a lot of popularity in recent decades, largely due to the work of Daniel Kahneman and Amos Tversky. Their empirical research of human decision making led them to the conclusion that there are two basic systems of reasoning: one that "operates automatically and quickly, with little or no effort and no sense of voluntary control", and one that "allocates attention to the effortful mental activities that demand it, including complex computations" (Kahneman 2013, p. 21). The former generates intuitive judgments by making use of certain rules of thumb. For example, we tend to be more concerned about losses than enthusiastic about comparable gains. When offered a 0.5 chance to loose $100 and 0.5 chance to win $150, most people reject the gamble, even though according to the standard expected utility theory of rationality, they should accept it (ibid., p. 238). Furthermore, eliminating the perception of a loss can make an offer more appealing. Compare "Would you accept a gamble that offers a 0.1 chance to win $95 and a 0.9 chance to lose $5?" to "Would you pay $5 to participate in a lottery that offers a 0.1 chance to win $100 and a 0.9 chance to win nothing?" – participants in a study were much more likely to respond positively to the latter question, apparently because they did not conceptualise buying a $5 lottery ticket as a loss (Kahneman & Tversky 1984, p. 349). Bias can therefore be understood in terms of *inconsistency*: sometimes we want to maximise our expected utility, sometimes we do not. The same offer sometimes seems acceptable to us, sometimes it does not. A heuristic is what makes us inconsistent.

2.2 Evolutionary explanations

Where do heuristics such as loss aversion come from? According to Kahneman, they are products of our evolutionary past, and we inherit them genetically: "organisms that treat threats as more urgent than opportunities have a better chance to survive and reproduce" (Kahneman 2013, p. 237). Joshua Greene, who has conducted extensive research on the psychology of trolley judgments, offers a similar explanation of why in the footbridge scenario sacrificing one person to save five seems wrong, unlike in the bystander scenario and other scenarios which do not involve killing *with one's bare hands*:

> "Up close and personal" violence has been around for a very long time, reaching far back into our primate lineage (Wrangham & Peterson, 1996). Given that personal violence is evolutionarily ancient, predating our recently evolved human capacities for complex abstract reasoning, it should come as no surprise if we have innate responses to personal violence that are powerful but rather primitive. That is, we might expect humans to have negative emotional responses to certain basic forms of interpersonal violence, where these responses evolved as a means of regulating the behavior of creatures who are capable of intentionally harming one

another, but whose survival depends on cooperation and individual restraint (Sober & Wilson, 1998; Trivers, 1971). In contrast, when a harm is *impersonal*, it should fail to trigger this alarmlike emotional response, allowing people to respond in a more "cognitive" way, perhaps employing a cost-benefit analysis. (Greene 2008, p. 43)

These explanations suggest that a heuristic is not necessarily a rule that produces correct answers in most cases. It might be a rule that *used to* produce correct answers in most cases at some point in the history of a species. Moreover, "correct" in this context does not mean "true", but rather "conducive to survival and reproduction". But this creates a problem for someone who wants to use the framework to filter out bad intuitions for philosophical purposes: philosophy is primarily about finding out which beliefs are true, not which beliefs help us reproduce. One can, of course, argue that there is a deal of overlap between the two. However to make this argument, one needs an independent criterion of truth, and the framework itself does not offer any such criterion.

Furthermore, there are reasons to be deeply sceptical about Kahneman's explanation of loss aversion, Greene's explanation of aversion to "personal violence", and similar adaptationist explanations of our intuitive judgments in general (for a review of these criticisms, see Downes 2021). Admittedly, one does not need to accept Kahneman's or Greene's evolutionary psychology to accept their basic heuristics framework. Perhaps heuristics are learned, or maybe we simply do not know how they came about. What we know is they still shape our intuitions, and some of these intuitions are incorrect.

2.3 The circularity problem

But which intuitions are they? Again, the framework is of no use without a criterion of intuition correctness. Granted, it is not impossible to adopt such criterion with some degree of plausibility. One can, for instance, argue that the expected utility theory is true: there are reasons to think it is the correct normative theory of rational choice (see Briggs 2019). As it follows from the theory that accepting the "loose $100 or win $150" 50/50 gamble is rational (at least under certain assumptions, like every $1 = 1 unit of utility), our intuition that we should not accept it must be flawed, along with numerous other intuitions shaped by the loss aversion heuristic. One might also want to adopt the principle of utility as the correct moral standard. As it follows from the principle (again, under certain assumptions) that we should push the fat man off the footbridge to stop the tram, our intuition that we should not do so must be flawed, along with some – but perhaps not too many –

intuitions shaped by the "pushing aversion" heuristic. This way we can clearly separate the wheat from the chaff.

However the question that arises here is: what would be the point of this exercise? If we know the criterion of rationality, morality, or whatever it might be, we already have what we are searching for. Why would anyone bother to test any intuitions against the criterion, if their goal is to find this very criterion? As I mentioned in the previous chapter, if Climenhaga is right about what he calls "error theories", then consequentialists defend consequentialism simply by assuming the truth of consequentialism: without this assumption they would be unable to tell whether the footbridge intuition is incorrect. And this point can be generalised to the whole enterprise of debunking intuitions for philosophical purposes – it amounts to putting forward circular arguments.

Here proponents of DE might bite the bullet and admit that when philosophers reject certain intuitions to rely on other intuitions, they are arguing in a circle. One philosopher who seems to hold this view is Robert Cummins:

> Philosophical intuition is epistemologically useless, since it can be calibrated only when it is not needed. Once we are in a position to identify artifacts and errors in intuition, philosophy no longer has any use of it. (Cummins 1998, p. 125)

Cummins believes that despite the manifest uselessness of intuition, philosophers heavily rely on it. It is possible – and advisable – for them to abandon their intuition-based methodology, however this would require a major revision of how philosophy is done. For example, it would mean eliminating all discussion of the trolley problem and other well-known thought experiments. I agree with Cummins that intuition is, in a sense, epistemologically useless, but I disagree that philosophers ever rely on it. In my view, he tries to fix a problem that does not exist. One reason to think that is simple: philosophers are not stupid. It is not very plausible that they routinely argue for their theories by showing they are in line with their favourite intuitions, and for their favourite intuitions by showing they are in line with their theories. Anyone who endorses this picture owes us an account of why philosophers are oblivious to this bizarre circularity of their argumentation – and, as far as I know, no such account has been proposed.

2.4 Two ways of debunking intuitions

There are more reasons to reject DE in the context of appealing to heuristics and biases, but they are somewhat less straightforward. The cognitive bias defence of DE is based on the idea that philosophers defend their views by accusing their opponents of succumbing to cognitive biases, which amounts to debunking their opponents' intuitions – and this would not make sense were DE to be false. However, as I argued in the previous chapter, even if DE is false, debunking intuitions is not out of place in a philosophical publication. It can serve as a device of persuasion, or it can be a rhetorically convenient way of introducing arguments. Identifying biases in one's opponents' thinking often goes hand in hand with identifying reasons behind their views, and problems with those reasons. This practice should be sharply distinguished from any DE-friendly interpretation of intuition debunking.

To shed some light on the criteria for deciding between the two interpretations, I will now illustrate each with an example. First, let us go back to another case brought up by Climenhaga: Parfit's defence of objectivism about practical reasons. Climenhaga believes – mistakenly, in my view – that the subjectivists attacked by Parfit must be relying on intuitions as evidence, and that Parfit recognises it as he tries to explain away their intuitions in order to undermine subjectivism. The relevant chapter starts off with the following question:

> Since so many people believe that *all* practical reasons are desire based, aim-based, or choice-based, how could it be true that, as objective theories claim, there are *no* such reasons? How could all these people be so mistaken? (Parfit 2011, p. 65)

Let us grant that these subjectivist beliefs can be characterised as intuitions, as Climenhaga understands the term. Parfit offers ten reasons why he thinks these intuitions are so prevalent. Here is one of them:

> Ninth, some people mistakenly believe that hedonic reasons are desire-based. When these people think about sensations that are painful or unpleasant, they do not distinguish between our dislike of these present sensations and our meta-hedonic desires not to be having sensations that we dislike. It is our dislike, I have claimed, that makes our conscious state bad, and gives us our reason to try to end our pain, or our unpleasant state. Since these people do not distinguish between our dislike and our meta-hedonic desire, they believe that this desire gives us this reason. Similar claims apply to pleasures, and to some other good or bad conscious states. (ibid., p. 67)

To be sure, Parfit is interested in the psychology of people he disagrees with. He explains that they are prone to confuse what actually grounds a practical reason with something similar that does not ground it. However, in order to claim that any confusion takes place at all, he needs to *argue* that it is one of those things, rather than the other, that grounds a reason. And he argues for this position at length – just not in the passage quoted above. According to one of his arguments, we all must have a reason to want to avoid all future agony, but if what makes our bad conscious states bad is a desire of some kind, then there are situations when we have no such reason (ibid., pp. 73-82). This is why Parfit believes that subjectivism is false – it is not because subjectivist intuitions are biased in some way. In other words, Parfit does not provide any evidence against subjectivism merely by analysing the psychology of his opponents. Even if their intuitions or biases were entirely different, Parfit's case for objectivism would remain the same. His argument confirms neither that he adopts DE, nor that he assumes anyone else does.

Let me now turn to the other example. As I have mentioned several times so far, my rejection of DE is not entirely unqualified – I think that experimental philosophy is probably the only kind of philosophy that relies on intuitions as evidence. To be more specific, the DE-friendly interpretation of the practice of debunking intuitions seems to describe large swathes of experimental philosophy quite accurately. In their "Experimental Philosophy Manifesto" Joshua Knobe and Stephen Nichols argue that one of the goals of the movement is to empirically study intuitive beliefs to "determine whether the psychological sources of the beliefs undercut the warrant for the beliefs." (Knobe & Nichols 2008, p. 7) Here is how Joshua Alexander and Jonathan Weinberg think this undercutting works in practice:

> We want [our sources of evidence] to be sensitive to all and only the right kinds of things; that is, whatever is relevant to the truth or falsity of the relevant set of claims. It turns out that at least some epistemic intuitions are sensitive to more than just these kinds of things; they are sensitive to aspects of who we are, what we are being asked to do, and how we are being asked to do it. There is a large range of well-motivated and prima facie–substantiated hypotheses about such sources of noise in various sorts of philosophical intuitions, far more than just ethnicity, gender, and order effects, including such demographic dimensions as personality (Feltz & Cokely, 2009) and such seemingly philosophically irrelevant differences as whether people are asked to imagine themselves thinking about the case "in a few days" versus "in a few years" (Weigel, 2011), or even what font the case is presented in (Weinberg, Alexander, Gonnerman, & Reuter, 2012). (Alexander & Weinberg 2014, p. 132)

As certain studies have revealed that people's intuitions about cases, like trolley judgments or Gettier judgments, are sensitive to epistemically irrelevant factors, we should reject, or at least become more sceptical, about these judgments. Gettier's case against the justified true belief theory of knowledge is thus weakened, and so are points made by Thomson in her series of papers about the trolley problem.

I think by this point it should be clear that there is a sharp difference between Parfit and experimental philosophers. They all engage in debunking intuitions and exposing biases, but for Parfit this practice does not *amount to* offering evidence for his philosophical position, or against the position of his opponents. In contrast, experimental philosophers think that when a judgement is shown to be influenced by a bias, it should be discarded as poor evidence. Some also suggest we should find out which intuitions are more "stable" – that is, resistant to biases – and use them as evidence instead (Wright 2014). All this means they not only think that armchair philosophy relies on intuitions as evidence – they also rely on intuitions as evidence themselves.

Therefore, to test the cognitive bias defence of DE as an explanation of why intuitions are dismissed in a given case, it is not enough to just check if cognitive biases are appealed to in the process. Rather, we need to check whether the dismissal itself plays an evidentiary role, like in experimental philosophers' work, or merely accompanies evidence, like in Parfit's work. DE can only be confirmed when certain intuitions are concluded to be false, or less likely to be true, *because of* being subject to a bias.

3. Conceptual analysis

Another way to restrict DE is to argue that philosophers only rely on intuitions as evidence when they engage in the practice called *conceptual analysis*. When philosophers do not engage in it, the response goes, they are free to dismiss intuitions. This is how counterintuitive conclusions in philosophy are possible. Cappelen suggests that Goldman and Pust in their "more cautious moments" subscribe to this kind of restricted view of relying on intuitions in philosophy (Cappelen 2012, p. 205). Another example might be the so-called Canberra Plan, which outlines two stages of philosophical enquiry. In step one, the philosopher is meant to find out how a concept is normally used by appealing to intuitions. In step two, she is meant to relate the concept to the actual world. This might involve consulting empirical results to determine "which of the available options is the

"best deserver" to be the phenomenon under discussion—which property or thing satisfies the most of the role specified, or the most important aspects of the role satisfied" (Nolan 2009, p. 269). The outcome of this procedure might be quite surprising and counterintuitive. For example, Paul Churchland's argument for eliminative materialism (Churchland 1981) could be interpreted as a confrontation of his analysis of the concepts of "belief", "desire", "intention" etc. with our best neuroscience, or perhaps with what we can expect neuroscience to be in the future, which results in the counterintuitive claim that there is no such thing as a belief, desire, intention etc.

3.1 What is conceptual analysis?

The main difficulty with the hypothesis is that it is far from clear to what extent philosophers actually engage in conceptual analysis. Views on that range from "philosophy, correctly conceived, simply *is* conceptual analysis" (McGinn 2012, p. 11), through "conceptual analysis is very widely practised—though not under the name of conceptual analysis" (Jackson 1998, p. vii) to the idea that philosophers never engage in conceptual analysis and those who say they do "misdescribe their own practice" (Papineau 2009, p. 4). The disagreement might stem from the fact that "conceptual analysis" can mean different things to different philosophers. It might also be explained by the fact that many philosophers are simply wrong about what they do. In any case, to make sense of the response we first need to explain what conceptual analysis is. Below I outline its basic characteristics, as it is typically described.

(1) It is a priori.

In order to do conceptual analysis, we do not need to appeal to any kind of sense experience. One can discover that all bachelors are men that have never married without empirically checking if this is the case, for example by finding as many bachelors as possible and enquiring about their marital status in the register office. What is possible to do with the concept of a bachelor is also possible (at least in principle) to do with concepts that philosophers are typically interested in, like knowledge, causality, reference, justice, consciousness, personhood, truth etc.

(2) Its product is an analytic statement.

One way to characterise the distinction between an analytic and a synthetic statement is to say that the former is "one whose truth depends upon the meanings of its constituent terms (and how they are combined) alone", while the latter is one "whose truth depends also upon the facts about the world that the sentence represents" (Rey 2023). This is sometimes referred to as the metaphysical conception of analyticity. On a slightly different account, known as the epistemological conception, one can *know* that an analytic statement is true merely by knowing what its constituent terms mean, without having to know anything about what they represent (Williamson 2022, pp. 54-5). The claim that bachelors are unmarried men would therefore be true in virtue of what terms like "bachelor", "man" or "marry" mean, or known to be true in virtue of knowing what those terms mean. On the other hand, the truth of a synthetic statement, like "Some bachelors are untidy", depends on the state of affairs that its constituent terms refer to. Conceptual analysis is then supposed to allow philosophers to discover truths about knowledge, causality, reference etc. just by pondering the meanings of these terms.

(3) If its product is true, it is *necessarily* true.

Most, if not all, true statements that are analytic and justified a priori are also supposed to be necessarily true, as opposed to contingently true. There are several ways of making sense of this distinction, the most popular being the possible worlds approach. The claim that bachelors are unmarried men is true in every possible world, while the claim that some bachelors are untidy is true in merely some possible worlds. If the product of conceptual analysis is true, it must be true in every possible world.

(4) It assumes the classical theory of concepts.

According to the classical theory of concepts "a lexical concept *C* has definitional structure in that it is composed of simpler concepts that express necessary and sufficient conditions for falling under *C"* (Margolis & Laurence 2023). To be a bachelor, it is necessary to never have been married and it is also necessary to be a man. Being both is sufficient. We can therefore explain what the concept of a bachelor is by listing all necessary and sufficient conditions for falling under the category of a

bachelor. And the same can be said of concepts of knowledge, causality etc. Alternative theories of concepts include the prototype theory, the theory theory, the atomistic theory and the eliminativist theory. All of them – in different ways – reject the idea that a concept is constituted by the necessary and sufficient conditions of its applicability.

(5) It is "mentalist".

Alvin Goldman and Joel Pust (1998) introduce the distinction between "mentalist" and "extra-mentalist" types of philosophical analysis. Their subjects are respectively an "in-the-head psychological entity" and "outside-the-head nonpsychological entity", the latter being a Platonic universal, a modal equivalence, or a natural kind.
Conceptual analysis, as the very name suggests, is meant to target concepts, which are clearly psychological entities. However the advocates of conceptual analysis often argue that its major strength lies in allowing us to discover truths about the external world. Moreover, the critics of conceptual analysis often argue its major weakness lies in failing to discover truths about the external world. Both groups then seem to assume conceptual analysis is ultimately not about concepts. So is it mentalist or extra-mentalist, after all? Perhaps the correct answer is it is mentalist as its *primary* target is always a concept. Optimists about conceptual analysis seem to think that there is a kind of correspondence between the concept of x and x itself that makes it possible to learn something about x by analysing the concept of x. Pessimists seem to think there is no such correspondence, however few deny that what is analysed is a psychological entity.

(6) It is descriptive.

The target of the orthodox conceptual analysis is a concept as it is, not as it *ought to be*. As Edouard Machery points out, there is also a kind of conceptual analysis that aims at reforming concepts rather than describing them – his examples are Carnapian explication and Gramscian analysis (Machery 2017, p. 312-20). Perhaps it would be more accurate to say that such reformist programs can be broken down into two components: the descriptive and prescriptive. But, in any case, this is not the mainstream approach. Frank Jackson insists that he is only interested in what x (be it free action, intentional state, etc.) is "*according to our ordinary conception,* or something suitably close

to our ordinary conception" (1998, p. 31). He is therefore not trying to modify or improve concepts, nor is he trying to suggest how the already existing concepts *should* be used.

(7) It relies on the "method of cases" and uses intuitions as evidence.

Jackson argues that what A. J. Ayer and Roderick Chisholm wrote about knowledge as a justified true belief "counted as a piece of conceptual analysis because it was intended to survive the method of possible cases. They sought to deliver an account of when various possible cases should be described as cases of knowledge that squared with our clear intuitions." (ibid., p. 28).
This is the standard account of how intuitions are used as evidence in philosophy discussed in detail in the previous chapter.

3.2 Analysing moral concepts

I have argued that some of the most celebrated examples of conceptual analysis, like Gettier's analysis of knowledge or Plato's analysis of justice, do not meet (7) – they do not involve appealing to intuition. It might be objected that even if I am right, Gettier or Plato might still be analysing concepts, perhaps on a less orthodox account of analysing concepts, which only satisfies *some* of the criteria listed above. I do not necessarily disagree with this objection. My goal in this thesis is to show that intuitions are not used as evidence, not that conceptual analysis in some form never occurs in philosophy. However it has to be pointed out that even if one adopts a weaker account, there would still be doubts about how widespread the practice is, or even whether it takes place at all. Consider the following passage on the role of the concept of justice in philosophical enquiry:

> Practical philosophy, as conceived by Kant and Rawls, is not a matter of finding knowledge to apply in practice. It is rather the use of reason to solve practical problems. The concepts of moral and political philosophy are the names of those problems, or more precisely of their solutions. This is made clear by the way Rawls employs the concept/conception distinction in *A Theory of Justice*. There, the *concept* of justice refers to the solution to a problem. The problem is what we might call the distribution problem: people join together in a cooperative scheme because it will be better for all of them, but they must decide how its benefits and burdens are to be distributed. A *conception* of justice is a principle that is proposed as a solution to the distribution problem, arrived at by reflecting on the nature of the problem itself. The concept *refers* to *whatever solves the problem*, the conception proposes a particular solution. (Korsgaard 2008, pp. 321-2)

If this reflects the nature of philosophical debates on justice in general, then – contrary to the common view – it is hard to see how they could count as examples of conceptual analysis. They are not aimed at capturing the ordinary meaning of "justice", they need not be divorced from sense experience in any sense, etc. Admittedly, Korsgaard does not argue that the concept of justice as the solution to the distribution problem must be the only concept of justice that is ever dealt with in ethics. Maybe philosophers other than Kant and Rawls focus on a different kind of concept, and they analyse it the way I have just outlined. But even this is problematic.

Recall once again the analysis of justice in Plato's *Republic*. Imagine that someone disagrees with Plato's claim that acting justly towards one's friend can never be harmful to them. Plato offers an interesting reply: when men are harmed "they become worse by human standards", and justice is one of such standards (*arete*). (Plato/Emlyn-Jones & Preddy 2013, p. 37, 335c). However justice cannot bring about injustice, just like heat cannot chill or dryness cannot moisten. If Plato is engaging in conceptual analysis, then his consideration should be knowable a priori. But it does not seem a priori at all. Rather, it appears that in order to define justice, Plato relies on an empirical hypothesis about what causes people to behave a certain way. Moreover, there is no indication that something that exists in the mind is the target on his analysis, or even that to understand the "outside-the-head" entity he must first understand the corresponding "inside-the-head" one. He is interested in justice itself, not in how we think about justice. In short, it is very hard to see how the passage is meant to meet the criteria I have listed. And yet it remains one of the canonical cases of conceptual analysis.

2.3 Quine and Williamson against analyticity

One might also object that the structure of non-moral concepts might be different. Perhaps knowledge, reference or causation are often analysed a priori, as psychological entities, in terms of necessary and sufficient conditions etc. However there are reasons to be wholesale sceptical about even the *possibility* of conceptual analysis. The most prominent one is arguably Quine's argument against analyticity (Quine 1951). According to Quine, what makes an analytic statement analytic must be the fact it is synonymous with a logical truth. For example, "Bachelors are unmarried men" is analytic because it is synonymous with "Unmarried men are unmarried men". But how do know whether an expression is a synonym of another expression? Quine argues we lack a satisfactory criterion – none of the available options can be understood without appealing to the notion of analyticity. We should therefore abandon the notions of analyticity and synonymy, together with the

hope for being able to draw a line between analysing the concept of x and consulting empirical data to argue that something counts or does not count as x. In fact, empirical data is needed for verifying *all* statements, which "face the tribunal of sense experience not individually but only as a corporate body" (ibid., p. 38).

Of course, not everyone has been convinced by Quine's argument. For example, Timothy Williamson argues that "although [Quine] may succeed in showing that "analytic" is caught in a circle with other semantic terms, such as "synonymous," he does not adequately motivate his jump from that point to the conclusion that the terms in the circle all lack scientific respectability, as opposed to the contrary conclusion that they all have it" (Williamson 2022, p. 52). However Williamson agrees with Quine's conclusion – he is only critical of the way Quine arrives at it. He concedes that we can intuitively classify many sentences as analytic and synthetic without much disagreement. But any analytic sentence can be rationally doubted by a competent speaker of a language who understands what its constituent terms mean. For example, we can imagine a native speaker of English who does not assent to "furze is gorse". She fails to realise that both terms refer to the same bush, perhaps because she learnt "furze" by ostention in summer and "gorse" the same way in winter, when the bush's appearance is very different. According to the standard account of analyticity, this situation must be explained by not understanding the ordinary meaning of "furze" and "gorse", not speaking English well, or being irrational. Williamson argues that none of these is the case. It turns out that convincing someone that furze is gorse does is not fundamentally different from convincing someone that furze has yellow flowers, or that furze grows in Scotland. In each case one needs to appeal to sense experience. It follows that the analytic vs synthetic distinction, just like the a priori vs a posteriori one, does not "cut at the cognitive or epistemological joints" (ibid., p. xxviii).

Assessing Quine's and Williamson's arguments against analyticity would go beyond the scope of this thesis. My point in this section is merely that it is far from clear what conceptual analysis is and whether it plays any role in philosophy. If conceptual truths do not exist – and there are serious reasons to think they do not – then the conceptual analysis hypothesis cannot be the answer to the problem of counterintuitive conclusions, unless the idea of conceptual analysis is radically reformulated. One such revisionist proposal has recently been put forward by Deutsch, who argues that conceptual analysis not only does not involve appealing to intuitions, but also that it does not target concepts (Deutsch 2020). The term has a referent, but it is a misnomer. In a sense, Deutsch saves conceptual analysis from Quine's and Williamson's attack. But he does it at the cost of

discarding DE. Proposals that reconcile DE with the idea that there are no conceptual truths seem much harder to come by.

2.4 The psychology of concepts

There is another doubt about the orthodox picture of conceptual analysis. The classical theory of concepts, which constitutes an important part of the picture, has been discarded by empirical science decades ago. As Cappelen points out, "psychologists disagree widely about just what concepts are—one thing they don't disagree about is the rejection of the view that concepts are represented as neat little bundles of necessary and sufficient conditions inside the speakers' heads." (Cappelen 2012, p. 209) If contemporary philosophers' goal is to find necessary and sufficient conditions of knowledge, justice, reference etc., then their enterprise seems not only futile, but somewhat anti-scientific. However the interaction between the philosophical community and the community of psycholinguists hardly resembles that of, for example, climatologists and global warming deniers. There is little hostility or isolation between the two. Philosophers frequently invoke psychological discoveries that shed light on their projects, and very rarely, if ever, dismiss the consensus in any field of psychology. It is of course possible that both sides are somehow oblivious to the conflict, but the more plausible explanation seems to be that philosophers are less committed to the classical theory of concepts than it is often claimed.
It seems that what is described as the search for necessary and sufficient conditions can in many cases be plausibly interpreted as, for example, the search for "family resemblance" type of characteristics, which are neither necessary nor sufficient. Proponents of DE might object that one still needs to rely on intuitions to find these. They might be correct, however the relation between the two has to be explained. And more generally, anyone who wishes to appeal to the idea of conceptual analysis to defend DE has to clarify which theory of concepts she adopts and how possessing a certain concept implies having certain intuitions on that theory (I am going to discuss this difficulty in more detail in chapter 5). Without such theory, it is impossible to make much sense of the conceptual analysis hypothesis.

From what I have said so far, it should be clear that trying to first find an instance of conceptual analysis and then check whether it involves relying on any intuitions would be unwieldy. Fortunately, this is unnecessary for my purposes. Since the hypothesis assumes that philosophers at some point of their enquiry rely on intuitions as evidence in the DE-sense, I can use my criteria,

developed in chapter 1, for testing DE in its both varieties, explicit and tacit. The criteria may not help us determine whether a concept is being analysed in this or that sense, but they will help us find the answer to the question whether intuitions are being appealed to in the sense it is typically claimed they are.

4. Reflective equilibrium

Perhaps it would be ideal to account for all intuitions, but, sadly, it is often impossible, as intuitions can come into conflict with each other. We therefore need to sacrifice some intuitions so that other intuitions can be preserved. Allegedly philosophers have a well-established method of doing exactly that: what is called "the method of reflective equilibrium". Here is how Norman Daniels outlines the idea:

> The method of reflective equilibrium consists in working back and forth among our considered judgments (some say our "intuitions," though Rawls (1971), the namer of the method, avoided the term "intuitions" in this context) about particular instances or cases, the principles or rules that we believe govern them, and the theoretical considerations that we believe bear on accepting these considered judgments, principles, or rules, revising any of these elements wherever necessary in order to achieve an acceptable coherence among them. (Daniels 2020)

This may seem perfectly in line with DE: first certain judgments are taken to be supported, explicitly or implicitly, by the fact they are intuitive, and then philosophers work back and forth among them until some sort of equilibrium state is reached. Perhaps what I described as the justification interpretation – that is offering reasons for and against intuitive judgments – is part of the method, and should not be held against DE.

4.1 What is reflective equilibrium?

How do we determine whether philosophers actually engage in reflective equilibrium seeking? The question is not easy to answer. "Reflective equilibrium" may sound like a well-defined philosophical term of art, but in fact different philosophers interpret it differently, and in many cases it is hard to tell how the method is understood. I am now going to present a number of problems with defining the idea.

First, is the method coherentist or foundationalist? The dominant view is that it is the former: no proposition is immune from revision in the process of seeking coherence. However Pust argues that on any reasonable interpretation we are dealing with a method "within which the process of justification is linear and stops with intuitions" (Pust 2000, p. 13). For him, some claims must be treated as non-negotiable, otherwise the method would suffer from vicious circularity, and would not be worth a serious consideration.

Secondly, for a judgment to be treated as an input, is being intuitive a necessary or a sufficient condition, or perhaps neither? Rawls makes it clear that intuitiveness is merely one of a number of properties required: considered judgments also need to remain uninfluenced by certain emotional states, self-interest or threats, they need to be made by people with a degree of intelligence, certain kind of understanding of how human interests can conflict etc. (Rawls 1951, pp. 178-183, Rawls 1971, p. 47) However all these additional requirements are often neglected in contemporary discussions of the method. The status of principles, rules and theoretical considerations mentioned by Daniels is also far from clear: are they all suitable to be used as starting points because they also are, in some sense, intuitive? To make things even more confusing, some philosophers argue that reflective equilibrium can, but does not have to involve working with intuitions (Brun 2014), and others argue it never involves working with intuitions at all (Bealer 1998, p. 206).

Third, what do we mean by "intuitive"? As I mentioned in the previous chapter, Rawls's understanding of the term is radically different from that of most contemporary theorists: for him intuitions are judgments that are not derived by consciously applying ethical principles. Should we adopt this account, or should we go with one of the modern ones? If so, which one?

Fourth, if aside from being intuitive judgments need to meet other conditions to be considered, what exactly are these conditions? Stefan Sencerz outlines four broad answers to this question. Judgments may be need to be "made under relevant cognitive conditions", such as calmness or familiarity with relevant facts. They may need to be "formally correct", for example made with a measure of conceptual clarity, impartiality or rationality. They may need to "result from cognition of objectively existing moral properties", and, finally, they may need to simply be made with confidence (Sencerz 1983, pp. 83-90). There seems to be little consensus over which of these criteria, if any, should be applied.

Fifth, does "seeking equilibrium" refer to a creative process that ultimately leads to the publication of a philosophical work, or perhaps to the work's content? As I have argued in chapter 1, the same publication can be the outcome of many different thinking processes. The reverse is also possible:

the same thinking process can be captured by many different publications. This means that much depends on whether we are talking about philosophers' psychology or philosophers' work – or perhaps some combination of both.

Sixth, is there a separate stage of inquiry when a set of starting points is determined and another one when the process of confronting them with each other takes place? Or perhaps the two are intertwined? Should we expect reflective equilibrium seekers to provide us with a list of input-intuitions at the beginning of a text? Or should we expect them to introduce them gradually as they go along?

Seventh, is the process carried out by an individual, or is it dialogical? In other words, is it even possible to focus on a single philosophical publication and determine whether it utilises the method of reflective equilibrium? Or maybe the presence of the method can only be confirmed by examining the responses to the publication written by other authors, then responses to these responses etc.? Daniels claims that both uses are possible, however the common understanding seems to be the individualist one.

Eighth, are we talking about the so-called narrow or wide reflective equilibrium? According to Rawls, the latter does not involve investigating "principles people would acknowledge and accept the consequences of when they have had an opportunity to consider other plausible conceptions and to assess their supporting grounds" (Rawls 1974, p. 8) – which is essentially what Daniels refers to as "theoretical considerations" in the passage quoted above, and also what can be to an extent identified with theoretical virtues I am going to discuss in more detail in the next section. The idea behind the distinction is that for a set of judgments several different narrow equilibria can be be found, and the method of seeking wide equilibrium would allow us to eliminate some of them. But it is not entirely clear whether we should understand seeking wide equilibrium as a further stage of enquiry, possible only after some narrow equilibria have been found, or rather as an extended version of seeking narrow equilibrium in which some additional claims are treated as input.

Ninth, are we talking about a method of ethics, or a method of philosophy in general? Rawls thinks it is the former, however something closely resembling Rawls's idea can be found in Nelson Goodman's work on the justification of deductive and inductive inferences:

The point is that rules and particular inferences alike are justified by being brought into agreement with each other. *A rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend.* The process of justification is the delicate one of making mutual

adjustments between rules and accepted inferences; and in the agreement achieved lies the only justification needed for either. (Goodman 1955, p. 67)

This suggests that seeking equilibrium is successfully carried out outside ethics. Daniels credits Goodman as one of the originators of the method, but he points out that since 1950s the discussion has been almost exclusively focused on its application in ethics and political philosophy. It is not clear what explains this fact. Is it just a historical contingency, or is there something about ethics that makes the method more suitable for this particular area?

Tenth, how exactly are we supposed to solve conflicts between judgments to reach the equilibrium state? Are there any rules for deciding which judgments have to go? If so, what are they? Some, like Pust and Bealer, argue that intuitions about particular cases are generally better evidence than intuitions about abstract principles (Pust 2000, p. 12, Bealer 1998, p. 205), which suggests the latter should give way to the former. On the contrary, Jeff McMahan writes that the method in its standard form "assigns the same epistemic status to our intuitions about particular cases that it assigns to the deeper principles of which the intuitions are expressions" (McMahan 2013, p. 113). The same applies to possible conflicts between intuitions about cases and principles on the one hand, and theoretical considerations taken into account in seeking a wide reflective equilibrium on the other. Should we assign equal status to both, or is one set weightier than the other?

Finally, to answer all of the above, what are the canonical texts that we should we turn to? And do these texts all refer to one phenomenon, or perhaps to several different ones? As Daniels points out, Rawls is widely considered to be the father of the method. But which of Rawls's publications are the most relevant? He first uses the term "reflective equilibrium" in *The theory of justice* (1971/1999), and then in *The independence of moral theory* (1974). However what he says earlier in *The outline of a decision procedure for ethics* (1951) is often taken to describe roughly the same method. A similar problem applies to the distinction between wide and narrow reflective equilibrium – the terminology is first found in *The independence*, however what Rawls says in *The theory of justice* seems to refer to the same idea. Another example of what appears to be reflective equilibrium *avant la lettre* is the above mentioned work of Goodman. This raises the question: does the fact that the term "reflective equilibrium" or "wide reflective equilibrium" is not used in a text makes it less authoritative? Again, no obvious answer can be given.

Pust believes there are five versions of the method of reflective equilibrium: three different ones proposed by Rawls, one proposed by Goodman, and one proposed by Daniels (Pust 2000, p. 13). Daniels himself believes that essentially there are only two: the method of seeking narrow

equilibrium and the method of seeking wide equilibrium. And many of who those use the term seem to think there is only one. Who is right? Everything depends on which description of the method one relies on and how one interprets it.

4.2 Ambiguous input

Given all this confusion, how is the reflective equilibrium hypothesis to be tested? It seems that we would need to either develop testing criteria for each of the numerous varieties of the method, or to reject some of the varieties as implausible, and focus on evaluating the remaining ones. However for my purposes this would be unnecessary. Just like with the conceptual analysis hypothesis, I am happy to concede that philosophers seek reflective equilibrium – that is, they work back and forth between judgments, some of which, or all of which, are intuitive in some sense. What I deny is that these judgments are meant to be supported by the fact they are intuitive. In other words, I deny that philosophers seek reflective equilibrium *in a DE-friendly way*.

To test the DE-friendly variety of the hypothesis, we only need to employ the criteria for testing DE in its open and tacit forms outlined in chapter 1. However it might be objected that while testing for explicit DE seems straightforward, deciding between tacit DE and the DE-friendly reflective equilibrium might be impossible. I argued that to test the tacit variety of DE one needs to decide whether a community of philosophers would be likely to evaluate a suspected proposition by appealing to its intuitiveness, or rather by appealing to other considerations. The latter would weigh against DE. But if the DE-friendly reflective equilibrium hypothesis is true, then other considerations would always be brought up to evaluate the proposition – after all, this is the very point of the method of reflective equilibrium. This way the reflective equilibrium hypothesis could be used to turn both tacit DE and the denial of DE into something of an unverifiable article of faith.

To illustrate, suppose that the proponent of DE argues that the bystander judgment ("it is permissible to throw the switch and divert the tram onto a sidetrack") is treated as tacitly supported by its own intuitiveness, and as part of an input in the process of reflective equilibrium seeking. Eventually, the judgment is dismissed, as it can be seen in Thomson's 2008 paper, because it is impossible to reconcile it with other judgments treated as tacitly supported by their own intuitiveness, like "negative duties are more significant than positive duties". In contrast, my position is that neither Thomson nor others who have discussed this problem take the bystander judgment to be supported by its own intuitiveness, as they are all eager to examine various reasons for and against the judgment, like "negative duties are more significant than positive duties", and

none of those reasons has anything to do with the fact that the judgment is intuitive. However the proponent of DE may not be moved by my evidence against DE. For her, the presence of intuition-unrelated considerations can be just as well explained under DE: different propositional contents attested by their own intuitiveness are being confronted with each other to reach the equilibrium state. Moreover, the absence of intuition-related considerations can also be explained under DE: they are simply too obvious to make them explicit.

This might look like an impasse – two contradictory accounts of philosophical methodology have been proposed and there is no way of deciding which one is correct. But the situation is not as hopeless as it may seem. This is because what according to the DE-friendly picture is too obvious to be stated can be sometimes called into question, and philosophers' reaction to this kind of challenge can tell us a lot about the plausibility of DE. Suppose someone objects that they do not find the bystander judgment intuitive at all. If DE is true, then we should expect the objection to prompt a certain kind of response from those who discuss the example. As it is far from clear what kind of consensus over intuitiveness is required by DE, the response could take different forms. Someone might argue that *enough* people find the judgment intuitive, or that *the right kind of people,* perhaps people with philosophical expertise, find it intuitive. Someone might point to the ambiguity of "intuitive" and argue that the judgment is still intuitive in the relevant sense. Someone might back down and withdraw their argument. In any case, the tacit, DE-based assumptions would likely come to the surface. If, on the other hand, philosophers would struggle to see how the intuitiveness of the bystander judgment is relevant to what they do, we would gain evidence against DE.

4.3 Input and demographic variety

The example I have just given is not purely hypothetical. As I mentioned, one of the goals of the experimental philosophy movement is to systematically examine how judgments about famous philosophical cases vary across different demographic groups, or how they are influenced by different factors. In several studies it has been revealed that Americans and Europeans are significantly more likely to agree with the bystander judgment than the Chinese – and that there is still a fair number of Americans and Europeans who reject it (Awad et al. 2020, Ahlenius & Tännsjö 2012). Given that participants were asked for spontaneous responses, the responses would count as intuitions on most accounts of the intuitive. On the DE-friendly interpretation of the reflective equilibrium hypothesis, these results should spark a lively debate among philosophers writing about the trolley problem. Is it acceptable to treat the bystander judgment as an input in the process of

reflective equilibrium seeking, given that it does not seem true to many people? Should the Chinese have a separate trolleyology, one that is more in line with Chinese intuitions? How intuitive are other propositions that have been used as an input in the debate? If DE is true, we would expect philosophers to quickly become preoccupied with questions of this kind.

Furthermore, even if empirical studies on intuitiveness did not exist, philosophers would probably think of carrying out such studies themselves. In other words, there would be a tendency to transform philosophy into experimental philosophy. Note that if DE is true, it is hard to imagine how many philosophical debates could go on for many years without someone questioning the intuitiveness of various argumentative starting points. For instance, it is far from obvious whether the bystander judgment seems true to people in all demographic groups, or that it cannot be made less appealing by various means, like reframing the scenario. Even if there existed a tacit agreement over the bystander judgment being supported by its own intuitiveness, this agreement could not remain universal for very long.

My view, on the other hand, implies that studies carried out by experimental philosophers are largely irrelevant, argument-wise. Of course, if someone wishes to place the bystander judgment in the common ground, they might be interested in how many of their potential readers accept the bystander judgment, and perhaps modify the common ground accordingly. However nobody treats the intuitiveness of this or any other judgment they place in the common ground as evidence for the content of the judgment. This means that learning what seems true, to whom it seems true, under which conditions it seems true etc. is not of primary interest to philosophers, as it does not affect the substance of their arguments. We can therefore test the DE-friendly version of the reflective equilibrium hypothesis by looking into how philosophers react to experimental philosophy and whether they are inclined to engage in experimental philosophy themselves. This provides us with an additional criterion to what I offered in chapter 1.

5. Theoretical virtues

Proponents of DE generally believe that philosophical theories are meant to accommodate intuitions: if intuitions do not fit a theory, so much the worse for the theory. However in some cases they might be inclined to agree with the opposite: if intuitions do not fit a theory, so much the worse

for the intuitions. This is because what we expect from a good theory is not *only* being able to accommodate as many intuitions as possible, but also to have certain other features. And as sometimes having these desirable features is not achievable without dismissing intuitions, some intuitions have to go. An example of such situation is given by Williamson:

> Some revisionary metaphysicians deny that, strictly and literally, there are mountains. (…) The claim that there are no mountains is usually regarded as counterintuitive. Even its proponents may concede that it is counterintuitive, arguing that the cost to intuition is worth paying for the overall gain in simplicity, strength, logical coherence, and consonance with science they attribute to their total metaphysical system, which entails the claim. If their system also entails that there could not have been mountains, it contradicts the modal "intuition" that there could have been mountains. But even without the claim of necessity, the non-modal claim that there are no mountains is already counterintuitive as many philosophers use the term, because it contradicts the common sense judgment that there are mountains, for example in Switzerland. (Williamson 2022, pp. 220-1)

Williamson is suggesting that a theory is *ceteris paribus* considered better when it entails an intuitive claim like "mountains exist". One can hold this view without committing oneself to DE: for example, one can assume that claims like "mountains exist" are part of the common ground for a community of enquirers. But it is also a view that proponents of DE find attractive, if not indispensable: according to them, the intuitiveness of "mountains exist" lends some support to the claim that mountains exist, and this is why a successful theory should generally explain why mountains exist, unless, some might add, the cost of explaining it is too high. And what makes it too high is lack of certain theoretical virtues. In chapter 1 I briefly discussed what is often called "explanatory desiderata" or "explanatory virtues" – these seem largely synonymous with theoretical virtues, although it can also be argued that the latter category is broader, as theories not born out of abductive reasoning are also expected to have them.

5.1 The list

It is notoriously difficult to define theoretical or explanatory virtues. Lycan points out that while virtually everyone agrees that simplicity should make the list, there are multiple ways of understanding it: it can refer to linearity of mathematical function, "elegance of structure; parsimony of posits and/ or of ontology; fewer principles taken as primitive; and no doubt more" (Lycan 2002, p. 415). Worse still, none of these characteristics is easily measured. The same seems to apply to four other virtues distinguished by Lycan:

Testability.
Other things being equal, a hypothesis H will be preferred to a competitor H' if H has more readily testable implications. (...) Intuitively, if a hypothesis makes no testable predictions, it has little explanatory force. (...)
Fecundity.
H will be preferred to H' if H is more fruitful in suggesting further related hypotheses, or parallel hypotheses in other areas. (Perhaps this is a higher-order form of simplicity again.)
Neatness.
H will be preferred to H' if H leaves fewer messy unanswered questions behind, and especially if H does not itself raise such questions.
Conservativeness.
H will be preferred to H' if H fits better with what we already believe. If this sounds dogmatic or pigheaded, notice again that, inescapably, we never even consider competing hypotheses that would strike us as grossly implausible (...) All inquiry is conducted against a background of existing beliefs, and we have no choice but to rely on some of them while modifying or abandoning others—else how could any such revisions be motivated? (ibid., pp. 415-6)

It is worth noting that Lycan is focusing on virtues of scientific theories, which might be somewhat different from those of philosophical ones. Most obviously, the virtue of being consonant with science listed by Williamson cannot apply to something that is part of science. A more controversial case would be testability – many would be inclined to think that philosophical theories are not supposed to be testable the way that scientific theories are, or perhaps even not testable at all.

5.2 Commensurability

Another problem with making sense of the hypothesis has to do with the commensurability of the virtues, or lack thereof. According to Lycan, theoretical virtues often come into conflict with each other – for example, revolutionary scientific theories tend to score high on simplicity, but low on conservativeness – and there is no obvious standard of comparison. This suggests that in many situations it might not be easy to determine whether something is treated as a theoretical virtue outweighed by another theoretical virtue, or as a non-virtue. A non-conservative theory is likely to be presented in a way that downplays conservativeness, regardless of whether its author sees conservativeness as something desirable.

Similarly, if theoretical virtues are ever used to outweigh the intuitiveness of judgments, in a DE-friendly sense, it is far from clear would count as evidence for such outweighing. Suppose what Williamson says about metaphysicians who argue that mountains do not exist is fully accurate –

their rejection of the intuition that there are mountains is treated as a price worth paying for gains in simplicity, coherence etc. of their theories. As I mentioned, this can mean the metaphysicians are relying on the intuition that mountains exist as defeasible evidence for its content. But it can also mean according to them the typical common ground in the debate, which includes the claim that mountains exist, needs to be reconsidered. The latter option does not confirm DE in any sense.

5.3 Testing the hypothesis

How can we decide which one is correct? The theoretical virtues defence of DE seems to come in two basic varieties. According to one, it is only the intuitiveness of what is ultimately rejected that is treated as evidence. According to the other, some intuitions withstand the process of argumentation, in a DE-friendly way. Perhaps the intuitiveness of the theoretical virtues themselves is used as evidence – for example, that fact that it seems to us that the correct theory should be simple is used as evidence that it should be simple. One way of testing the latter variety would be to identify the prevailing intuitions and checking whether any of them has been used as evidence, according to the criteria I laid out in chapter 1.

Additionally, both varieties can be tested by focusing on intuitions that end up dismissed. For example, finding assertions like “it seems to us that mountains exist, which gives us a reason to believe they exist” or “our intuition that mountains exist indicates that they exist” would constitute straightforward evidence in favour of the DE-friendly interpretation of the hypothesis. However, faced with a lack of explicit assertions like this, proponents of DE might want to defend its tacit variety, according to which claims like “our intuition that mountains exist indicates that they exist” are universally recognised assumptions that need not to be stated.

It might be objected that my criteria for testing the tacit variety of DE are not applicable, for reasons described in the previous section. This is because the theoretical virtues hypothesis can be interpreted as an instance of the reflective equilibrium hypothesis. But if this is the case, the additional criteria specified in the previous section could be employed: we could check to what extent and in what ways philosophers are interested in investigating the psychology of relevant intuitions. If, for instance, they tend to ask questions about how strongly or why it seems to us that mountains exist, whether there are individuals or cultures that do not take the existence of mountains for granted etc., it would give us a reason to think that at least some of them might be relying on the intuitiveness of this proposition as evidence.

6. Arbitrariness

In *The Philosophy of Logical Atomism* Bertrand Russell writes:

> I am trying as far as possible again this time, as I did last time, to start with perfectly plain truisms. My desire and wish is that the things I start with should be so obvious that you wonder why I spend my time stating them. This is what I aim at, because the point of philosophy is to start with something so simple as not to seem worth stating, and to end with something so paradoxical that no one will believe it. (Russell 1914/2009, p. 20)

It is not entirely clear whether Russell's intention here was to express a genuine metaphilosophical view or perhaps just to quip without making too much of a commitment. However if we take his claim seriously and if we also assume we can identify "truisms", "obvious" and "simple" with the intuitive, and "paradoxical" with the counterintuitive, Russell's claim can inspire another answer to my research question: philosophers are only interested in using intuitions as starting premises of their arguments, and they do not care if they dismiss intuitions in their conclusions. Or perhaps it is even desirable for them to do so.

Someone might point out this is a rather bleak picture of philosophy, at least so long as one expects philosophy to be about finding true answers to philosophical questions. Suppose we have three intuitive propositions: p, q and r. If we can show that not-r follows from p and q, then we can also show that not-p follows from q and r, and that not-q follows from p and r ($[(p \wedge q) \rightarrow \sim r] \equiv [(p \wedge r) \rightarrow \sim q] \equiv [(q \wedge r) \rightarrow \sim p]$). If the process of selecting initial intuitions is *arbitrary*, then we are bound to be left with a host of contradictory claims and no way of telling which one is correct. Everything depends on what a philosopher happens to start with.

The fact that some view is bleak does not of course mean it is false. Perhaps philosophers do not care about the truth of their premises and their conclusions after all – or, somewhat less plausibly, they fail to realise that their methodology makes it impossible to establish what is true. One might try to argue that at least some examples of philosophical philosophical practice fit into this picture.

6.1 Sorites and its starting points

Take arguably the most popular example of an inconsistent set of intuitive propositions in philosophy, the Sorites paradox. It can be portrayed as a conflict between three statements:

(1) On grain does not make a heap.
(2) One grain does not make a difference between a heap and a non-heap.
(3) One million grains makes a heap.

All three seem undeniably true, and yet they cannot all be true at the same time – at least not unless standard logic is rejected. However the laws of standard logic seem undeniably true too, so, in any case, some seemingly true statements have to be denied in order to solve the paradox.
And when we look into philosophical solutions of the paradox, this seems to be the case. For example, Timothy Williamson accepts (1) and (3), but denies (2). According to him everything is either a heap or a non-heap as there is a sharp cut-off point between the two, we just do not know – and cannot know – where it lies (Williamson 1998). On the other hand, Peter Unger accepts (1) and (2), but denies (3). According to Unger heaps do not exist, and neither do "pieces of furniture, rocks and stones, planets and ordinary stars, and even lakes and mountains" (Unger 1979, p. 119). This is because if they existed, we would be forced to admit that one atom, or even no atoms at all, can constitute a stone, a planet etc. which, argues Unger, is "absurd". Perhaps Williamson arbitrarily picks (1) and (3) to attack (2), and Unger arbitrarily picks (1) and (2) to attack (3). Perhaps all that matters, form a methodological point of view, is that the initial premises are *intuitive.* If the premises then lead to counterintuitive conclusions – so be it.

6.2 Truth, intuitiveness, and indifference

The arbitrariness hypothesis comes in several varieties. We may distinguish between one according to which philosophers *deliberately* try to arrive at counterintuitive conclusions and one according to which philosophers *are indifferent* in this respect. The hypothesis can be also divided into two other, cross-cutting categories. According to one, philosophers do not care about the truth of their starting premises, all they care about is whether they are intuitive. According to the other, philosophers care about the truth of their starting premises – and consequently, the truth of their conclusions – however they are somehow oblivious to the fact that intuitions can contradict each other and, consequently, to the fact that simply relying on any intuitions is not likely to get them very far in their pursuit of philosophical truth.

Fortunately all these varieties require that starting points are meant to be supported by their own intuitiveness, which means my criteria for testing DE in both forms are applicable.Moreover, it may be useful to confront a particular counterintuitive conclusion with conflicting counterintuitive

conclusions and see how their advocates interact with one another. For example, if what I wrote about Williamson and Unger is correct, we should expect both of them to see each other's starting premises as equally legitimate, so long as both agree they are intuitive. They also should not be in principle opposed to the idea of accepting each other's conclusions as equally legitimate. The arbitrariness hypothesis, if true, should considerably limit the ways in which philosophers defending different views can disagree with each other. If, on the other hand, it turns out that philosophical disagreement goes beyond those limitations or that philosophers do not start their arguments with anything that simply happens to be intuitive – we will have grounds to reject the hypothesis.

7. Principles only

Peter Singer argues that our ethical intuitions, like the footbridge judgment or the bystander judgment, cannot be trusted. This is because they are tainted by their evolutionary origin, in the way I described in section 2. Instead of appealing to intuitions philosophers should appeal to *reason*. In evaluating different trolley cases one should carefully distinguish morally relevant factors, like the number of lives lost in each scenario, from morally irrelevant ones, like the means of killing one person to save the five. However some object that what Singer proposes is essentially trading one intuition for another:

> It might be said that the response that I have called "more reasoned" is still based on an intuition, for example the intuition that five deaths are worse than one, or more fundamentally, the intuition that it is a bad thing if a person is killed. (...) The "intuition" that tells us that the death of one person is a lesser tragedy than the death of five is not like the intuitions that tell us we may throw the switch, but not push the stranger off the footbridge. It may be closer to the truth to say that it is a rational intuition, something like the three "ethical axioms" or "intuitive propositions of real clearness and certainty" to which Henry Sidgwick appeals in his defense of utilitarianism in The Methods of Ethics. The third of these axioms is "the good of any one individual is of no more importance, from the point of view (if I may say so) of the Universe, than the good of any other." (Singer 2005, pp. 350-1)

As I argued in chapter 1, intuitionists like Sidgwick do not rely on intuitions in any DE-friendly sense. Neither does, in my view, Singer. However his words can be interpreted as an endorsement of a form of DE: only intuitions about general or abstract principles are used as evidence of their

contents. This hypothesis could neatly explain how counterintuitive conclusions are possible: it is perfectly legitimate to dismiss intuitions about particular cases.

To test the hypothesis, we need to distinguish intuitions about particular cases from intuitions about principles. This may not be as straightforward as it seems. Many of our particular cases are not, in a sense, particular at all: they are not descriptions of singular, concrete, spatio-temporal events. An infinite number of such events can satisfies the description of the footbridge scenario, and the footbridge judgment applies to each of them. In this respect, the claim that it is wrong to push the fat man off the bridge is no different from the claim that it is a bad thing if a person is killed. However the former seems to be a principle, while the latter does not.

7.1 Principles and frequency

Perhaps the difference boils down to the *frequency* with which the relevant events occur in the real world: we often come across cases of people being killed, while we hardly ever come across cases of people being pushed off bridges to stop trams from hitting other people. But this account has several shortcomings. First, if being a principle is a matter of degree, not of kind, then we are left with a problematic grey area of not fully-fledged principles. Second, assessing the frequency of relevant events might not be easy, or even possible. Third, it is often unclear which events are relevant. Consider what Parfit calls "the non-Hedonistic Impersonal Total Principle": "If other things are equal, the best outcome is the one in which there would be the greatest quantity of whatever makes life worth living" (Parfit 1987, p. 387). What exactly is the scope of this principle? Does it apply to all outcomes? Or maybe only to outcomes *of actions*, as opposed to outcomes of natural events? Or perhaps only to outcomes of actions that can be reasonably assessed in terms of how they affect the quantity of what makes life worth living? Or something still more narrow?

Parfit's principle, like numerous other principles, contains a *ceteris paribus*, or "other things being equal" clause. The problem with the clause is that in practice other things are almost never equal. It is hard to imagine a real-life choice between maximising and not maximising whatever makes life worth living that would leave other important aspects of the situation intact. In this sense, a typical principle with a *ceteris paribus* clause has the scope of zero: it is impossible to apply it automatically. But this clearly does not mean that a typical principle is not a principle. One might be tempted to solve the problem by arguing simply that anything containing a *ceteris paribus* clause is

a principle. But this is implausible. I can say: “other things being equal, it is worse to push the fat man off the footbridge than do nothing”. If this manoeuvre can turn the footbridge judgment into a principle, that the distinction between particular cases and principles collapses.

7.2 Principles and scope

When it comes to normative principles, some philosophers distinguish between principles *qua standards* and principles *qua guides* (see Ridge & McKeever 2020). The former *explain* why certain actions – or whatever happens to be the subject of evaluation – are right or wrong, the latter tell us how we should act. Parfit’s Impersonal Total Principle can be interpreted as either of these, and its scope would probably differ depending on which interpretation is chosen. Understood as a standard it would explain a range of cases that it would not be (easily) applicable to as a guide for action, for reasons I have just outlined. It might still not be entirely clear which cases should count under the *qua standard* reading, but it is clear they would be abundant. This proposal might give us a plausible account of one kind of principles, but what about the other kind? If certain principles are principles in the guide sense only, what makes them principles?

Finally, it is doubtful that it is the scope that makes the difference in the first place. For example, we know that killing people in England nowadays is generally less common than it was centuries ago, but this does not seem to make “it is a bad thing if a person is killed” any less of a principle. Maybe the difference is not big enough to be significant. Or perhaps we still come across numerous killings in fiction, or often imagine such cases, and these cases should be taken into account?

7.3 Principles and explanation

Trying to account for principles solely in terms of their scope might be a dead end. But what would be an alternative? As I mentioned, principles – or at least principles *qua standards* – are supposed to *explain*, or give reasons. This might be what distinguishes principles from non-principles. Jonathan Dancy argues that this view must be correct at least in the area of morality:

> Moral principles, however we conceive of them, seem all to be in the business of specifying features as *general* reasons. The principle that it is wrong to lie, for instance, presumably claims that mendacity is

always a wrong-making feature wherever it occurs (that is, it always makes the same negative contribution, though it often does not succeed in making the action wrong overall). The principle that it is wrong to lie cannot be merely a generalization, a claim that lies are mostly the worse for being lies (…) (Dancy 2009, p. 76)

On the face of it, this proposal looks attractive. The difference between "it is wrong to push the fat man" and "it is a bad thing if a person is killed" would be that the latter specifies *what makes* certain actions wrong, while the former does not. But is it really the case? If the latter claim offers a reason, it is certainly not the *ultimate* reason. Many philosophers, Singer included, believe that there are things that make killing wrong: we are not dealing with any kind of rock-bottom principle with which moral justification has to stop. However, if this is so, what prevents us from concluding that the footbridge judgment also offers a non-ultimate reason? Why not interpret it as the claim that pushing the fat man, under specified conditions, "is always a wrong-making feature" of a situation? This would mean that the difference between our two claims is not that one offers a reason while the other does not. Someone like Dancy might concede that, but still point out that one claim offers a *general* reason while the other offers a non-general one. But what makes a reason a general reason? The problem of scope looms large again: just like it is hard to capture the nature of a general claim, or a principle, in terms of how many cases it is applicable to, it is hard to capture the nature of a general reason this way.

However for the purposes of my thesis drawing a sharp line between principles and non-principles is not necessary. First, we may assume that we simply know a principle when we see it – that is, we can rely on our intuition to recognise one, at least in more straightforward cases. Secondly, just like with the previous hypotheses, ones needs to remember that my ultimate goal to show that DE is false *in all possible varieties*. I admit that due to the difficulties described above it might be impossible to tell whether a particular proposition is a principle, and consequently whether the principles only hypothesis is falsified if the proposition turns out to be supported by its own intuitiveness. But this does not mean, of course, that it would be impossible to tell whether it is supported by its own intuitiveness – and to make my point, this is all I need.

8. Expertise

Perhaps not all intuitions are created equal. Maybe philosophers, due to their expertise in philosophy, have better intuitions about philosophical cases than the ordinary folk. And maybe for that reason philosophers only treat *their own* intuitions as evidence. This claim is often referred to as the "expertise defence" of the use of intuitions – dubbed a "defence" as it is meant to defend philosophy against the objection according to which philosophical intuitions are unreliable, and therefore poor source of evidence. Steven Hales argues that just like professional scientists' physical intuitions are more trustworthy that of undergraduates, the modal intuitions of professional philosophers are more trustworhty (Hales 2006, p. 171). A similar view is expressed by Kirk Ludwig:

> Philosophers are best suited by training and expertise to conduct thought experiments in their areas of expertise and to sort out the methodological and conceptual issues that arise in trying to get clear about the complex structure of concepts with which we confront the world. A lot of the problems we confront are very difficult, and so it is not surprising that in many areas there is lively debate, but it would be a mistake to think that the way to resolve such debates is to return to questioning untutored subjects in just the places where there is evidence of the most difficulty in coming to a clear view. (…) the logical end point of this process is to give the subjects we want to run our tests on an education in philosophy, and it is to admit that training in philosophy puts one in a better position in general to sort out what the proper response is to a scenario in a thought experiment. (Ludwig 2007, p. 150-1)

The expertise defence can inspire another answer to my question: counterintuitive conclusions are possible as they only contradict folk intuitions, but not philosophers' intuitions.

8.1 Conceptual competence

What is meant to make expert intuitions superior? Ludwig argues that experts "respond just to the scenario on the basis of one's competence in the use of the relevant concepts" and have "relevant background in the conceptual field intuitions about which we are interested in so that one can bring to bear a sophisticated understanding of what the issues are" (ibid.). A more substantial account of the difference between relevant and irrelevant intuitions has been offered by Antti Kauppinen (2014). On his view, there are two kinds of intuitions: *robust* and *surface* and only the former are used as evidence in philosophy. Robust intuitions are different from surface intuitions in three respects.

First, Kauppinen agrees with Ludwig that they are only generated in the minds of *competent users*. It should be clear that it is possible to misapply a concept, also in a systematic way: either due to not being sufficiently familiar with it, or perhaps for other, less obvious reasons. How do we tell whether someone is not sufficiently familiar with a concept, or otherwise likely to misapply it, and therefore incompetent? One answer could be that competent users of a concept know its meaning, which is determined by being *disposed to apply the concept in a certain way*. This, however, is problematic. Kauppinen brings up the example originally used by Kripke: people who forget to "carry" while adding large numbers do not apply the concept of "adding" as they should apply it, even though they seem to be perfectly aware of what "adding" means. What is the correct answer then? A brief review of theories in philosophy of language shows it is easier to say what does not make one competent than what it does:

> Competent users are those whose application of the concept generally matches the conceptual norms prevailing in the linguistic community. To sort out incompetent users, one must therefore identify at least the most important norms governing the concept. These norms cannot be derived from either actual use or simple dispositions, individual or collective, since the very notion of normative constraint opens a gap between what people are inclined to say about a particular case and what they should, by their own lights, say about it. (Kauppinen 2014, p. 103)

No simple theory of how to identify these norms seems to work. As I argued in section 3 of this chapter, it may be because the norms do not exist.

8.2 Performance errors and pragmatics

Secondly, robust intuitions are "generated in sufficiently favourable conditions". This means "there are no perturbing, warping or distorting factors or limits of information, access or ability" (ibid., p. 104), which cause even competent users of a concept to make what linguists sometimes call "performance errors". For example, a judgment about moral responsibility made by somebody influenced by strong emotions may not count as a judgment made in sufficiently favourable conditions.

Thirdly, robust intuitions are "based entirely on semantic considerations". The same judgment about a given philosophical case can be appropriate *semantically* and inappropriate *pragmatically*. To use Kauppinen's example, "I voluntarily had lunch yesterday" could be a semantically appropriate, but not pragmatically appropriate judgment about what I did yesterday afternoon. Having lunch is typically a voluntary activity, and because language requires us to provide only the right amount of

information, I would normally omit "voluntarily" in a conversation in this case, even though literally the judgment is true. Robust intuitions concern only what is semantically correct. A philosopher who studies voluntary action does not let intuitions about "what we would say" about my action get in her way, if "what we would say" refers to pragmatic appropriateness.

8.3 Empirical challenges

Some experimental philosophers argue that professional philosophers' intuitions are actually not very robust in Kauppinen's sense – in fact, they seem just as easily malleable by epistemically irrelevant factors as non-philosophers' intuitions. For example, one study has revealed that philosophers asked to evaluate different versions of the trolley scenario give different answers depending on the order in which the vignettes are presented to them, and on the wording of the vignettes (Schwitzgebel and Cushman 2015). Another has shown that philosophers' intuitions actually differ from non-philosophers' intuitions, however both groups fall prey to the "Actor-Observer bias": their judgments about the same situation (in this case, a variety of the trolley scenario and Bernard Williams's "Jim and the natives" scenario) differ depending on whether they are portrayed as participants in the situation or whether the situation is described from a third-person perspective (Tobia et al. 2012).

The expertise defence proponents can reply by arguing that while experts' judgments might not be much better while made *in experimental situations*, they are still better while made in the comfort of experts' armchairs, as at least some experiments are, in a sense, designed to elicit surface intuitions. For example, testing for "order effects" is bound to blur the boundary between the pragmatically appropriate and the semantically appropriate judgment (Deutsch 2009). Or perhaps testing for the Actor-Observer bias can be interpreted as deliberately introducing "perturbing, warping or distorting factors" by asking participants to imagine themselves as actors in the scenario. Experimental data might not be relevant to what experts do in their published material.

8.4 Methodological expertise

Another, more recent response to the empirical challenge is to endorse a different view of intuition-expertise. According to it, philosophers are better at *using intuitions in reasoning* rather than at *having* them. Here is how Chudnoff outlines the idea:

Suppose philosophers do not possess philosophical expertise that manifests itself in superior philosophical intuitions. Even without such a standing capacity for superior intuitions philosophers—and non-philosophers—might deliberately improve their intuitions so that they are expert-like in content. They might do this by drawing distinctions, clarifying the meanings of terms, evaluating analogies, highlighting logical form, engaging in dialectic, articulating principles, exploring models, considering extreme cases, etc. I do think there is such a deliberate effort to improve intuitions in philosophy (Chudnoff 2021, pp. 185-6)

Unlike the standard expertise view, it is unclear whether the methodological view can lend much support to DE. I agree with Chudnoff that philosophers draw distinctions, clarify meanings, evaluate analogies etc. But I think they do it indiscriminately to improve *all* judgments, intuitive and non-intuitive alike. In chapter 1 I have also argued that sometimes intuitions play a special role in philosophical methodology: philosophers rely on intuitions to clarify, persuade, or discover. I believe that because of their expertise, philosophers are generally better at doing all those things than the ordinary folk, and doing them involves relying on means listed by Chudnoff. But, of course, none of that has anything to do with DE.

I am therefore only going to assess the merits of the old expertise view. And the method of assessing it not going to differ from that of the previous hypotheses. If a philosopher relies on her own expert intuitions, one needs to look for inferences from "*p* is intuitive" to *p* in the text, and, if these inferences are absent, for reasons to think that any premises are tacitly assumed to be supported by the fact they are intuitive. The fact that the intuitiveness in question is a different kind of intuitiveness does not affect my criteria for testing DE.

9. Dualism

So far I have assumed that philosophical methodology is relatively homogeneous. Perhaps this assumption is wrong. Some have argued there are two distinct kinds of philosophical enquiry, and only one of them involves relying on intuitions as evidence. Call it the *dualism hypothesis*. The hypothesis can inspire another reply to the problem of couterintuitive conclusions: these conclusions are possible in the intuition-free kind of philosophy.

9.1 Brandt's dualism

As I mentioned chapter 1, dualism has only been proposed as a claim about the methodology of *ethics*. The first philosopher to flesh it out might have been Richard Brandt. In his 1979 *A Theory of the Good and the Right* he writes:

Historically philosophers have tried to answer the traditional questions about the good and the right in basically two ways. (Sometimes the two have been combined.) The first way is to rephrase these questions in terminology sufficiently clear and precise for one to answer them by some mode of scientific or observational procedure, or at least by some clearly stateable and familiar mode of reasoning. One of them uses this procedure (kind of reasoning) to find answers, sometimes surprising, to the stated questions. (...)

The second tradition, which I shall call 'intuitionism', can take quite different forms. Roughly the idea is that we already have presumably well-justified opinions about the answers to the several traditional questions, although these opinions need to be systematised and hence, in some cases, revised to some degree. However, the idea is not first to frame our questions clearly and then go out to find answers, letting the chips fall where they may; but rather that we roughly already know most of the answers, and if we want to know more precisely what our questions are, the best way to find out is by looking at the principles we already know, and seeing what construction of the questions is consistent with the truth or acceptability of these principles. (Brandt 1979, p. 2-3)

Brandt understands intuitionism as a kind of reflective equilibrium-seeking. The alternative approach can take two forms: what he calls "the method of appeal to linguistic intuitions" and "the method of reforming definitions". The terminology might be misleading here, as Brandt does not consider applying the former method to be an instance of intuitionism, even though it involves relying on intuitions. These are not, however, *normative moral* intuitions, which is what intuitionism treats as evidence. The method of appealing to linguistic intuitions consists in reformulating traditional moral questions to make them answerable, at least in principle, by empirical testing. The role of linguistic intuitions is to decide whether the reformulation is accurate. For example, when J. S. Mill argues that thinking of something as desirable is the same as thinking of something as pleasant, he does so supposedly because the two statements are intuitively synonymous to him. The method of reforming definitions is somewhat different: it consists in *proposing* to define moral language in naturalistic terms and accepting whatever ethical conclusions follow from this kind of proposal. Brandt's own project is to define "the best thing to do" as "the rational thing to do", according to his own understanding of "rational".

9.2 Unger's dualism

A slightly different form of methodological dualism has been put forward by Peter Unger. In his *Living high and letting die* he argues that most moral philosophers adopt the approach he calls "Preservationism", according to which "at least at first glance, our moral responses to particular cases appear to reflect accurately our deepest moral commitments, or our *Basic Moral Values*, from

which the intuitive reactions primarily derive; with all these case-specific responses, or almost all, the Preservationist seeks to preserve these appearances". This is contrasted with Liberationism, embraced by philosophers like Peter Singer or Unger himself, which assumes that "folks' intuitive moral responses to many specific cases derive from sources far removed from our Values and, so, they fail to reflect the Values, often even pointing in the opposite direction." (Unger 1996, p. 11)

To liberate our moral thinking from the deceptive influence of moral intuitions, we need to turn to what Unger calls "our general common sense" (ibid., p. 28) aimed at more general moral truths which better reflect our values, like "[one should not] contribute to the serious suffering of an innocent other, neither its initiation nor its continuation" (ibid., p. 31).

9.3 McMahan's dualism

Yet another form of dualism has been proposed by Jeff McMahan, who draws a distinction between what he calls the Theoretical Approach and the Intuitive Approach in ethics:

> According to [the Theoretical] approach, if our concern is to understand the morality of abortion, our first task must be to discover the correct moral theory. Moral inquiry is initially and primarily theoretical; only at the end of this theoretical inquiry is it possible to address moral problems such as abortion competently, bringing the theory to bear and extracting from it the knowledge we initially sought. This general approach therefore contrasts with the first approach I sketched, according to which moral inquiry begins with problems and cases and our intuitions about them, seeks principles that unify and explain the intuitions, and proceeds through adjustment and modification of both the principles and intuitions until consistency and harmony are achieved. On this approach, a moral theory in which we are entitled to have confidence is something that we can hope to have only near the end of the process of inquiry into problems of substantive morality. (McMahan 2013, p. 106)

McMahan thinks that both traditions have a significant representation in the history of philosophy: Plato's Socrates exemplifies the latter, while Kant and Hobbes exemplify the former. In contemporary philosophy the Intuitive Approach is adopted by "most philosophers working on problems of practical ethics", while Richard Hare, Richard Brandt, and "an assortment of theorists in the contractualist and consequentialist traditions" are mentioned as those who represent the Theoretical Approach.

These three dualistic accounts are by no means identical, or even compatible with each other. For example, Brandt argues that intuitionists rely on intuitions "of any level of generality" (Brandt 1979, p.18) while Unger argues that Preservationists rely on intuitions about specific cases only. It

might also be doubted whether all three contrast some form of DE with something that is not DE. For example, Unger's Liberationism might be interpreted as a way of relying on intuitions about abstract principles, along the lines of what has been discussed in section 7. However, we do not need to delve into the exegetic details here. Based on what Brandt, Unger and McMahan are saying, it is not unreasonable to suppose that there might be two different kinds of ethics, one DE-based and one DE-free. Anything that falls under this description will count as dualism.

9.4 Dualism: transparent vs opaque

Testing the dualism hypothesis seems more challenging than testing the previously described seven hypotheses. After all, no matter if it turns out that intuitions have or have not been used as evidence in a given text, both discoveries would be consistent with dualism. Suppose that testing for the previous seven hypotheses has revealed that no intuitions are used as evidence in a given text. How do we tell whether this discovery confirms dualism, or the view according to which intuitions are *never* used as evidence in ethics? It is of course impossible to answer this question without reaching beyond the text that is being analysed. One might think that the only solution is to randomly select a large number of philosophical writings and carefully examine whether any of them involves using intuitions as evidence. This, however, seems unwieldy. Fortunately, there is another solution.

To explain, let me first note that if dualism is true, either philosophers are generally aware of the fact there are two kinds of ethics, or they are not. If they are, we would expect some sort of institutional divide, similar to the one we have between analytic and continental philosophy. Intuition-based ethics would be practised in separate ethics journals, at separate ethics conferences etc. After all, if one philosopher is trying to address the problem of, for example, global poverty by accounting for intuitions that have to do with global poverty, and another philosopher addresses the same problem, but is not interested in accounting for intuitions in any way, they should probably conclude there is little point in arguing, or otherwise interacting with one another over the issue of global poverty. However it should be clear there is no such institutional divide, which means this option can be ruled out.

We are then left with the opaque version of the dualist hypothesis: dualism is true, but philosophers are to a significant extent oblivious to its truth. To find out whether this is the case, we can take a closer look at the *reception* of a given text, as we would expect the intuition-based camp to attack philosophers who reach counterintuitive conclusions. If they adopt a crude version of DE and assume intuitions can never be dismissed, their criticism would likely amount to simply pointing

out that the conclusion is counterintuitive and therefore cannot be correct. If they are more nuanced and accept that intuitions can sometimes be sacrificed, they could still complain that, for example, what has been preserved is overall less intuitive than what has been rejected, or that it is not intuitive to expert intuitiers, or that its intuitiveness is born out of a cognitive bias, etc. Each criticism can be examined in the light of the reconciliation theories I have outlined. If, on the other hand, it turns out that no such DE-based criticism have been offered, this fact would provide strong evidence against the dualism hypothesis.

CHAPTER 4. Case studies

1. Introduction

In the previous chapter I have described eight hypotheses – Cognitive Bias, Conceptual Analysis, Reflective Equilibrium, Theoretical Virtues, Arbitrariness, Principles Only, Expertise and Dualism – that reconcile the idea of relying on intuitions with the fact of dismissing intuitions. Except the first and the last one, all of them have one thing in common: they assume that some premises in the argument must be treated as supported by the fact they are intuitive. The cognitive bias hypothesis is slightly different – it assumes that some intuitive propositions must treated as less plausible because their intuitiveness stems from a bias. Finally, the dualism hypothesis implies either an institutional divide, or a specific kind of criticism of philosophers who reach counterintuitve conclusions. It is important to stress that I am only interested in these hypotheses *in their DE-friendly form*. I am not in principle opposed to the idea that philosophers seek reflective equilibrium, or analyse concepts, or utilise their philosophical expertise, or arbitrarily select their argumentative starting points etc. *in some sense*. Neither am I opposed to the idea that philosophers rely on intuitions *in some sense*. My thesis is that what is typically meant by "relying on intuitions in philosophy" is DE, and DE is false.

In this chapter I focus on three cases of reaching a strongly counterintuitive conclusion in contemporary ethics: Michael Tooley's defence of infanticide, David Benatar's defence of antinatalism, and John Taurek's attack on the idea of moral quantification. I am going to test each of these arguments against each of my eight hypotheses to see whether any of the arguments has anything to do with DE. The outcome is going to be thoroughly negative: not only Tooley, Benatar and Taurek fail to rely on any intuitions, the same is true of the philosophers who have critically interacted with them or debated their unargued assumptions.

2. Tooley on infanticide

If we were to make a list of things that just seem wrong to almost everyone, killing babies would surely be a top contender. Irrespective of which theory of intuitiveness one adopts, the claim that killing babies is wrong seems to easily meet the criteria: people tend to make this judgment spontaneously and without consciously deriving it from any general principles, it has a strong "feels wrong" phenomenology to it, its content "presents itself as necessary", it is not perceptual or memory-based, and so on. And yet in 1972 Michael Tooley published an article in which he argues that infanticide is in fact morally permissible. How was it possible for him to reach this conclusion?

2.1 Tooley's metaphilosophical views

First, let us ask: does Tooley comment on the fact that the claim he is arguing against is strongly intuitive? The closest he gets to addressing the issue is the following passage:

> The typical reaction to infanticide is like the reaction to incest or cannibalism, or the reaction of previous generations to masturbation or oral sex. The response, rather than appealing to carefully formulated moral principles, is primarily visceral. When philosophers themselves respond in this way, offering no arguments, and dismissing infanticide out of hand, it is reasonable to suspect that one is dealing with a taboo rather than with a rational prohibition. (Tooley 1972, pp. 39-40)

It appears that Tooley does not make much of the bare intuitiveness of the prohibition against infanticide, or the bare intuitiveness of any other moral claim. On the other hand, in his 1983 book – a substantially expanded version of the 1972 article – he writes that "agreement with the moral feelings of people makes it at least somewhat more likely, other things being equal, that a given moral principle is correct" (Tooley 1983, p. 27). This is supposed to apply only to basic, non-derivative principles that are not peculiar to a culture or a historical period. Nevertheless, it may appear that Tooley endorses DE in some restricted form. However, as I have argued in the previous chapters, when it comes to questions about philosophical methodology, we should not take philosophers' opinions for granted. Many philosophers hold mistaken metaphilosophical views, in particular many think they rely on intuitions, while in fact they do not. To find out what Tooley's methodology is, we need to take a closer look at his argument.

2.2 Tooley's argument

Tooley believes that the main reason why people think infanticide is wrong is that infants have a serious right to life. But, according to him, this claim is mistaken. From that it does not, of course, follow that infanticide is morally permissible – there might be other reasons to denounce it, but Tooley puts this problem aside by simply assuming there are no such reasons. The central premise of the argument is this: "An organism possesses a serious right to life only if it possesses the concept of a self as a continuing subject of experiences and other mental states, and believes that it is itself such a continuing entity" (Tooley 1972, p. 44). Since infants are not organisms of this kind, they do not possess a serious right to life. Tooley supports his central premise with the following claims: "To ascribe a right to an individual is to assert something about the prima facie obligations of other individuals to act, or to refrain from acting, in certain ways" and "the obligations in question are conditional ones, being dependent upon the existence of certain desires of the individual to whom the right is ascribed" (ibid.). These two premises appear to be his starting points – Tooley does not justify either of them in the text.

After establishing what rights are, he moves on to explain what it means to have a right *to life*. He argues that "life" is not to be understood in terms of a continued existence of a biological organism: neurosurgical intervention leading to a complete change of someone's beliefs, desires, memory, personality etc. would count as a violation of someone's right to life even though it would not kill the biological organism. Rather, "life" should be understood as a continued existence of, as Tooley puts it, "a subject of experiences and other mental states".

This results in the following analysis: "A has a right to life" is synonymous with "A is a subject of experiences and other mental states, A is capable of desiring to continue to exist as a subject of experiences and other mental states, and if A does desire to continue to exist as such an entity, then others are under a prima facie obligation not to prevent him from doing so" (ibid., p. 46). Tooley admits that this analysis is not entirely correct, as there are three types of situations when it is possible to violate someone's right to life even when they do not desire to continue to exist, namely "(i) situations in which an individual's desires reflect a state of emotional disturbance; (ii) situations in which a previously conscious individual is temporarily unconscious; (iii) situations in which an individual's desires have been distorted by conditioning or by indoctrination" (ibid., p. 47). He provides us with an examples of each situation. Here is one of his two examples of the first:

consider a case in which an adult human falls into a state of depression which his psychiatrist recognizes as temporary. While in the state he tells people he wishes he were dead. His psychiatrist, accepting the view that there can be no violation of an individual's right to life unless the individual has a desire to live, decides to let his patient have his way and kills him. (ibid.)

Of course, Tooley thinks that the psychiatrist violates his patient's right to life. He makes a number of similar judgments about other scenarios. These judgments can be interpreted as Tooley's starting points, but also as mere *illustrations* of more general claims that serve as starting points, like "if a desire to continue to exist is absent because of a certain emotional disturbance, the individual who lacks this desire has the same right to life as someone who does not". For reasons I detailed in chapter 1, it is not easy to determine which reading is correct. Since I do not need to solve this problem to make my point, I am going to stay neutral on this issue.

Tooley does not come up with a definite, clear-cut modification of his analysis of the right to life. He only points out the analysis would have to accommodate his exceptions, and that the exceptions have one thing in common: the presence of "the conceptual capability of desiring the thing in question" (ibid., p. 49). The problem with infants is they lack this capability – they cannot desire to continue to exist as subjects of experiences and other mental states. This means they cannot possess a right to life.

2.3 Judgments about cases

Let us now ask the fundamental question: can we reconcile the fact that Tooley dismisses the intuition that infanticide is wrong with DE? In the previous chapter I described eight possible ways of such reconciliation. Six of them require that the author at some point relies on at least one intuition as evidence for its content. Can we find it? Proponents of DE would most likely be tempted to argue that Tooley's thought experiments, such as the story about the psychiatrist killing his patient, are excellent examples of relying on intuitions: Tooley presents us with a scenario that triggers the intuition that killing the patient violates his right to life, and this intuition is used as evidence that the patient's right to life is indeed being violated.

But one can also think of DE-unfriendly interpretations of Tooley's use of thought experiments. First, the claim that killing the patient violates his right to life may be *inferred from* a more general claim, and this more general claim would serve as a starting point in Tooley's argument. Secondly, the claim that killing the patient violates his right to life may be treated as part of the common ground – something Tooley assumes his readers agree with him about, irrespective of whether anyone finds it intuitive. Which interpretation is correct?

I have made the distinction between explicit and tacit DE. If the latter is true, then we would expect Tooley to make claims like "it seems that the psychiatrist violates his patient's right to life by killing him, which supports the claim that the psychiatrist violates his patient's right to life by killing him".

But nothing remotely like that can be found in the text. This leaves us with the tacit variety. To test it, we should imagine asking Tooley: why think that killing the depressed patient violates his right to life? How likely would he be to answer: "because it is intuitive"? Moreover, how likely would he be to say that this argument is so obvious to his readers he did not need to make it explicit?

In chapter 1 I mentioned David Boonin and Don Marquis and their debate over why people who do not want to live may have a right to life. To remind the reader – Boonin endorses the following principle: "If an individual P has a future-like-ours F and if P has a present, dispositional and ideal desire that F be preserved, then P is an individual with the same right to life as you or I." As the depressed patient meets these criteria, he has a right to life. Marquis thinks the patient has a right to life because he will later desire to continue having experiences contained in F. Boonin and Marquis clearly disagree about why Tooley's judgment is true, but neither of them seems to believe that the intuitiveness of the judgment has any evidentiary role to play – they simply never bring it up. Nor does any other philosopher, as far as I am aware, in the context of this discussion. This does not bode well for DE.

Here proponents of DE might object that Boonin and Marquis, as well as Tooley, start with the judgment about the impermissibility of killing any depressed person and abductively infer their principles from it. And the fact they all start with the same judgment is best explained by DE: they all tacitly assume it is supported by its own intuitiveness. However it is not impossible to find a justification of the judgment that is not circular in this way. For example, James Griffin argues that human rights, including the right to life, are grounded in our normative agency, which is what makes them universal. One possesses their human rights irrespective of whether they can or want to exercise them. The fact that someone is very shy and does not mind not being allowed to speak does not waive their right to free expression, as "to be a tolerably successful self-decider typically requires an ability to ask questions, hear what others think, and so on" (Griffin 2008, p. 49). Thus silencing the shy person may well violate her right. Similarly, we may suppose, one does not lose their right to life when they wish they were dead – especially if the wish is only temporary – and killing them may well violate this right. This shows that Tooley's judgment can be justified without being presupposed, and, again, that considerations brought up to justify it have nothing to do with its intuitiveness.

There is also another, even stronger reason to question the DE-friendly interpretation: the fact that Tooley's starting point can be *rejected*. Admittedly, I am not aware of a philosopher who has explicitly argued that killing a temporarily depressed person who wishes to die does not violate their right to life. However we can easily imagine a utilitarian saying that under certain

assumptions, such as the unhappiness experienced in the depressive episode being much greater than the happiness experienced afterwards, killing the person might be permissible. Moreover, many utilitarians would argue that there is no such thing as a right to life, as it is commonly conceived (see, for example, Brandt 1984). And, again, these arguments never appeal to intuitions in any DE-friendly sense.

This shows that while our judgment seems plausible enough to be used as a starting premise in an argument, it is far from being universally accepted. If DE were true, then we would expect philosophers, both those who endorse the judgment and those who reject it, to turn to questions about the judgement's intuitiveness. Surely the latter would try to undermine the judgment by arguing that it does not seem true to everyone, or it does not seem true as strongly as one might think, or it conflicts with judgments that seem true more strongly or more universally. But this never seems to happen. In short, Tooley's starting point (if it is in fact a starting point) is contested, but contesting it never involves investigating whether or why it is intuitive.

2.4 Core principles

What I have just said about Tooley's judgment about the psychiatrist killing his patient seems to apply to all other judgments about particular cases he makes. How about his more abstract claims? Consider the first two starting premises I mentioned: one according to which rights have to be defined in terms of corresponding obligations and one according to which these obligations are conditional, depending on the existence of certain desires of right-holders. Some proponents of DE may argue that Tooley's evidence for these claims is the fact that the claims are intuitive. This would perhaps be consistent with how Tooley characterises his own method in the 1983 book: while he argues that it is irrelevant how judgments about particular cases feel to us, it is not entirely irrelevant how we feel about judgments about basic principles.

Here, again, making a case for the explicit variety of DE looks hopeless. Tooley does not say anything about the intuitiveness of either of these claims, let alone suggest that their intuitiveness constitutes a reason to accept them. Does he treat this point as too obvious to be stated? This seems highly implausible. Philosophers treat Tooley's principles just like they treat the judgment about killing a person who wishes to die, namely they contest both their justification and their truth, and they never seem to appeal to these principles' intuitiveness in the process. For example, Onora O'Neill writes that rights can be understood as existing independently of anyone's obligations (O'Neill 2005, pp. 429-30). She would most likely admit that in the case of a right to life this

account is unworkable, but it remains the fact that what she says is incompatible with Tooley's first principle, as presented in his paper.

Or consider the interest theory of rights, defended by philosophers such as Joseph Raz. Raz argues that "X has a right if X can have rights, and, other things being equal, an aspect of X's well-being (his interest) is a sufficient reason for holding some other person(s) to be under a duty" (Raz 1986, p. 166). This definition does not mention desires of any kind. It is true that well-being can be understood in terms of informed or otherwise idealised desires, but it can also be understood in a desire-independent way (see Crisp 2021). Adopting the latter option would threaten Tooley's second principle. Were DE to be true, we would expect conflicts like this to be addressed by appeal to the intuitiveness of competing claims. The problem is – they simply are not. O'Neill does not argue that the possibility of obligation-independent rights is real because it is more intuitive than Tooley's first principle, Raz does not argue that his interest-based account of rights is more intuitive than Tooley's second principle, and so forth. The idea that there is some sort of widespread consensus over Tooley's abstract intuitions being treated as evidence of their contents is just as far-fetched as the idea of a similar consensus with respect to his intuitions about particular cases.

2.5 Reception

By this point it should be clear that none of the six ways of reconciling DE with the counterintuitiveness of a conclusion works in the case of Tooley's argument. The cognitive bias hypothesis can also be rejected – Tooley does not try to argue that our aversion to infanticide is evidentially irrelevant because it is born out of a bias. The proponent of DE's last resort is therefore the dualism hypothesis. Perhaps Tooley does not rely on any intuitions at all, and neither do many others who discuss the morality of infanticide. But it is only because they represent the intuition-free style of doing ethics, which exists along the intuition-based style.

In the previous chapter I distinguished two varieties of the dualism hypothesis: the transparent and the opaque. According to the former, philosophers realise that there are two inconsistent methodologies of philosophy, or perhaps just of ethics. If it is true, we would expect them to keep the two institutionally apart. But this is clearly not the case: Tooley did not publish his article in any kind of intuition-free ethics journal, he did not present his argument at intuition-free ethics conferences etc. No such journals or conferences seem to exist. How about the oblivious variety? If there are philosophers who standardly account for intuitions, in the DE sense, and they are unaware that other philosophers are not interested in accounting for intuitions, we would expect the former to

react to Tooley's work in a particular way. They would chastise Tooley for dismissing what he was supposed to account for, or, at the very least, they would compare the intuitiveness of what Tooley accepts with the intuitiveness of what he dismisses.

But if we examine reactions to Tooley's article or his book, they are nothing like this. Sure enough, his conclusion sparked extensive criticism, however none of it had much to do with accusing Tooley of failing to account for intuitions. For example, in their reply Mark Tushnet and Louis Michael Seidman write that even if Tooley is right that we should not refrain from killing infants *for the sake of the infants themselves*, we clearly should refrain from killing them for the sake of their parents, other persons, or the society as a whole. Any successful defence of infanticide should deal with these third party-centred claims, but Tooley's argument does not address them (Tushnet & Seidman 1983). L. W. Sumner complains about how Tooley simply presupposes that rights must be tied to desires, without offering any argument. According to Sumner, it would be more plausible to base rights in sentience, or the ability to suffer and feel enjoyment. And this analysis would probably not lead to the conclusion that infants do not have a right to life (Sumner 1983, p. 539). Christopher Kaczor argues that when Tooley rejects the idea of infanticide being wrong because of the infant's *potential* desires, he conflates "active potentiality" with "passive potentiality". Infants are eventually going to gain relevant desires through the process of "self-propelled" development, which makes them morally different from beings that can only gain these desires via an intervention from the outside – and Tooley only focusses on the latter in his critique of the potentiality argument (Kaczor 2015, p. 28). All these objections make perfect sense on the assumption that DE is false, but not so much sense on the assumption that DE is one of the two significant approaches to ethics.

Towards the end of his reply, Sumner briefly addresses the issue of intuitiveness. He writes that he is not going to "settle the large issue of how much weight should be assigned to counterintuitive results in deciding whether to accept a moral principle", but "if intuitions are to be given any weight at all, then it is not easy to discount intuitions about infanticide" (ibid., p. 543). It may seem that at last we have found an argument – albeit vague, cautious and qualified – that fits into the DE picture. However before reaching this conclusion, we should ask ourselves: why is this qualified statement the best we can get? Here is my tentative answer: when philosophers argue, they instinctively reject DE. When philosophers think about how they argue, they often find DE plausible. I think Sumner's remarks reflect this tension. He is not prepared to openly accuse Tooley of failing to account for intuitions, in the DE-sense, because as soon as this objection is formulated, it just looks bizarre and unphilosophical. On the other hand, he thinks others should be making this kind of objection. So long as DE remains in the sphere of metaphilosophical theorising, it does not sound unreasonable,

however put into practice it immediately falls apart. Disentangling first-order philosophical practice from metaphilosophical comments can be challenging, but when it is done successfully, we are left with no reasons to accept DE.

3. Benatar on the harm of existence

Tooley's argument, as subversive as it may be, does not undermine the idea that it is fine to have children. But not all philosophers agree. For example, David Benatar has written a book in which he argues that coming into existence is always a serious harm, which means that the universe would be better off without conscious life forms, procreation is always immoral, and we should die out as quickly as possible. Just like Tooley's claim, this conclusion is deeply counterintuive on any major account of intuitiveness. How was it possible for Benatar to reach it?

3.1 Benatar's metaphilosophical views

Let us begin with Benatar's stance on appealing to intuitions in ethics. In his book, he provides us with a helpful section called "Countering the counter-intuitiveness objection", where he writes:

> At the outset, it is noteworthy that a view's counter-intuitiveness cannot by itself constitute a decisive consideration against it. This is because intuitions are often profoundly unreliable—a product of mere prejudice. Views that are taken to be deeply counter-intuitive in one place and time are often taken to be obviously true in another. The view that slavery is wrong, or the view that there is nothing wrong with 'miscegenation', were once thought to be highly implausible and counter-intuitive. They are now taken, at least in many parts of the world, to be self-evident. It is not enough, therefore, to find a view or its implications counter-intuitive, or even offensive. One has to examine the arguments for the disliked conclusion. Most of those who have rejected the view that it is wrong to create more people have done so without assessing the argument for that conclusion. They have simply assumed that this view must be false. (Benatar 2006, p. 203)

According to Benatar, we have a good reason to be sceptical about our pro-natal intuitions: they are a product of our evolutionary past:

> Those who do not have this belief are less likely to reproduce. Those with reproduction-enhancing beliefs are more likely to breed and pass on whatever attributes incline one to such beliefs. (ibid., p. 204)

Since there is no obvious link between making one more likely to reproduce and being true, the fact that procreation generally seems at least acceptable to us does not constitute good evidence in favour of the acceptability of procreation. Having said that, Benatar believes that intuitions count for something. By writing that a view's counterintuitiveness "cannot by itself constitute a decisive consideration" he implies that it can constitute a non-decisive consideration, possibly in a DE-friendly sense. Moreover, he points out that accusing him of running counter to deeply held intuitions is a case of the pot calling the kettle black. Those who reject his conclusion are committed to a number of strongly counterintuitive views themselves, whether they realise it or not. This reply can be interpreted in both a DE-friendly, and a DE-unfriendly way – along the lines of "we all place intuitive propositional contents in the common ground, and end up negating other intuitive contents". All in all, at minimum Benatar occasionally says things that sound like endorsing some variety of DE. But just like Tooley and many others, he can be wrong. The only way to determine whether and how he relies on intuitions is to examine his argument.

3.2 Benatar's argument

Benatar's case for antinatalism has three main steps. First, he argues that there is a certain asymmetry of harms, such as pain, and benefits, such as pleasure: the presence of the former is bad and the presence of the latter is good, but while the absence of the former is good, the absence of the latter is *not bad*. Secondly, he argues that from this asymmetry it follows that coming into existence is always a harm. Finally, he argues that even the best lives are very bad lives, which means that coming into existence is not only always harmful, but always *seriously* harmful. It follows that there are no circumstances that can justify seriously harming one's children by bringing them to this world.

Some philosophers believe that coming into existence can *never* be a harm. This is because harm must be a "worse off" relation between two states one can find oneself in, but non-existence is not a state of this kind. Benatar starts off by responding to this argument. One might argue that a harm does not necessarily need to be a "worse off" relation – it can simply be something non-comparatively *bad*. However Benatar prefers a view according to which a harm can be a "worse off" relation between existence and non-existence. The fact that someone can be harmed (or benefited) by dying aptly illustrates this point. It might be pointed out that Epicurus has famously challenged this view. According to him, "death, the most frightening of bad things, is nothing to us; since when we exist, death is not yet present, and when death is present, then we do not exist"

(Epicurus 1994, p. 29). Benatar does not refute the argument. Instead, he writes "we seem to have an impasse" between those who accept it and those who dismiss it (ibid., p. 217). It appears that the former are not going be convinced by Benatar's reasoning – but clearly most people who disagree with antinatalism belong to the latter group.

After establishing that it is not impossible for coming into existence to be harmful, Benatar moves on to argue that coming into existence is *always* harmful by defending his axiological asymmetry of pain and pleasure. He does so by arguing it best explains four widely accepted judgments. First: "there is a duty to avoid bringing suffering people into existence, there is no duty to bring happy people into being" (ibid., p. 32). Second: "whereas it is strange (if not incoherent) to give as a reason for having a child that the child one has will thereby be benefited, it is not strange to cite a potential child's interests as a basis for avoiding bringing a child into existence" (ibid., p. 34). Third:

> Bringing people into existence as well as failing to bring people into existence can be regretted. However, only bringing people into existence can be regretted *for* the sake of the person whose existence was contingent on our decision. (ibid., p. 34)

And fourth:

> Whereas, at least when we think of them, we rightly are sad for inhabitants of a foreign land whose lives are characterized by suffering, when we hear that some island is unpopulated, we are not similarly sad for the happy people who, had they existed, would have populated this island. Similarly, nobody really mourns for those who do not exist on Mars, feeling sorry for potential such beings that they cannot enjoy life. (ibid., p. 35)

Benatar admits that the asymmetry is not the only possible explanation of some of these judgments, so he tries to undermine alternative explanations. Most notably, the first judgment can make sense if one believes that negative duties are weightier than positive duties, without assuming that the absence of pain is good while the absence of pleasure is not bad. The duty not to create suffering beings is simply a negative one, that is a duty *to refrain from doing something*, and this is what makes it more stringent than the counterpart positive (hypothetical) duty to create happy people. However, replies Benatar, those who believe that negative duties are weightier typically also believe that *some* positive duties exist, and what makes the former weightier is the fact they can be fulfilled *without making sacrifices*, unlike their counterpart positive (hypothetical) duties. It takes little effort not to procreate, but it takes a considerable effort to bear a child. This means that if it were possible to create a happy person at no great cost to oneself, there would be a duty to create this person. But

this, according to Benatar, is implausible, as there is still *an asymmetry of procreative moral reasons* which dictates that no reason to procreate exists irrespective of sacrifices involved (ibid., pp. 33-34).

The next step in the argument is to show to the asymmetry implies that coming into existence is always a harm. One reason why people are resistant to this conclusion is they find it difficult to understand how the presence of pleasure of an existing person X (which is good) is not advantageous over the absence of X's pleasure in a scenario when X does not exist (which is neither good nor bad). To make this point, Benatar offers the following analogy:

> S (Sick) is prone to regular bouts of illness. Fortunately for him, he is also so constituted that he recovers quickly. H (Healthy) lacks the capacity for quick recovery, but he *never* gets sick. It is bad for S that he gets sick and it is good for him that he recovers quickly. It is good that H never gets sick, but it is not bad that he lacks the capacity to heal speedily. The capacity for quick recovery, although a good for S, is not a real advantage over H. (ibid., p. 42)

Similarly, the presence of pleasure in the "existence" scenario cannot be an advantage over the absence of pleasure in the "non existence" scenario. On the other hand, the absence of pain in the latter is a clear advantage over the presence of pain in the former.

So far Benatar has argued that there is an asymmetry of pain and pleasure, which implies that coming into existence is always a harm. However concluding that procreation is always morally wrong would be premature at this point. Perhaps the net badness of coming into existence is not always very significant, and can be outweighed by benefits such as satisfaction that comes with rearing a child. Benatar rejects this idea in step three by arguing that this is not the case. He does so by arguing that all lives go very badly, which we typically fail to realise.

This is because of three psychological phenomena. First, there is what Benatar calls Pollyannaism – a general tendency to be irrationally optimistic. When we think about our past, our future and our present state, we typically focus on the good and ignore the bad. Secondly, there is the phenomenon of adaptation: when our lives take a turn for the worse, we quickly get used to the new situation. Thirdly, when we think about the quality of our lives, we tend to compare ourselves with others rather than stick to some objective standard. Benatar cites a host of empirical research to support these claims – for example, studies showing that most people assess the level of their well-being as above average (ibid., p. 66).

We need to overcome these biases and concentrate on how well our lives go according to major philosophical theories of well-being: the hedonistic accounts, the desire-satisfaction accounts and

the so-called objective list accounts. And irrespective of which of these accounts is adopted, it turns out that there is no such thing as a good life in the actual world. For example, according to hedonism a life goes well if positive mental states contained in it dominate over negative mental states. However, argues Benatar, even the best lives are marked by a multitude of negative states, albeit often mild, such as pain, discomfort, irritation, boredom, shame, tiredness, hunger, thirst etc. On the other hand, many of life's pleasures are merely *relief pleasures*: relatively short moments of good possible only because they conclude long periods of bad. Moreover, non-relief pleasures are also few and far between. The overall balance of the positive and the negative must always be negative. The conclusions for non-hedonistic theories are similar: if we apply the criteria in an unbiased way, no actual life can be deemed good.

3.3 Judgments about cases

Is Benatar appealing to intuitions in any DE-friendly sense? As proponents of DE tend to focus on judgments about cases, typically thought experiments, let us examine these first. I have cited Benatar's scenario about H and S and his judgment about it: the ability to quickly recover is not S's advantage over H. It might be argued that the judgment is intuitive. But does Benatar treat it as supported by the fact it is intuitive? This is highly dubious. First, he never makes any claims in the form of "it seems to us that S's ability to recover quickly is not an advantage over H, therefore S's ability to recover quickly is not an advantage over H". Secondly, he offers a justification for the judgment:

> This is because the absence of that capacity is not bad for H. This, in turn, is because the absence of that capacity is not a deprivation for H. H is not worse off than he would have been had he had the recuperative powers of S. S is not better off than H in any way, even though S is better off than he himself would have been had he lacked the capacity for rapid recovery. (ibid., p. 42)

The language used in this passage ("this is because") strongly suggests that the justification interpretation I described in chapter 1 is true of the judgment. What makes the judgment true, or gives us a reason to believe it is true, has nothing to do with the judgment's intuitiveness, and everything to do with whether the absence of the capacity constitutes a deprivation. Put another way, the judgment is not even an argumentative starting point, let alone a starting point taken to be tacitly supported by its own intuitiveness. This is not to say that its intuitiveness is irrelevant in every respect. I think it is fair to say that Benatar uses an intuition as a device of clarification, and

possibly also as a device of persuasion. The concrete example makes the argument more accessible – but this does not change the argument substance, and does not support DE in any way.

While the scenario about H and S seems to be the most important thought experiment in the book, Benatar makes several other judgments that can be classified as judgments about cases. To avoid making this section excessively long, I will leave these them for the reader to evaluate. In my view, a little scrutiny reveals that none of these judgments is relied on in any DE-friendly sense.

3.4 Core principles

How about more abstract claims? The obvious candidates for intuitions being relied on in a DE-friendly sense would be the four judgments used to support the axiological asymmetry. It is worth noting that Benatar explicitly calls them his starting points. But he also stresses that he treats them as starting points *as they are widely accepted*, and that the fact they are widely accepted *is not a reason to think they are true* (ibid., p. 36).

Moreover, he points out that none of the four judgments is shared *universally*. For example, he writes that certain utilitarians may be inclined to reject the first judgment and argue that there is a kind of duty to procreate. Benatar does not name any names, but it is not hard to find examples. Stuart Rachels argues that it is good to bring happy people into existence and that potential people's happiness is as morally important as that of actual people. He points out that this does not necessarily mean that we are *obliged* to create happy people (Rachels 1998, p. 94), however he does not seem to be opposed to the idea either. Another, less ambiguous example is an article by Torbjörn Tännsjö, who argues that "to the extent that we add creatures living lives worth living, our ambition to replenish the universe not only is part of our quest for meaning, but also means that we comply with our duties as moral agents" (Tännsjö 2002, p. 355).

Neither Rachels nor Tännsjö seems to assume that the intuitiveness of what they question matters in evidential terms. Rachels examines eight arguments against his position, and none of them is remotely anything like "it is intuitive that there is no duty to make happy people, which suggests that there is no duty to make happy people". On DE, this is strange. Why is Rachels ignoring the most basic and universally recognised evidence against the view he is defending? How did the editor and the reviewers of *Bioethics* overlook this omission? Even if Rachels does not, after all, commit himself to the stronger, "duty" version of his view, we would expect him to address this

point, making it clear that what he actually rejects is not very intuitive, or that the intuitiveness of what he rejects is trumped by his arguments, etc.

Similarly, if DE were true, we would expect Tännsjö to argue that the intuitiveness of "there is no duty to create happy people" is either no evidence that there is no duty to create happy people – despite what is commonly believed – or that it is trumped by some evidence to the contrary. But Tännsjö says neither of these things. Sure enough, he tries to explain why his conclusion seems false to many of us, but he spends no time explaining why the fact it seems false is no reason to dismiss it, or a relatively weak reason to dismiss it. This is most likely because hardly anyone takes it to be such reason – there is no need to undermine the consensus as the consensus does not exist.

It might be objected that what Benatar means by "there is no duty to bring happy people into being" is "there is no duty to bring happy people into being *for the sake of those happy people*", and utilitarians like Rachels or Tännsjö do not necessarily deny it. Rather, they tend to think we only have a duty to make happy people *for the sake of the sum of all happiness*, whose maximisation is the sole goal of morality. However even if this objection is valid, and even if all philosophers on Earth agree with Benatar's starting premises, there are still no grounds to think that these premises are meant to be tacitly supported by the fact they are intuitive. It is perfectly legitimate to question and defend them by appealing to all sorts of considerations. For example, one might argue that fulfilling the alleged duty would turn many women into constantly pregnant procreation-machines, which would violate their autonomy, or that we cannot have duties towards non-existent beings whose existence is dependent on the fulfilment of the alleged duty. On the other hand, it would be bizarre to argue that the duty to create happy people does not exist because we tend to form this judgment spontaneously, or because making it is accompanied by a distinct phenomenology, or because we cannot think of why it is true, etc. There is something deeply unphilosophical about assertions of this kind, and they simply would not count as reasons in a serious conversation. However DE in its tacit form implies not only that such reasons can be respectable, but that they constitute the most obvious and universally recognised evidence.

Moreover, if Benatar believed there were a consensus over the fact that these judgments are true, or more likely to be true, because they are intuitive, he would not be discussing the relation between their popularity and their truth without mentioning the relation between their intuitiveness and their truth. The way he presents the judgments is simply at odds with DE, and explicitly in line with the common ground interpretation.

There are, of course, other abstract claims that Benatar seems to be assuming without argument. For example, at one point he invokes “a principle of equality”, which dictates that the same morally relevant interests count equally, irrespective of factors like species membership (ibid., p. 143). He never tries to prove or argue for this principle, and yet clearly thinks it should be accepted. Is it plausible that he tacitly assumes the principle to be supported by the fact it is intuitive? There is no evidence he does. Again, it is helpful to ask whether the principle is or can be challenged. An example of such challenge – at least on one interpretation of the principle – would be a book by Stephen Schwarz, who argues that human interests matter more than similar interests of other species. This is because every homo sapiens has “the basic inherent capacity to function as a person, regardless of how developed this capacity is, or whether or not it is blocked” (Schwarz 1990. p. 101). Schwarz never mentions any consensus over the fact that the principle of equality is supported by its own intuitiveness. Nor does he try to outweigh the intuitiveness of the principle with some evidence against it. This strongly suggests the consensus does not exist. Someone might object that perhaps Schwarz accepts the principle as an “other things being equal” claim – he only argues that when we compare human and non-human interests, other things are not equal. But even if this is the case, on DE we would expect him to clarify his position, explaining that what he dismisses is not exactly what intuition supports. The fact that he does not offer any explanation of this kind does not bode well for DE. On the other hand, the common ground interpretation of how Benatar treats the principle of equality works perfectly – just like it does applied to all other starting premises in Benatar’s argument.

3.5 Cognitive bias defence

Throughout his book Benatar talks about the pro-natal bias, by which he means Pollyannaism, the ease of adaptation and the assumption of the comparative nature of well-being, together with a possible evolutionary story behind these psychological phenomena. The bias is supposed to explain why both the conclusion – that coming into existence is always a serious harm – and one of the crucial premises – that even the best lives are very bad lives – seem false to us. Proponents of DE might be tempted to use this fact as evidence for DE: our pro-natal intuitions must be treated as evidence of their contents, as Benatar tries to undermine them by showing they are shaped by a faulty heuristic.

However, as I argued in the previous chapter, the fact that someone explains away certain intuitions this way does not yet weigh in favour of DE. For DE to be supported, the author would need argue

that certain intuitive propositions are false, or more likely to be false, *because* what makes them intuitive is an unreliable psychological mechanism. But it should be clear that Benatar does not do anything like that. He is not trying to say that even the best lives are very bad lives, or that coming into existence is always a serious harm, or that procreation is always morally wrong, *because* our intuitions to the contrary are shaped by natural selection, or whatever it might be. According to Benatar, even the best lives are very bad lives because correctly applying criteria for having a good life, as specified by different theories of well-being, leads to this conclusion. Coming into existence is always a serious harm because this is what follows from the axiological asymmetry of pleasure and pain, etc.

Even if the pro-natal bias did not exist at all, the substance of Benatar's argument would not change one iota. Perhaps he would not have written his book, as in this situation antinatalism would be a common wisdom, or, more plausibly, there would be nobody to write it in the first place, as humanity would be long extinct. In any case, the bias is clearly not *why* antinatalism is true. Here proponents of DE might ask: if this is so, why does Benatar goes to such lengths to discuss the pro-natal bias? But this question has a fairly straightforward answer: realising that the bias exists is supposed to *help us understand* why his argument is sound.

Benatar could have chosen to ignore the issue of psychology and focus solely on his premises and how his conclusion follows from them. For example, he could have left out the following passage:

> Of course, we tend *not* to think about how much of our lives is marked by these states. The three psychological phenomena, outlined in the previous section, explain why this is so. Because of Pollyannaism we overlook the bad (and especially the relatively mildly bad). Adaptation also plays a role. People are *so* used to the discomforts of daily life that they overlook them entirely, even though they are so pervasive. Finally, since these discomforts are experienced by everybody else too, they do not serve to differentiate the quality of one's own life from the quality of the lives of others. The result is that normal discomforts are not detected on the radar of subjective assessment of well-being. (ibid., p. 72)

The argument would have stayed intact, but it would probably be less accessible and less appealing. Merely stating that our daily discomforts are pervasive, no matter how meticulously, does not have the same effect as stating it *and* explaining why we tend to overlook them. However trying to achieve such an effect clearly does not amount to treating intuitions as evidence of their contents.

3.6 Reception

So far I have rejected seven of the eight hypotheses described in the previous chapter. They all fail because Benatar does not treat any of his premises as supported by the fact they are intuitive, or any of his opponents' intuitive claims as less plausible because of being born out of a cognitive bias. The last-ditch defence of DE would therefore be the hypothesis of methodological dualism. According to it, DE may be false with respect to Benatar's book, but it is still true with respect to other philosophical work. As I argued in the previous section, this hypothesis can only gain a whiff of plausibility in its opaque variety, according to which philosophers generally do not realise they are divided into two methodological camps. This means we should expect Benatar to be accused of failing to account for intuitions. Can we find any such responses?

Benatar addresses major criticisms of his book in his 2013 article *Still better never to have been*. Here is a very concise list of these replies. David DeGrazia doubts that merely possible persons can be harmed or benefited by not being brought into existence. Elizabeth Harman argues that the four judgments can be, after all, adequately explained by the fact that positive duties are weaker than negative duties, rather than by the axiological asymmetry. Chris Kaposy argues that what best explains the first two of the four judgments is that that there is value in *avoiding being the cause of suffering*, rather than that there is value in the absence of suffering. Tim Bayne argues that the four judgments are better explained by an asymmetry between good and bad *lives*, not an asymmetry of good and bad experiences. Ben Bradley argues that Benatar's asymmetry is incoherent. Campbell Brown also argues that the asymmetry is incoherent, but in a different way.

From this short description it should already be clear that none of these criticisms has much to do with the counterintuitiveness of Benatar's conclusion. All Benatar's critics seem to share his intuition-free methodology, which is hard to explain under the dualism hypothesis. However, towards the end of the article, Benatar also mentions "smug, dismissive, and often vituperative responses, many of which attack only the conclusions and not the arguments" (Benatar 2013, p. 150). Perhaps what he refers to are replies coming from the intuition-based camp? We are only provided with one example: an article by Christopher Cowley, who calls Benatar's book "the work of a crack-pot", suggests that philosophers should not seriously engage with it, and accuses Benatar of corrupting the youth. But it is hard to find traces of DE in this response. What seems to be Cowley's most substantive objection is that something can only be deemed harmful "within the context of an on-going human life, with all the background meanings that characterise that human life" (Cowley 2011, p. 24). This has clearly nothing to do with accusing Benatar of failing to account for the intuition that coming into existence is not always a harm. The only time when intuitions are brought up is when Cowley writes about Benatar's argument's destructive effect on

young people's moral intuitions, which are already "in a state of flux". It is obvious that this point concerns practical consequences of spreading Benatar's views, not the method of establishing them. The reason why Benatar decided not to engage with Cowley's criticism is therefore clearly not because it relied on an alien, DE-based methodology. Rather, it is exactly what Benatar claims it is: the criticism does not seriously address any of his arguments. So we are left with the conclusion that neither Benatar nor any of his critics treats intuitions as evidence of their contents in the context of justification.

4. Taurek on whether the numbers count

In the previous chapter I quoted Peter Singer, who argues that even if his work relies on intuitions, these must be general and abstract intuitions, like the intuition that five deaths are worse than one death, other things being equal (Singer 2005, p. 350). Many people find this claim self-evident. To oppose it would be to to oppose something fundamental about morality, or perhaps morality itself. And yet there are philosophers who do oppose it – without endorsing moral nihilism. John Taurek in his famous article *Should the numbers count?* argues that no number of deaths can be said to be worse than one death, and that someone who faces a choice between saving more lives and saving fewer lives, and wishes to act impartially, should toss a coin to make the decision.

4.1 Taurek's argument

Unlike Tooley or Benatar, Taurek does not make metaphilosophical comments about the role of intuitions in philosophy or ethics. Let us then go straight to the argument. Taurek starts off by describing the following scenario:

> The situation is that I have a supply of some life-saving drug. Six people will all certainly die if they are not treated with the drug. But one of the six requires all of the drug if he is to survive. Each of the other five requires only one-fifth of the drug. What ought I to do? (Taurek 1977, p. 294)

It might seem obvious that I ought to save the five *special considerations apart*. But then I am asked to imagine that the person who needs all of the drug, named David, is someone I know and

like, while the other five are complete strangers. It is also stipulated that I have no *duty* to save David's life. It seems that while I may have a reason to give the drug to David – I like him and care about him, after all – this reason does not count as a special consideration which can override the obligation to save the five. This, argues Taurek, shows there is simply no such obligation. "It is the absence of any moral requirement to save these others rather than David that makes my doing so morally permissible" (ibid., p. 297).

This is how some people see it. However others would argue that giving the drug to David would be wrong, which only proves that five deaths are worse than one death, all else being equal. To explain why this is not the case, Taurek asks them to imagine that David himself is the owner of the drug. Does he have any conclusive reasons not to save his own life? The answer is: he does not. This is because the relation between David and his own life is fundamentally different from the relation between David and each of the other five lives.

> He values his own life more than he values any of theirs. This is, of course, not to say that he thinks he is more valuable, period, than any one of them, or than all of them taken together. (Whatever could such a remark mean?) (ibid., p. 300)

Taurek then argues that if it is permissible for David to spare his own life, it must be permissible for a third party, who owns the drug, to spare David's life. If good moral reasons to save the five are absent for David, then they must be absent for any other agent.

The central argument of the article is not made fully explicit, but I think it can be summarised as follows. Things are good or bad only because they are good or bad *for someone.* David's death is bad because it is bad for David, and each of the other five's death is bad because it is bad for that person. But there is no person for whom the sum of five deaths is bad. For this reason the sum of five deaths cannot have any moral significance, let alone a significance greater than that of one death. Those who believe that five deaths are worse than one death are therefore adding up what cannot be added, and comparing what cannot be compared. Furthermore, it cannot be said that more deaths are worse because they translate into more overall suffering, or a greater overall loss of future happiness. This is because what goes for the badness of death also goes for the badness of any other bad thing, like pain or suffering. "The discomfort of each of a large number of individuals experiencing a minor headache does not add up to anyone's experiencing a migraine" (ibid., p.

308). The collective suffering (if this expression has any meaning at all) cannot be something that matters morally, since there is nobody to suffer it.
In his reply – written in the late 1970s, but only published recently – to Derek Parfit's criticisms, Taurek clarifies that he is not even ready to allow that expressions like "collective suffering" or "the sum of several people's suffering" refer to anything tangible. Suffering, or happiness, is like physical attractiveness or boxing skill – it can only be meaningfully ascribed to an individual.

> If I call your attention to the great beauty in this woman's face, and having acknowledged it, you reply 'But her beauty pales when compared to the awesome beauty we contemplate when we add together or sum the beauty found in each of a sea of ordinary faces,' I will not know what to make of this. (Taurek 2020, p. 313)

Similarly, Taurek does not know what to make of the claim "the sum of many minor pains of different persons can be greater than a very intense pain of one person" – even if the badness of pain is set aside.
What about the situation in which all six people who need the drug are strangers to the owner of the drug? It is not entirely clear if Taurek thinks *any* decision would be equally acceptable. Nevertheless he points out that personally he would be in favour of tossing a coin, which "best express[es] [his] equal concern and respect for each person", as it gives each person equal chance of surviving (Taurek 1977, p. 303).

To many people the suggestion that preventing more deaths is not morally better than preventing less deaths, irrespective of the actual numbers, seems preposterous. However it might become more plausible when we realise that in many concrete situations one still has a duty to prevent more deaths rather than less deaths. There is no paradox here. Taurek illustrates this point with the following scenario:

> Volcanic eruptions have placed the lives of many in immediate jeopardy. A large number is gathered at the north end of the island, awaiting evacuation. A handful find themselves on the southern tip. Imagine the captain of the only Coast Guard evacuation ship in the area finding himself midway between. Where shall he head first? Having been persuaded by my argument, to the amazement of his crew and fellow officers, the consternation of the government, and the subsequent outrage in the press, he flips a coin and makes for the south. (ibid., p. 310)

The captain might be misapplying Taurek's principle. He might actually be obliged to head north. We should note that the difference between him and the agent in the drug scenario is that the captain

does not *own* the ship. Rather, he is at his community's service and has to carry out whatever the community has agreed on – which may well be saving the larger number:

> A number of people have joined to invest in a resource, the chief purpose of which is to serve the interests of those who have invested. Whether each has invested an absolutely equal amount, or whether individual investments are scaled to individual resources, is neither here not there. Theoretically at least, each person's investment (or status) is seen as entitling him to an equal share, an equal claim on the use of that resource or on the benefits form its use. Now a policy for the employment of that resource in just such contingencies as this present trade-off situation must be adopted. And it must be a policy agreeable in advance to all those who are supposed to see their interests as equally served. The captain's duty, then, whatever it is, is seen as deriving form this agreement. (ibid., p. 312)

This example might to some extent alleviate the outrage: the practical implications of the view that the numbers do not count are not entirely different from those of the opposite view. But, to be sure, some of them are different. For example, it is hard to see how a captain of a private boat with no special relation to the islanders could have a duty to save the larger number.

4.2 Judgments about cases

Does Taurek appeal to intuitions in a DE-sense? Again, let us start by focusing on judgments about thought experiments. First, we have the three versions of the drug scenario: one in which you are deciding whether to save one stranger or five strangers, one in which you are deciding whether to save David or five strangers, and one in which David is deciding whether to save himself or five strangers. In the first case the response advocated by Taurek – toss a coin – is quite counterintuitive. This poses a challenge to proponents of DE. Is Taurek trying to undermine his own case? As he duly notes, "to many it seems obvious that in such cases, special considerations apart, one ought to save the larger number" (ibid., p. 294). But there is no indication that anyone could take this fact to be evidence that this is what in fact one ought to do. Taurek simply proceeds to offer an argument against the claim, without explaining how the argument is meant to weaken or surpass the evidential force of the claim's intuitiveness. This is most likely because neither Taurek nor anyone else assumes that any such force exists.

The same applies to verdicts about the two other versions of the scenario. As for whether it is morally permissible to save David, Taurek notices that opinions are *split*. But he is not interested in

finding out how many people support which opinion, or which opinion can be said to more intuitive in one sense or another. He simply gives reasons to think that by saving David one does not do anything wrong. The same goes for the last case, where David decides to save himself and *justifies* his choice: he says his life is far more valuable to him than the lives of the five strangers, and that each them is, of course, in the same position. This means nobody should expect him to give up his life, just like nobody should expect any of the strangers to give up theirs, were they to decide.

In short, all three judgments about different versions of the drug scenario are supported with evidence. It would be extremely far-fetched to argue that by providing this evidence Taurek engages in some sort of abductive reasoning which starts with the judgments serving as data assumed to be supported with their own intuitiveness. For one thing, the first two judgments are not even intuitive in any plausible sense of the word. Whether the third one can be described as such is debatable. But even if it can, Taurek never brings up the question of intuitiveness. Nor does he appeal to any explanatory virtues or mention alternative explanations when he spells out why David is justified in choosing to save his own life. Moreover, Taurek acknowledges that not everyone agrees with the justification – he writes that according to "the usual sort of utilitarian reasoning" David should sacrifice his life to produce more overall happiness (ibid., pp. 299-300). If DE were true, we would expect Taurek to counter this reasoning by pointing out how intuitive it is that David cannot be required to make the sacrifice. However Taurek says nothing of this kind. The only evidence against the utilitarian conclusion is the problem with the idea of "overall happiness", nothing turns on whether the conclusion is intuitive.

How about the boat scenario? Here Taurek points out that while it seems obvious that the Coast Guard captain ought to save the greater number, what he actually ought to do depends on the circumstances not specified in the original scenario. It is plausible that there is a contract which binds him to save more people, but it is still not impossible that he should toss a coin. Once again the correctness of the answer is entirely independent of its intuitiveness. All that matters is what the people involved have agreed upon.

4.3 Denial of impersonal goodness

I have argued that Taurek's argument can be interpreted as relying on the premise that anything that is good must be good *for someone*. Some proponents of DE might be tempted to argue that this is

the actual intuition used as evidence in the article: Taurek tacitly treats the premise as supported by the fact it seems true, or is intuitive in some other sense. He does not make this inference explicit, but it is only because it is too obvious to be stated.

The best way to examine this hypothesis is to find out whether any philosophers challenge or defend Taurek's premise, and whether they ever invoke its intuitiveness in the process. As it turns out, Christine Korsgaard has recently argued in favour of the claim (Korsgaard 2014). She starts her article by pointing out that it is not difficult to find apparent counterexamples to it. For example, we tend to think that the world full of happy creatures is better than the world full of miserable ones, *even if the creatures are different in the two cases*, and also better than the world with not inhabitants at all. This might look like a clash of intuitions. On the one hand, it seems to us that whatever is good must be good for someone. One the other, it seems to us that the happy world is a good thing. But if it is in fact good, it cannot always be good for someone in particular. How can the conflict be solved? If DE were true, we would expect Korsgaard to compare the two propositions in terms of how intuitive they are, and to offer criteria for weighing their intuitiveness against each other and against any other relevant kind of evidence. Or, at the very least, we would expect her to mention philosophers who give some importance to the intuitiveness of these propositions.

But this is clearly not Korsgaard's approach. She argues that if impersonal goods exist, then having a good must be understood as a special relation between an individual and a good which, absent the special relation, would be free-floating or belong to someone else. Some believe that this special relation can be understood in terms of appreciation, enjoyment or ownership of the good. But Korsgaard finds all these proposals wanting. The upshot is "not merely that everything that is good must be someone's good: it is that everything that is good must be related to someone in a particular way before it can really *be* something good at all" (ibid., pp. 411-12). This does not mean that we are wrong to think the universe full of happiness is good, we are only wrong to think it is good impersonally. Korsgaard suggests that while thinking about the problem we typically imagine ourselves as creators facing the choice between bringing about this or that particular universe. She then argues that if we were in this position, we would have a duty to do as well as possible *for whomever we create* (ibid., p. 426).

In the end Korsgaaard may be said to have vindicated both seemingly contradictory intuitions. However she clearly has not done so in any DE-friendly sense. Neither of the claims is used as a starting point for abductive reasoning – this is for the same reasons none of Taurek's judgments

about cases is used in this fashion in his article. And even if they were, there are clearly not meant to be supported by their own intuitiveness. Nothing in Korsgaard's argument depends on how intuitive it is that whatever is good must be good for someone, or how intuitive it is that the world full of happiness is better than the world full of misery. Moreover, this seems typical. I am aware of no philosophical argument which uses any of these facts as a premise. And even if such DE-based argument exists somewhere in the literature, its existence would only prove that DE is a rather niche philosophical methodology, not that Taurek adopts it in his article.

Taurek's principle can also be challenged in a more radical way: one might argue that no goods are, strictly speaking, personal. On this view the statement "anything that is good must be good for someone" would be false because there is no such thing as "someone", or because "someone" is not the kind of thing that can have a good. This may sound odd, however the idea that the self is an illusion has been defended by a number prominent thinkers over centuries – from the Buddha and Nāgārjuna, through Hume, to Parfit.
Let us focus on Parfit's argument, which seems most elaborate. It starts by attacking "non-reductionism": the view that there exists some sort of "Cartesian Ego" or a soul which unifies one's experiences while being ontologically separate from these experiences, as well as from the body where it resides. Parfit argues that non-reductionism lacks any evidential basis. First, some people believe in reincarnation, and what reincarnates may well be a non-reducible ego, however we have no memories of our past lives that would make this view plausible. Secondly, we know that brain damage and mental illness does not affect personality in an all-or-nothing way, which is what could have been the case had egos existed separately from brains. Finally, because the separate ego must be an all-or-nothing phenomenon, non-reductionism seems to lead to absurd conclusions. Imagine Derek Parfit being gradually transformed into an exact replica of Greta Garbo at 30 by having his brain and body cells replaced one by one. It is implausible that Parfit and the final replica are the same person. It is also implausible that Parfit suddenly dies at one stage of the process by losing just one or a few of his cells. However these are the only two options that non-reductionism offers.

If non-reductionism is rejected, we are left with mere facts about series of events, mental and physical, which "can be described without either presupposing the identity of [a] person, or explicitly claiming that the experiences in this person's life are had by this person, or even explicitly claiming that this person exists" (Parfit 1987, p. 210). Reductionists do not have to hold that there are no persons. However, on their view a person is much more fuzzy an entity than it is commonly

assumed. The question of whether someone died often becomes what Parfit calls an "empty question" – both "yes" and "no" can be consistent with what happened.

A reductionist can adopt a psychological criterion of identity, according to which one's survival is a matter of preserving one's mental states, like beliefs, memories, preferences, interests, plans, character traits etc. They may also be drawn to a physical criterion, according to which it is a matter of preserving a particular biological organism. Or they might go for some combination of the two. However on any of these accounts, persons can gradually cease to exist and no determinate cut-off point can be found.

Moreover, on the common sense view personal identity is a *numerical* kind of identity: one person at a specific point in time can only be identical with exactly one person at a different point in time. We also tend to think it should be decidable whether this relation holds between any two entities. However if personal identity can be reduced to a bundle of mental states, this bundle can be *multiplied*, for example by destroying someone's body and creating several exact copies of it. The same goes for physical substance: a brain can be split and its hemispheres can be transplanted into two separate bodies. There is no one correct answer to the question of whether one would continue to exist as all of the resulting branches, just one of them, or perhaps none.

Reductionists must satisfy themselves with the possibility of adopting different conventions about what counts as death. For example, I can stipulate that I continue to exist only if I do not divide – otherwise I die and a number of new persons are created. But surely this kind of death is very unlike ordinary death. There are still beings who inherit my mental life: they perform what I intended to do, they know what I learnt, etc. This consideration leads Parfit to think it is not personal identity, but rather a kind of psychological connectedness and continuity – what he calls "Relation R" – that morally matters. In consequence, it "becomes more plausible, when thinking morally, to focus less upon the person, the subject of experiences, and instead to focus more upon the experiences themselves" (ibid., p. 341). One might object that each of these experiences is still not free-floating, but always *had* by someone – at the very least some momentary subject of experience – and therefore always good or bad for someone. However, even if this is correct, this kind of ephemeral subject cannot count as a person, and cannot correspond to "someone" in Taurek's premise. For example, it would make little sense to say David's death is bad only because it is bad for David, if David turned out to be a series of different short-lived entities.

Needless to say, not everyone has been convinced by Parfit's attempt to dissociate the good from its subject. A common objection is that if Relation R is what morally matters, it does not follow that we

should “focus more upon the experiences themselves”. Rather, we should focus more on the chains of psychological connectedness that these experiences belong to. Diane Jeske points out that Parfit’s version of reductionism supports the commonsense belief we have special obligations to our intimates – as they are psychologically connected to us to a greater degree than strangers (Jeske 1993). However Parfit himself believes that his view is at least consistent with some sort of agent-neutral, impersonal, utilitarian ethics, in which every experience has an equal status. If Jeske is correct, then perhaps Parfit’s move from personal identity to Relation R does not offer us grounds to undermine Taurek’s argument. Maybe the premise “anything that is good must be good for someone” can still be true, if interpreted as “anything that is good must be good relative to a chain of psychological connectedness”, and then claims about the badness of death are true if “death” is interpreted as “end of a chain of psychological connectedness”.

Another response has been offered by Korsgaard, who argues that even if “there is no deep sense in which I am identical to the subject of experiences who will occupy my body in the future (...) I nevertheless have reasons for regarding myself as the same rational agent as the one who will occupy my body in the future. These reasons are not metaphysical, but practical” (Korsgaard 1996, p. 369). First, I must act, and I have only one body to act with. Second, I must deliberate and choose, which has to be done from a certain *standpoint*. Third, choosing any action must carry me into the future. For these reasons I cannot help but regard myself as a unified agent – and this kind of unity is what “someone” in Taurek’s premise refers to. Some might object that being an agent is merely an *illusion*, as it is shown by Parfit’s argument against non-reductionism, or perhaps by arguments against free will, or some other metaphysical considerations: it only seems to me that I persist over time as a separate entity, or that I am the author of my doings. But Korsgaard insists that the agent-centred picture is always a legitimate description of reality *from a practical perspective*. There is also a theoretical perspective, to which the aforementioned metaphysical arguments belong. These two are not necessarily in harmony with each other. However “the incongruity need not become contradiction, so long as we keep in mind that the two views of ourselves spring from two different relations in which we stand to our actions” (ibid., p. 378). If we want to explain and predict, we need the latter. If we want to choose and justify, we need the former. And ethics is primarily about choosing and justifying actions, which means there is no escape from relating the good to separately existing agents.

How about Parfit’s, Jeske’s and Korsgaard’s evidence? From what I have said so far, it should be clear that none of them makes much of the *intuitiveness* of impersonal goodness, or any other idea they attack or defend. Parfit writes he may never be able to completely erase his “intuitive belief” in

non-reductionism (Parfit 1987, p. 280), but this is merely a psychological remark. He is not trying to weigh the fact that his belief is intuitive against his reasons to reject the belief. Rather, he thinks that his reasons compel him to fully embrace the reductionist view with all its implications, and his intuition-based reluctance to do so is simply irrational. Similarly, Jeske is not stating or implying that the intuitive view about our moral obligations to friends or family members is plausible because it is intuitive. She believes it is plausible because it is supported by the claim about the importance of psychological connectedness, which in turn is well-supported by evidence provided by Parfit. Finally, Korsgaard's point is not that her pragmatic account of agency and its moral relevance is more intuitive than the alternative views, nor that her arguments in favour of the pragmatic agency trump the intuitiveness of any of the alternative views. Her point is that that agents can be said to exist in the sense that is relevant for ethics.

In sum, philosophers examine all sorts of reasons to accept the claim that anything that is good must be good for someone, but the fact that the claim is intuitive is simply not among them. This undercuts the hypothesis of tacit DE, which requires not only that this fact must be taken into consideration, but that there is a consensus over its evidential status. It would be absurd to assume that Taurek was somehow oblivious to the fact there was no such consensus when he was writing his article.

4.4 Semantic defectiveness thesis

Another general claim at the core of Taurek's argument is that utterances about aggregated happiness or aggregated suffering are somehow semantically defective. Unfortunately it is not easy to pinpoint what exactly this defectiveness amounts to. According to Taurek when we compare two sentences like "Individual A suffers more pain than individual B" and "Individual A and individual B taken together suffer more pain than individual C", the phrase "more pain" undergoes a change in meaning. Taurek says he understands the meaning of "more pain" in the first statement, but he does not understand its meaning in the second. He also repeatedly calls whatever the phrase in the second statement refers to a "metaphysical fiction" (Taurek 2020, pp. 313-4).
One way to interpret this complaint would be to argue that the sentence "A and B taken together suffer more pain than C", while syntactically well-formed, is *meaningless*, akin to Carnap's "This stone is thinking about Vienna", Russell's "Quadruplicity drinks procrastination", or Chomsky's "Colourless green ideas sleep furiously". Suppose this is the correct reading. According to DE, the

fact that it is intuitive that the sentence is meaningless is treated as evidence that it is in fact meaningless. Admittedly Taurek does make this inference explicit, but this is only because he assumes it is unnecessary, as his readers take it for granted.

The main problem with this view is that the sentence "A and B taken together suffer more pain than C" does not seem meaningless in the way that "Quadruplicity drinks procrastination" does. Many people find it perfectly meaningful. To this proponents of DE might reply that according to Taurek the only reason why the sentence seems meaningful to many is that it is similar to some actually meaningful sentences, like "A suffers more pain than B". When we notice that "more pain" in the meaningful sentence is a different "more pain", suddenly our sentence does not seem meaningful to us any more. This hypothesis may be consistent with the expertise version of DE I described in the previous chapter: Taurek is not relying on the widely shared intuition, but only on his own superior, expert intuition.

This, however, is deeply problematic. On this account Taurek could have stayed silent about the inference he was making only if he had believed his readers were able to easily, perhaps non-consciously, recognise it. But of course for many of them it would have been impossible to recognise it as they did not even accept the premise. A reply might be that Taurek is addressing his paper only to experts who share his expert intuition. But in this case another question arises: who are they and what reasons do we have to believe they are the actual target audience? Surely they are not academic philosophers in general – Taurek was perfectly aware that many utilitarians took the sentence to be meaningful when his was writing his original paper. His later clarifications were made in response to Parfit, who, according to Taurek, equivocated between different senses of "more pain". If DE were true, at this point Taurek would have responded by writing something like "after considering all relevant data, one forms an intuition that of the sentences is meaningless, which indicates that it is indeed meaningless". But this is not what he did.

There is also a more general problem with the meaninglessness interpretation: philosophical debates about this issue never appear to involve appealing to intuition in any DE-friendly sense. For example, Quine argues that sentences like "Quadruplicity drinks procrastination" are not in fact meaningless. Rather, they are simply false. He points out that the urge to label them as meaningless can stem from a "spontaneous revulsion against silly sentences" (Quine 1960/2013, p. 210). But he does not treat this revulsion as evidence against his view that has to be somehow dealt with. According to Quine, those who dismiss these sentences as meaningless are still likely to accept mathematical falsehoods as meaningful, and the task of accounting for this difference is more challenging than one might expect. It appears that any satisfactorily parsimonious theory would

likely classify both kinds of sentences as false. The question of intuitiveness of such classification never enters the discussion.

Some argue that meaninglessness is different from contentlessness, and that the two should not be conflated with each other (Magidor 2022). Drawing on this idea one can interpret Taurek as saying that the sentence "A and B taken together suffer more pain than C" has a meaning, but does not have a content. That is, it does not express a proposition. In certain contexts the sentence does express a proposition – which is what makes it meaningful – but *these are not the contexts in which it is normally used*, as A and B are taken to refer to humans, or perhaps some non-human animals, in the actual world. If A and B referred to some hypothetical creatures who were capable of merging their individual pains and co-experiencing them, the sentence would have a content. We might also suppose that A and B coexist with some meta-experiencer, like a being who suffers all the pains suffered by individual people. William MacAskill in his recent book asks the reader to imagine living "through the life of every human being who has ever lived" (MacAskill 2022, p. 1). Perhaps this kind of existence, assuming it is metaphysically possible, could give some propositional content to the sentence. One might also think of Hindu beliefs about the all-encompassing "world soul". The problem is, of course, that the proposition discussed by Taurek is not about pain experienced by hypothetical or dubious creatures. It is about flesh and blood human beings. The contentlessness reading may therefore allow us to make more sense of Taurek's words about a "metaphysical fiction".

However this view changes nothing in terms of the plausibility of DE. All problems with the meaninglessness interpretation apply to the contentlessness interpretation with the same, if not greater, force. It is far from clear whether it is intuitive, in any sense and to anyone, that the sentence in question has no content. But even if it is – for example according to someone's expert intuition – Taurek could not have simply assumed any consensus over this fact.

To sum up: it is undeniable that at least some of Taurek's argumentative starting points are, in some sense, intuitive. But this does not mean that their intuitiveness is tacitly understood to support them. In all likelihood, they are placed in the common ground – Taurek decided that they were plausible enough to his readers, for whatever reasons. As I have tried to show, these reasons vary. What is invoked to defend Taurek's starting premises largely depends on how the premises happen to be attacked. It is also not impossible that a number of readers just accept the premises unreflectively, and would struggle to explain why if challenged. However the fact they would struggle, or any

other characteristic of the intuitive, is never treated as a reason to accept the content of the premises, at least not in any kind of philosophical setting.

4.5 Reception

None of the varieties of DE which require that Taurek's argument starts with something meant to be supported by its own intuitiveness can therefore be upheld. One can also safely rule out the cognitive bias variety, which implies that the counterintuitiveness of Taurek's conclusion is dismissed as a product of a faulty cognitive process. By discussing the Coast Guard case Taurek offers some explanation as to why it wrongly seems to us that we ought to save the greater number all else equal, but the explanation itself clearly does not function as a reason to reject the claim. This leaves us with the dualism hypothesis, or, to be more precise, with the opaque version of the dualism hypothesis. Perhaps Taurek ignores our moral intuitions while other ethicists respect them, but without realising that their methodology is not universally shared. If this is true, we would expect Taurek to be accused of failing to see how intuitive it is that more deaths are worse than less deaths, or how the intuitiveness of whatever he relies on does not outweigh the counterintuitiveness of his conclusions.

Let us then have a closer look at the major critiques of Taurek's article. Gregory Kavka offers three interrelated objections. First, one can hold that giving the drug to David is morally permissible and still reject the claim that the numbers do not count. Secondly, if we should be indifferent between saving one stranger and saving five strangers, then we should also be indifferent between saving four of the five strangers and saving all of the five strangers, which is unacceptable. Third, Taurek fails to explain how the numbers matter in rational prudence while they do not matter in morality (Kavka 1979). None of this looks promising to the advocate of the dualism hypothesis. Perhaps she could argue that Kavka is listing one more counterintuitive implication of Taurek's view – wasting some of the drug and saving less than five people is morally equivalent to not wasting any and saving all five – to show that this implication tips the intuition scales, so to speak, in favour of the claim that the numbers should count. But it is clear that Kavka is not interested in making any comparisons of this kind. He never mentions the intuitiveness of what Taurek accepts, or the intuitiveness of what Taurek rejects, or any criteria to measure the two. In all likelihood Kavka is placing the claim that wasting some of the drug would be worse *in the common ground*: "none of us [...] is likely to find this implication of Taurek's view acceptable" (ibid., p. 292).

Derek Parfit in his aforementioned reply identifies a number of controversial assumptions that Taurek's argument rests on. For example, Taurek assumes the agent-neutrality of David's permission to prioritise his own life. To Parfit it is more plausible that the permission is agent-relative: while it is not wrong for David to keep the drug for himself, it would be wrong for the third party to give the drug to David and let the five die (Parfit 1978, p. 291). But Parfit is not saying that agent-relativity is more plausible than agent-neutrality because it is more intuitive. He is saying that unless Taurek offers us an argument for agent-neutrality, we have no good reason to favour it over agent-relativity. Ultimately, which view is correct depends on the quality of evidence for each view, and there is no indication that facts about the intuitiveness of a view can be part of this evidence. The same can be said about each of the other assumptions listed by Parfit.

According to John Sanders, Taurek's "principle of equal concern must rule out any thought that persons are worth saving because they are persons, or that human life is valuable or worth saving in and of itself" (Sanders 1987, p. 13). However such thoughts should not be easily ruled out. Sanders's argument can be summarised as follows. First, an object can be valuable instrumentally, but also valuable in its own right. Second, persons are always objects of the former kind. Third, whenever a loss of an object is at stake and that object is valuable, the numbers should count. Therefore, in Taurek's scenario the numbers should count and it is obligatory to save the greater number of people, other things equal. Granted, there is plenty of intuition-talk in Sanders's article. For example, he describes the disagreement between himself and Taurek as "the war of intuitions". Or he points out that he "cannot help feeling that the world is a better place with people in it than it would be without them" (ibid., p. 12), which, he thinks, supports his second premise. But what he is expressing here is not an inference from "I feel that *p*" to *p*. Rather, just like Parfit he is saying he is inclined to accept the view, while Taurek is inclined to reject it, and neither of them can offer conclusive evidence for or against. Taurek might still be correct, but one of his assumptions needs to be substantiated, as it is far from obvious.

Several philosophers have argued that by trying to salvage the value of the individual Taurek somewhat paradoxically neglects the value of the individual. For example, Frances Kamm writes:

> If we (...) toss a coin between one person and any number on the other side, giving each person an equal chance, we would behave no differently than if it were a contest between one and one. If the presence of each additional person would make no difference, when this affects their good, this seems to deny the equal significance of each person. (Kamm 2007, p. 33)

A similar point has been made by T. M. Scanlon, who thinks it is unacceptable that "the presence of the additional person (...) makes no difference to what the agent is required to do" (Scanlon 1998, p. 232), and by Jens Timmermann, who complains that "for Taurek, one person counts for one but two or five or fifty million equally count for one" (Timmermann 2004, p. 110). Kamm and Scanlon believe that even if the idea of aggregating the good is dismissed as incoherent, one still have an obligation to save the larger number. Timmermann, following Michel Otsuka's argument, believes that Kamm and Scanlon are wrong as they still covertly aggregate the good (Otsuka 2000, Otsuka 2013). He then proposes a third solution: each individual should be given a $1/n$ chance of being prioritised, where n is the number of individuals affected by the decision. In Taurek's scenario, instead of tossing a coin, one should *roll a dice*: there would be a 5/6 chance of selecting a person who only needs one fifth of the drug to survive. After rescuing the selected person, one would either have to accept that everyone else perishes or, more likely, face a choice between saving four additional people and saving nobody, which is hardly a moral dilemma. This way no one can complain about not being treated as equally significant.

None of these criticisms have much to do with using the counterintuitiveness of the claim that the numbers do not count as evidence against the claim. The only reason to reject it is that Taurek relies on a defective account of impartiality: the moral worth of each individual cannot be described as equal. It is not unlikely that Kamm, Scanlon and Timmermann had engaged in relying on intuition as evidence in the context of discovery. For example, Timmermann writes that his solution "pays tribute to our unreflective feeling that the greater number ought to be saved" (2004, p. 112). However the feeling does not itself constitute a premise in his argument. Even if the feeling did not exist, Timmermann's solution would be exactly the same, and for the same reasons.

In short, no reactions to Taurek's article reveal any DE-based assumptions. Multiple alleged flaws have been spotted, but none of them is that the conclusion is counterintuitive. Just like Taurek appears to adopt the DE-free methodology, so does everyone else.

CHAPTER 5. Experimental philosophy

1. The dogma and the birth of x-phi

The recent pushback against the idea that philosophers appeal to intuitions has been largely a reaction to experimental philosophy (x-phi), which I have touched upon at various points of this thesis, arguing it constitutes an exception to my claim that DE is false. The movement is fairly young – what is often cited as the founding paper was only published in 2001 (Weinberg et al. 2001). This is curious, as DE has been around for much longer, at least since 1970s. In 1990s it was fleshed out and defended by people like George Bealer, Alvin Goldman, Joel Pust or Ernest Sosa. Why did we have to wait so long for DE to be attacked? One answer to this question is that up until the birth of x-phi DE had been a largely harmless metaphilosophical misconception. Experimental philosophers have changed that – their DE-based project can be seen as not only pointless, but actively harmful.

2. The pointlessness of x-phi

Let us begin with the pointlessness. X-phi often defines itself in opposition to mainstream analytic philosophy, which it calls "armchair philosophy". The central flaw of armchair philosophy is, we are told, being too unempirical, a priori, conceptual, introspective or intuition-based in nature. The goal of x-phi is to overcome this flaw by making philosophy more empirical. However, as Deutsch points out, this picture is heavily skewed:

> It is now common to hear the dispute between xphiles and those who defend analytic philosophy and its methods characterized as one over whether philosophy can be pursued "from the armchair." (…) the characterization of the dispute in terms of those for or against armchair philosophy is misleading in at least two ways. First, the characterization wrongly suggests that armchair philosophy is unscientific, or unconcerned with empirical results related to its subject matter. Second, the characterization unfairly casts xphi as a curative—a pro-science balm designed to counteract the tendency to simply sit in an armchair and think.

> Since typical survey-style xphi methods are clearly empirical, casting xphiles as opposed to armchair philosophy suggests that armchair methods are *not* empirical. But this is not true. By definition, a priori methods are not empirical. But sitting in an armchair does not prevent one from appealing to things one has learned a posteriori. (Deutsch 2015, p. 157)

Deutsch is correct: armchair philosophy is chock-full of unquestionably a posteriori claims. Some of them can be easily found in the case studies from the previous chapter. For example, Tooley's argument for the permissibility of infanticide relies on the premise that infants lack the concept of a continuing self. This is clearly not something knowable a priori, and no philosopher seems to treat it as such. Were empirical science to discover that infants do, after all, possess the concept of a continuing self, it would be hard to imagine Tooley digging in his heels and insisting they do not, as this is what he had established via a priori reasoning.

Experimental philosophers might concede this point, but still maintain there is some kind of a priori core armchair philosophy, which is based solely on consulting intuitions. Gettier's argument against the justified true belief account of knowledge is often cited as a prime example of this practice. However, as I have tried to show in this thesis, even if there exists something worth being called "a priori philosophy", it has nothing to do with consulting intuitions in any DE-friendly sense. And the problem with much x-phi is that it only makes sense under this assumption. For instance, experimental philosophers run questionnaire studies to check whether the Gettier judgment can be influenced by cultural background, or framing of the scenario. What motivates them is the belief that Gettier assumed, from his armchair, that most people would find "Smith does not know" intuitive – or perhaps he tried to impose this intuition on his readers – in order to justify the claim that Smith does not know (Weinberg et al. 2001, p. 434; Turri 2016, p. 339; Fischer & Collins 2015, p. 11). But this is simply not the case: facts about the intuitiveness of "Smith does not know" play no justificatory role in Gettier's argument. The same seems true of all other cases that have been tested.

As Joshua Alexander, one of the leading advocates of x-phi, rightly points out: "if it turned out that our intuitions weren't philosophically significant, then experimental philosophy would be left to occupy the unhappy position of taking seriously a way of thinking about philosophy not worthy of serious consideration in the first place – it would be philosophically insignificant" (Alexander 2012, p. 94). In my view, this is exactly what turns out. X-phi mistakenly assumes that armchair philosophy is based on intuitions, notices that these intuitions can be unstable or parochial, and tries to fix the non-existent problem by producing a great deal of philosophically irrelevant data.

3. X-phi and conceptual analysis

Here some experimental philosophers resort to the conceptual analysis defence. They argue that even if their research is unhelpful in terms of finding out what knowledge is, it is still helpful in terms of finding out what *the concept of knowledge* is – and the same goes for many other philosophically interesting concepts. For example, cross-cultural divergence in Gettier intuitions can be explained by the fact that different cultures use different concepts of knowledge, and cross-cultural convergence in Gettier intuitions can be explained by the fact that different cultures use the same, or perhaps a similar, concept of knowledge. This kind of research may or may not have bearing on the problem Gettier tried to solve – namely whether knowledge is justified true belief – but it is still philosophically valuable.

However this reply runs into the problem of circularity I described in chapter 3. To recap: if one wants to investigate the nature of knowledge, justice, reference etc. by studying our intuitions about what counts as knowledge, justice, reference etc., one needs to separate accurate intuitions from inaccurate ones. But this is only possible if one first assumes what knowledge, justice, reference etc. is. Studying intuitions is therefore a waste of time. The same applies to the project of conceptual analysis. If one wants to investigate *the concept of* knowledge, justice, reference etc. by studying our intuitions about what counts as knowledge, justice, reference etc., one needs to separate *conceptually* accurate intuitions from *conceptually* inaccurate ones. But this is only possible if one first assumes what *the concept of* knowledge, justice, reference etc. is. Studying intuitions is therefore a waste of time.

How can experimental philosophers respond to this challenge? One answer might be that if there are *multiple* concepts of x, then x-phi can at least reveal who possesses which concept of x. However the way experimental philosophers approach this issue often seems too casual – they tend to jump to conclusions without an adequate theoretical framework. For example, Machery et al. report that roughly a third of Bengali-speaking Indians presented with a Gettier-type vignette deny knowledge, which "strongly suggest[s] that for many Bengali speaking participants, their concept of knowledge requires more than JTB" (Machery at al. 2015, p. 652). But this simply does not follow. Setting aside the problem of a semantic gap between English terms and their Bengali equivalents, it is possible that being justified is not a component of the participants' concept of knowledge – perhaps, as some epistemologists suggest, it should be replaced with being produced by a reliable cognitive process (Dretske 1981), or by being causally linked to what makes the belief true

(Goldman 1967). It is possible that their concept of knowledge is atomistic and cannot be analysed at all (Fodor 1998). It is possible that even if their concept of knowledge has a structure, this structure cannot be expressed by a set of necessary and sufficient conditions of falling under the concept, like "JTB" or "JTB+x" (Hampton 2006). It is possible that the category of a concept is explanatorily deficient and should be eliminated from the vocabulary of psychology (the view defended, curiously enough, by Machery himself in Machery 2009). It is possible that for certain pragmatic reasons some participants have understood the question less literally than others (Deutsch 2009). It is possible that the concept of knowledge only requires justified true belief, and some participants have somehow misapplied it. It is possible that the question of misapplication is ill-defined as there is no way of separating the conceptual from the factual (Williamson 2022). In short, Machery et al. owe us an account of what concepts are and how being disposed to make certain judgments is related to possessing certain concepts. Without it, it is hard to say what to make of the collected data.

Someone might argue that experimental philosophers can always develop and defend a theory of the intuition-concept relation before they start conducting surveys. This is of course true, but it would mean transforming x-phi largely into armchair philosophy, or perhaps cognitive science – in any case into something very different from what it is today. Moreover, the philosophical significance of the remaining experimental part would be far from clear. Suppose that Machery et al. provided us with a theory of concepts from which it follows that, for example, responding positively to a question about a particular Gettier-type scenario suggests that one's concept of knowledge is JTB, and responding negatively that one's concept of knowledge requires more than JTB. Then suppose it turns out that two thirds of Bengali-speaking Indians possess the former concept, while one third possess the latter. What philosophical problem could that finding elucidate? Should it be considered a philosophical discovery in its own right? Sure enough, it would *raise* a number of interesting philosophical questions: how can members of the same linguistic community acquire different concepts referred to by one word? How can they fail to realise this conceptual difference? How can they even communicate with each other? Answering these questions would definitely constitute a valuable contribution, however – again – it would require moving beyond the kind of research that is currently done.

In their more cautious moments experimental philosophers seem to leave the questions about concept possession open. They merely collect the data and sketch some interpretative options without committing themselves to any of them. However this can hardly vindicate the project. Virtually any body of data can have a number of interesting philosophical explanations. The basic condition for a meaningful experimental inquiry is to formulate a research question and explain how

it can be answered by specific empirical findings. If what experimental philosophers do amounts either to obtaining data with unknown implications and for unknown reasons or drawing unjustified conclusions from the data, this is clearly bad news for x-phi.

4. The harmfulness of x-phi

So far I have argued that it is hard to see how x-phi can be philosophically significant, at least in terms of analysing philosophically interesting concepts. Cappelen takes a step further by arguing that experimental philosophers also *spoil* philosophy:

> Not only does x-phi promote a false picture of philosophy, but, insofar as the programme is successful, it will *change* philosophy. Suppose I am right in my claim that philosophers don't rely on intuitions or anything that can be measured by responses to x-phi surveys. With the institutionalization of x-phi, interest in and reliance on surveys have gradually become integral parts of professional philosophy. So though x-phi was born in sin, so to speak, its institutional foothold has made it a truism that philosophers care about survey responses, since *they (i.e. experimental philosophers) do*. I agree with one of Stich's motivating thoughts: philosophy should not be based on the kinds of judgements people make when responding to surveys. That is an awful (borderline absurd) way to do philosophy. The institutionalization of x-phi has made it the case that many philosophers now think those kinds of judgements are important. *They have, in effect, created the practice they set out to undermine.* If that is right, then even x-phi lovers should agree that their influence is damaging. (Cappelen 2014a, p. 285)

It has been nine years since Cappelen wrote it and I am not sure if his worries have materialised. Experimental philosophy is still far from mainstream. It might have influenced some armchair philosophers' thinking about the relevance of survey responses, but it has not dramatically affected their first-order philosophical practice. The idea of appealing to intuitions was already widespread before the advent of experimental philosophy, but it had a very limited effect on actual philosophising. This situation might have changed for worse, however the effect does not seem overwhelming. Nevertheless, I agree with Cappelen on his epistemological point. So far in this thesis I have only argued that, experimental philosophy aside, DE is false, and refrained from opining whether it would be a good thing if it were true. But I am strongly inclined to agree that it would not. There is no reason to give more evidential weight to judgments that are snap, spontaneous or unreflective. If anything, we should value them *less*.

Granted, on a certain account of intuitiveness, certain intuitions are clearly trustworthy. However I am unaware of any viable proposals to filter out these intuitions for philosophical use. Experimental philosophers think their project constitutes one such proposal, but that seems mistaken. They sometimes insist that "philosophical inquiry has never been a popularity contest, and experimental philosophy is not about to turn it into one" (Knobe & Nichols 2008, p. 6). However if the majority criterion is to be rejected, what criterion would be both plausible and justify x-phi's empirical work? In the next section I am going to offer more reasons to think no satisfactory answer to this question can be given.

5. X-phi and overcoming biases

Cappelen believes that x-phi, in its entirety, is rubbish – there is nothing philosophically valuable about it. But this view is not common among intuition deniers. Jonathan Ichikawa, who rejects DE convinced by Cappelen's arguments, writes:

> It will come as no surprise to anyone that our philosophical capacities are fallible—sometimes we make mistakes—but many specific ways in which we are fallible could well turn out to be surprises. And at least in theory, these surprises could easily bring with them radical methodological consequences for armchair philosophy. Consider the various sorts of fallacies to which we humans are sometimes susceptible. The extent to which we are subject to these fallacies is an empirical question; so too is the question, under what circumstances are we better and worse at avoiding them. (Ichikawa 2014, p. 241)

If experimental philosophy can succeed in tracking down these errors, argues Ichikawa, then surely it is valuable – it can help philosophers do philosophy more responsibly. Joshua Knobe makes a similar point. In a private conversation he has recently used the following example: imagine a particular proposition, *p*, which plays an important role in philosophical argumentation (perhaps used as a starting premise). Now suppose that the belief that *p* has been randomly assigned to philosophers by a clever neuroscientist, as part of an experiment. Discovering this fact would clearly be a reason to become more sceptical about *p*. Of course x-phi does not test outlandish hypotheses like this one, but it does test many other hypotheses about philosophers' susceptibility to errors and biases. And exposing errors and biases in philosophical work is important irrespective of whether this work relies on intuitions in any DE-friendly sense.

This argument, I think, is based on a misunderstanding. The fact that a group of people believe something on flimsy grounds does not mean that there are no good grounds to believe it. Even if the only reason why philosophers believe that *p* is that their brains have been secretly tampered with by a neuroscientist who chose *p* at random, *p* can still be true. The neuromechanism used to cause the belief that *p* can therefore be a part of a highly reliable, truth-conducive cognitive process, and manipulated philosophers might not be making any *error* after all. The circularity problem looms large again: to determine whether we are dealing with a judgment error, we need to first find out whether the judgment in question is true, or at least more likely to be true than not. If we can find it out, then investigating why certain people believe certain things is redundant. If we cannot, then it is futile. The discovery described by Knobe is irrelevant to the evaluation of *p*, and consequently to any arguments relying on *p* as a premise.

Perhaps Knobe could modify his thought experiment and specify that the secret neuromechanism gives philosophers *inconsistent* beliefs: one day they believe that *p*, another day that not-*p,* based on random selection. This might somewhat resemble actual experimental studies, for example those suggesting that professional philosophers are subject to order effects and framing effects when they evaluate trolley scenarios (Schwitzgebel & Cushman 2015). Surely the discovery that philosophers' judgments are inconsistent is philosophically relevant? In fact, it is not. As Cappelen points out, there is a difference between doing philosophy and responding to survey questions (2014, p. 283). Philosophers do not typically just shoot from the hip when they philosophise. Rather, they engage in careful deliberation, which makes them modify or reject many judgments that can seem plausible to them at first glance. It is of course possible that philosophers' biases or inconsistencies still sneak into their published work. But there is only one method of establishing whether this is the case: doing armchair philosophy. To find out whether a philosopher is committed to both *p* and not-*p,* one needs to carefully read her texts. Running experiments in which she might endorse both *p* and not-*p* cannot tell us anything about the quality of her published argumentation.

6. X-phi and understanding other cultures

Deutsch, like Ichikawa, thinks that the wholesale rejection of x-phi would be a mistake. But his reasoning is different. He argues that some data collected by the experimentalists "are relevant to how we should treat others and how, more fundamentally, we should understand the social practices of different groups of people" (Deutsch 2015, p. 160). I can see two problems with this view. First,

some experimental philosophers, like Knobe, argue that their studies do not reveal any *robust* cross-cultural differences in philosophical intuitions, even though they occasionally reveal differences that are *statistically significant* (Knobe 2023) – and Deutsch's point seems valid only if the robust differences exist. Secondly, even if some robust demographic variation has been discovered, it is far from clear whether the kind of understanding that Deutsch is talking about can be drawn from it.

Consider a study by Henrik Ahlenius and Torbjörn Tännsjö, which purportedly shows that the Chinese are more likely to judge that it is impermissible to divert the tram in the bystander scenario, and to push the man in the footbridge scenario, compared to Westerners. What do we make of this result? According to Ahlenius and Tännsjö the "Chinese are more prone to accept some form of nonconsequentialist ethics, according to which an action may be wrong even if it maximizes utility" (Ahlenius & Tännsjö 2012, p. 198). But this conclusion is unwarranted. To determine whether someone accepts some form of nonconsequentialist ethics we need to check how she *justifies* her judgments, however the study focuses exclusively on what judgments people of different ethnicities *make*. The same trolley judgment can be arrived at via both a consequentialist and a non-consequentialist reasoning – as I pointed out in chapter 1, Thomson herself makes it clear when she says she is only interested in non-consequentialist reasons to divert the tram. It is possible that the Western participants' thinking was not any less non-consequentialist than that of the Chinese participants.

Ahlenius and Tännsjö also speculate that the Chinese responses may have been influenced by socialism, Confucianism or Daoism. They admit they do not have much illuminating to say about the first two, but they offer an explanation of the connection between Daoism and the Chinese trolley judgments: under Daoism "inaction is seen as a virtue" as one is expected to accept the "flow of things" (ibid., p.189). However the shortcoming of this conjecture is fairly obvious: Daoists, or people influenced by Daoism, are not exactly known for staying passive in every possible situation. Any serious attempt at understanding Daoist ethics or Daoist culture should involve an account of what kind of inaction is seen as appropriate. But it is impossible to develop such an account just by asking Daoists about different trolley cases – the data is simply far too limited. It is not even clear why trolley cases in particular are a suitable choice of subject for someone who wishes to better understand Daoist ethics of inaction. The choice of participants does not seem adequate either – randomly selected individuals from Beijing, Chengdu, Guangzhou and Shanghai are surely going to be influenced by Daoism to different degrees. In short: not much can be learned about the Chinese culture from Ahlenius's and Tännsjö's study, beyond the fact that the

Chinese give somewhat different responses to trolley scenarios. And this problem is typical of experimental philosophy studies that report cross-cultural differences.

Some may argue that these objections only apply to x-phi in a fairly narrow sense. However there is also a broader view. Stich and Tobia write that experimental philosophy can be defined as "empirical work undertaken with the goal of contributing to a philosophical debate, though of course that may not be the only goal" (2016, p. 5). Their examples of an early work of this kind are Richard Brandt's study of Hopi ethics (Brandt 1954) and John Ladd's study of Navaho ethics (Ladd 1957). Now, it is true that my criticism does not apply to Brandt of Ladd. However it has to be noted that what is typically meant by "experimental philosophy" is experimental philosophy in the narrow sense, namely post-2001, mostly questionnaire-based studies that focus on philosophical thought experiments. I doubt whether using the term in the broad sense is very helpful. The empirical part of Brandt's and Ladd's studies consist in carefully observing indigenous social practices and asking detailed questions about them, which makes both an example of a fairly straightforward, traditional ethnography. It is undeniable that traditional ethnography or anthropology is valuable to philosophers who work on, for example, moral relativism, value pluralism, toleration, or multiculturalism. But it is valuable precisely because it is very different from typical x-phi. It is more comprehensive, it studies behaviour as well as judgments, it is interested in reasoning processes, not just outcomes of these processes, it uses in-depth, multi-faceted interviews to learn about core values of a given culture, etc. This is not to say the data obtained by experimental philosophers is worthless. It is, however, too scarce to offer the kind of understanding Deutsch seems to be hoping for.

Another line of defence of x-phi that appeals to the value of cross-cultural understanding may look like this: philosophers necessarily need starting points for their arguments. What counts as a suitable starting point is something that seems plausible to most readers. It might not be hard to guess what seems plausible to readers of a similar cultural background. But what if someone wishes to make their argument appealing to people of a completely different background? In this case some kind of experimental research seems inevitable. For example, in the previous chapter I wrote about David Benatar who starts his argument with the claim that there is no duty to create happy people, but there is a duty to avoid creating suffering people. He argues that the reason why he chose the claim is that it is widely accepted. This is probably a fair bet with respect to middle-class Westerners – but most people in the world are not middle-class Westerners. If Benatar hopes to convince people culturally different from himself, he needs to first check how many of them actually accept his

starting premise. Thus even if the original motivation behind cross-cultural x-phi is misguided, the project can still be valuable.

This argument seems stronger than other attempts to save x-phi from the charge of insignificance. It might be objected that experimental philosophers tend to focus on judgments that are typically *not* used as starting premises of arguments, like trolley judgments or Gettier judgments. But this problem can be easily solved simply by focusing on something else. Even if the argument cannot vindicate any actual x-phi, it can still vindicate a lot of possible x-phi. However it should also be noted that according to the argument the role of x-phi is much more modest than it is often advertised: philosophers might use it if they want to appeal to certain demographics whose views they are unsure of. On this account x-phi cannot give rise to any "methodological revolution" or constitute an alternative to the traditional way of justifying claims in philosophy.

7. X-phi without the dogma

I have argued that experimental philosophy is in trouble because it relies on DE, which is false. I have also argued against some of those who believe that experimental philosophy can be philosophically significant despite the fact it relies on DE. However we can also come across rare examples of what appears to be a DE-free x-phi. In chapter 1 I argued that "philosophers rely on intuition-states as evidence" is ambiguous between "philosophers rely on the fact that *p* is intuitive as evidence for *p*" and "philosophers rely on the fact that *p* is intuitive as evidence for *q*". I argued that while the former is always false, the latter is sometimes true. To illustrate this point I brought up Nozick's experience machine – I argued that Nozick does not treat the intuitiveness of "one should not plug oneself into the machine" as evidence that one should not plug oneself into the machine, however he does treat it as evidence that what we value, as a matter of fact, something beyond just pleasure or what gives us pleasure. In other words, he uses an intuition-state as evidence against *psychological* hedonism, as opposed to *ethical* hedonism. Now, it seems entirely appropriate to carry out an experimental study on whether and why the experience machine judgment is intuitive. If it is not, or if it is, but for reasons other than those specified by Nozick, then Nozick's evidence against psychological hedonism may be undermined.

As it turns out, one such study has been carried out. Experimental philosopher Felipe De Brigard presented a group of participants with what can be called "the reverse experience machine scenario": it is revealed that all your conscious life you have been plugged into the machine, and now you can choose to go back to reality. Would you like to be unplugged? Most participants

refused, unless it was specified that in the real life they were multimillionaire artists living in Monaco, in which case they were split roughly in half. De Brigard argues that these results can be explained by the status quo bias: people do not want to "to abandon the life they have been experiencing so far, regardless of whether such life is virtual or real" (De Brigard 2010, p. 43). This implies that Nozick is wrong to argue that we reject the machine because we care about contact with reality, which means we care about something other than pleasure. As for myself, I am inclined to think that Nozick is right and De Brigard is wrong. The simpler explanation seems to be that people do not want to plunge themselves into a life completely disconnected from the one currently have, as it would not be *their* life anymore, but a life of a different person. Simply put, they are afraid they would cease to exist. Hence I think that in addition to caring about contact with reality we also care about staying alive. However, irrespective of whose explanation is correct, De Brigard's results seem clearly relevant to the evaluation of Nozick's argument.

It might be objected that evaluating psychological hedonism is a matter for psychologists, not philosophers. But this would presuppose a view of philosophy that seems unnecessarily narrow. As several enthusiasts of experimental philosophy have pointed out, few people insist that, for example, Locke's empirical argument against innate ideas is not philosophical (Prinz 2008, p. 190). If it safe to assume that Locke is engaging in philosophy, there is a good reason to think that De Brigard is doing the same.

Finally, it is worth noting that not all x-phi is questionnaire-based. Recently a number of experimental philosophers have turned to methods like computational analysis of linguistic corpora or behavioural experimentation (Fischer & Curtis 2019). It seems to me that at least a portion of this kind of x-phi can be considered DE-free. Take a recent study by Justin Sytsma, Roland Bluhm, Pascale Willemsen and Kevin Reuter. It focuses on how expressions like "to cause" and "to be responsible for" are used in English texts to examine four hypotheses: "ordinary causal attributions are sensitive to normative information", "outcome valence matters for ordinary causal attributions", "ordinary causal attributions [are] similar to responsibility attributions" and "causal attributions of philosophers [are] different from causal attributions we find in corpora of more ordinary language" (Systma et al. 2019, p. 210). This work does not seem to assume or be motivated by DE in any way. A separate question may be asked about whether it constitutes a *philosophical* contribution. Perhaps to count as philosophy it would have to link the four hypotheses to deeper problems of the nature of causation, or at least of the nature of the concept of causation. I do not believe that a definite answer can be given, as I do not believe a sharp line can be drawn between philosophy and linguistics or psychology.

In any case, I think it would be unfair to argue that all x-phi is worthless. However it is hard not to think that to the extent x-phi is significant it amounts to putting a new label on something old and familiar: sometimes simply conducting empirical research in psychology, linguistics, economics etc., and sometimes conducting this kind and research together with drawing philosophical conclusions from the results.

CHAPTER 6. Conclusion

1. Dismissing intuitions vs explaining intuitions

I started off by sketching two popular pictures of philosophy. According to one, philosophers relentlessly question our common sense beliefs and refuse to take "it just seems true" for an answer. According to the other, philosophers try to come up with theories that account for what seems true to us. These two pictures appear to be fundamentally at odds with each other. My thesis can be read as an attempt to examine this conflict. The outcome is fairly straightforward: there is no way of reconciling the two pictures, the former is true, and the latter is false. However this conclusion has to be qualified. First, the intuition-friendly picture is often confused with similar claims about the importance of intuition in philosophical methodology, and some of those claims are true. It is not the case that intuitions are treated as worthless in every possible sense. Secondly, by saying that one cannot reconcile the two pictures I do not mean that one cannot reconcile them *in principle*. To be sure, it is logically and metaphysically possible to rely on intuitions and dismiss intuitions at the same time – it just happens that actually existing philosophers often dismiss them and never rely on them.

I am clearly not the first to attack the intuition-friendly picture – several dissenters have done it before me. I owe them a lot. They planted the first seeds of doubt in me and eventually changed my mind. That said, I thought I still had something important and original to say on the topic. In this final chapter I am going to highlight what is new in my thesis and explain where I and other dissenters diverge.

2. What is new in this thesis?

The bulk of the first chapter is devoted to indentifying different meanings of claims like "philosophers rely on intuitions", "philosophers appeal to intuitions", "philosophers use intuitions as evidence" etc. I describe one ambiguity that has long been acknowledged – namely the state vs content one – but I also distinguish two more: the persuasion or clarification device vs evidence and the context of discovery vs context of justification. This gives me four, not just two, ideas of relying on intuitions in philosophy, of which, I argue, only one is false. I then go to some lengths to show

that most claims like "philosophers rely on intuitions" reveal a commitment to the false reading – something I call "descriptive evidentialism" (DE). Next, I argue that philosophers often equivocate between different readings of "philosophers rely on intuitions", that is they tend to present reasons to accept one of the true readings as reasons to accept DE.

I also examine the notion of "treating something as evidence". So far it has been rather neglected – while some specific accounts have been offered, most participants in the debate would assume that the thesis stands or falls irrespective of how "treating as evidence" is understood, and for the same reasons. I argue that neither the theory-neutral approach nor the existing specific accounts are adequate. In response I distinguish three basic senses in which something can be treated as evidence – doxastic, experiential and factual – and argue that each of them can fit into DE only if it is somehow translated into what I call *inferential evidence*, which can be either explicit or tacit. I then develop criteria for establishing whether DE is true of particular examples of philosophical practice. The discussion about using intuitions in philosophy have been largely focused on a relatively small number of paradigm cases, and the so-called "method of cases" they are supposed to exemplify. My contribution to this discussion is twofold. First, I distinguish three main interpretations of "the method of cases", two of which – the abductive and the noninferential – seem more DE-friendly. I substantially expand on Deutsch's criticism of the former and offer a brand new criticism of the latter. Secondly, I describe the limitations of the "method of cases" approach to testing DE. My own criteria are designed to overcome those limitations – they allow to examine the claim that philosophers rely on intuitions independently of whether the alleged method exists and what falls under its scope.
That said, I also utilise the traditional approach of studying the paradigm cases. Most notably, I offer the first, to my knowledge, examination of Plato's alleged reliance on intuition in his analysis of justice in *The Republic*, and arguably the most detailed to date examination of Thomson's alleged reliance on intuition in her series of articles on the trolley problem.

In chapter 2 I address seven arguments for DE. Only two of them – the argument from inevitability and the argument from intuition-talk – have been met with replies. I expand on the existing critical responses to the last two, and offer first responses to the remaining five: the arguments from endorsement, non-coincidence, error theories, counterexample diversity and intuitionism.

In chapter 3 I describe eight theories that reconcile the fact of dismissing intuitions with the alleged fact of relying on intuitions. Of course I do not start from scratch – each of the theories is based on

an idea that can be found in the literature. However hardly any of these ideas have been explicitly used to address the problem of counterintuitive conclusions. Moreover, to the extent the problem has been addressed, a number of important ambiguities have been ignored. The chapter attempts to fix this problem. I argue that dismissing intuitions and relying on intuitions at the same time in a DE-friendly sense has to be sharply distinguished from dismissing intuitions and relying on intuitions at the same time in a DE-unfriendly sense, and I offer criteria to test both hypotheses.

In chapter 4 I apply my criteria to three concrete examples of reaching a counterintuitve conclusion in ethics: Michael Tooley's claim that infanticide is morally permissible, David Benatar's claim that coming into existence is always a serious harm and John Taurek's claim that more deaths are not worse than less deaths, all else equal. The outcome of the investigation is that none of these arguments involves appealing to intuitions in any DE-friendly sense. I also put these arguments in the wider context and argue that what is true of Tooley, Benatar and Taurek is also true of all philosophers who have openly attacked them or inadvertently undermined their premises. The chapter constitutes a novel contribution in two ways. First, it analyses the three arguments in terms of whether they combine appealing to intuitions with dismissing intuitions. Secondly, and more importantly, it provides a concrete alternative to the traditional analysis of paradigm cases as a method of evaluating DE.

In chapter 5 I argue that DE is the core assumption behind the project of experimental philosophy. I outline three arguments that can be given in its defence by those who concede my point: experimental philosophy, even if it rests on a mistake, can still be valuable in terms of the cross-cultural analysis of concepts, facilitating cross-cultural understanding, and overcoming philosophical biases. I then offer critical responses to all three. The reply to the last one expands on Cappelen's criticisms, the first two are new.

3. Delusion or mistake?

Let me now move on to the differences between myself and other "intuition deniers" (to borrow Jennifer Nado's term), most notably Cappelen and Deutsch. There seem to be five main points of disagreement: whether experimental philosophy is entirely worthless, whether the philosophical intuition-talk is gibberish, how to explain the popularity of the intuition-dogma, whether

philosophers *justify* their judgments about cases, and whether we should be optimistic about traditional philosophical methodology. The first point has already been addressed in the previous chapter, and the second one in chapter 2. To recap: Cappelen's claim that philosophers' use of "intuition" is semantically defective is undermined by Cappelen's own criteria of intuitiveness, employed in his second argument, and by his own claims about what the term refers to in philosophical texts, made in his first argument. What he says about the lack of agreed upon definitions or paradigm cases also seems unjustified. His scepticism should therefore be rejected. I will now discuss the remaining three points in turn.

According to Cappelen "on no sensible construal of 'intuition', 'rely on', 'philosophy', 'evidence', and 'philosopher' is it true that philosophers in general rely on intuitions as evidence when they do philosophy" (Cappelen 2012, p. 3). Furthermore, the view that philosophers rely on intuitions as evidence is not only false, it is "ridiculous" (Cappelen 2014b, p. 594). And yet a great number of philosophers endorse it. This raises the question: are they delusional? How is it possible that so many intelligent, open-minded people, whose job is to think carefully and rigorously, believe something so manifestly wrong about their own practice? And how is it possible that so many of them refuse to accept the overwhelming evidence to the contrary, as presented to them by Cappelen? Here is Cappelen's tentative answer:

> I am inclined to put weight on what I think of as a verbal tick (or virus): philosophers started to use expressions such as 'Intuitively, BLAH' a lot. The fact that philosophers started using such locutions created the illusion that Centrality is true. (…) There might be an interesting question to be answered about where this verbal tick originated and what allowed it to spread. (Cappelen 2012, p. 22)

Capplen mentions Jaakko Hintikka's suggestion that Chomskyan linguistics might be to blame. Other possible influences are ordinary language philosophy with its emphasis on "what we would say", G. E. Moore's intuitionism and Rawls's reflective equilibrium. However none of these answers are, in Cappelen's own view, satisfactory. He is not wrong. If philosophers are minimally rational, they should generally be able to see the verbal tick for what it is and realise it does not justify the claim that they rely on intuitions as evidence. Upon some reflection they should also be able to understand that different intuition-related theories, methods, or research programmes have little in common with Centrality.

Contrast this with my explanation. First, I disagree with Cappelen that the sentence "philosophers rely on intuitions as evidence" is false *on any sensible construal*. By making this claim, he seems to

be throwing the baby out with the bathwater. I have argued that intuitions can be relied on as evidence of something different than their contents, and, more importantly, as evidence of their contents in the context of discovery. To be fair to Cappelen, in a passing remark he allows the latter possibility (ibid., p. 230), but then the remark seems to contradict his "no sensible construal" claim. Moreover, I have argued that intuitions are often relied on not as evidence, but rather as devices of clarification and persuasion. This gives us several true claims that are easy to conflate with the prevalent intuition-dogma, namely DE. It is not merely *potentially* easy to equivocate between different meanings of "philosophers rely on intuitions" – in chapter 1 I gave concrete examples of this equivocation being committed. It seems that nobody is being irrational. Rather, we are dealing with an honest, understandable *mistake*. There is no great mystery in the fact the dogma is so widely endorsed.

Deutsch's position is slightly more similar to mine. Just like Cappelen, he argues that the widespread use of the intuition-talk is responsible for the popularity of the intuition-dogma. But he also adds that the state vs content ambiguity is to blame (Deutsch 2015, pp. 131-2) – which means there is at least one true interpretation of "philosophers rely on intuitions". I do not disagree with any of Deutsch's suggestions, however, as I argued in chapter 1, the popularity of DE is hard to explain if we only invoke this one ambiguity. The other two I introduce make the phenomenon much less puzzling.

4. Intuitions and abductive inferences

In chapter 1 I described three competing interpretations of how judgments about cases are treated by philosophers: the justification, the abductive and the non-inferential. Deutsch insists that the first one is *always* true: philosophers always provide evidence for case judgments and this evidence is unrelated to psychological facts about the judgments' intuitiveness. This view, I think, is mistaken. First, as I have argued, the category of "judgments about cases" is hopelessly vague. A great variety of judgments can fall under its scope, and it would be unreasonable to think that none of them are ever treated as argumentative starting points. Secondly, even if we put this problem aside, we can find fairly clear-cut examples of judgments about cases that *are* treated as starting points in abductive arguments. The example I used in chapter 1 was the exchange between Boonin and Marquis on the right to life. Jennifer Nado points out that in his discussion of the Gettier problem

literature Deutsch himself quotes Alvin Goldman, who writes that he tries to "account for the fact that Smith cannot be said to know that p" (Nado 2017, p. 395). She argues that even if Deutsch is right about how Gettier treats the Gettier judgment, he is not right about how Goldman treats it. I think she may well have a point. But this is no evidence for DE, which requires that the intuitiveness of the judgment is meant to support it. Put another way, the abductive interpretation is merely *compatible* with DE. We can – and should – test DE independently, making use of the criteria I specified in chapter 1. And the outcome of such enquiries is, in my experience, always the same: it is much more plausible that the case judgment has been placed in the common ground rather than assumed to be supported by the fact it is intuitive. In my view, Deutsch is too invested in his defence of the justification interpretation. His book creates the false impression that the viability of DE depends on whether this interpretation is correct.

Cappelen's position on this issue seems more aligned with my own. He argues that straightforward abduction from judgments about cases is possible, albeit not common, and that it does not confirm Centrality in any way (2012, pp. 122-4). On the other hand, elsewhere he says that "philosophical practice treats unjustified judgments about philosophical cases as worthless" (ibid., p. 223), which seems hard to square with the abductive interpretation. In any case, according to Cappelen in most examples analysed in his book "the right reading is probably to take the writer's intention to be that support goes both directions: she sees T [the theoretical framework] as providing support to c [the judgment about the case], and vice versa" (ibid., p. 123). Deutsch says something similar about the literature on Gettier cases: "some of the arguments for Gettier judgments [...] are not, perhaps, simply arguments for Gettier judgments; they are also meant to be abductive arguments that proceed from the truth of a Gettier judgment to the truth of the epistemic principle that best explains it" (Deutsch 2015, p. 96). However he is adamant that no "pure" abduction ever takes place.

As for myself, I am not sure what to make of the "both directions" interpretation. It seems to imply that philosophers offer *circular* arguments, which is not a very charitable reading of their work. If one is meant to accept p on the basis of q and at the same time accept q on the basis of p, what value can such reasoning possibly have? There might be a response to this basic objection, however neither Cappelen nor Deutsch provides it. I therefore remain sceptical about it, which is another difference between us.

5. A case for pessimism

As Cappelen points out, many proponents of DE are concerned, if not downright pessimistic, about the role of intuitions in philosophy. They argue that intuitions are generally not a great source of evidence for anything. Even if certain intuitions are reliable, it is often added, we are not in a position to tell which ones are they. And, worse still, even when we can find ourselves in such position, intuitions become epistemically useless. These claims are usually accompanied by the sense that intuitions are nevertheless *inevitable* in a philosophical inquiry. The picture of philosophy that emerges from these considerations is nothing but gloomy.

Fortunately, replies Cappelen, the considerations are ill-founded. This is good news for metaphilosophy: a number of its most pressing problems turn out to be pseudoproblems. From this it does not, of course, follow that we should be optimistic about philosophy in general – it can still be plagued by other fundamental difficulties. But, according to Cappelen and Deutsch, unwarranted complaining about intuitions is just a recent instance of a larger phenomenon, namely unwarranted – or at least exaggerated – complaining about philosophical methods in general:

> The literature on philosophical methodology is dominated by hyperbolic claims about philosophy's moral and intellectual decline and corruption. Since the discipline's beginning, its imminent death has been a constant theme. This tradition of flagellantism starts with the Platonic dialogues, goes through Hume, Kant, Wittgenstein, and the logical positivists, and continues through, for example, Rorty, and the experimental philosophy movement. (Cappelen & Deutsch 2018)

An example of an argument very much in this vein is the common objection that philosophical methodology cannot be reliable as philosophy does not make any progress in terms of widening our knowledge: philosophers keep multiplying theories, arguments, technical concepts and hair-splitting distinctions, but at the end of the day no consensus on any matter is ever reached. Cappelen rejects this view by arguing it hinges on a questionable assumption that, first, persistent disagreement among experts means no collective knowledge is possible, and, second, that collective knowledge is essential to philosophical progress (Cappelen 2017). In short, he seems optimistic about the way philosophy is done in general, and there is an association between this kind of optimism and the rejection of the intuition-based picture of the discipline.

Unlike Cappelen and Deutsch, I am not an optimist, especially not about normative ethics. Granted, complaints about relying on intuitions as evidence are misguided – they must be, as no relying on intuitions as evidence takes place after all. But there still is a grain of truth in this critique. I have argued that relying on intuitions in the DE sense is often confused with placing intuitive contents in the common ground. The problem is that these contents can, I think, easily be used to neutralise one

another. For example, at one point in his book David Benatar notices that arguments put forward by Nils Holtug and Peter Singer are based on precisely what he denies in his conclusion (Benatar 2006, p. 202). We therefore have arguments which start with the claim that it is not always morally wrong to procreate, and the argument which ends with the claim that it *is* always morally wrong to procreate, starting from other intuitive premises, like "there is no duty to create happy people". In my analysis of Benatar's book I also mentioned another argument by Tännsjö which ends with the conclusion that there *is* a duty to create happy people, starting with another set of intuitive premises.

How do we tell which of these arguments, if any, are sound? I cannot see how our current philosophical methodology can offer a reasonable solution. Intuitive starting points can always be challenged, which is good, and challenging them never involves assessing their intuitiveness, which is also good. It does, however, involve appealing to similar intuitive claims, which can also be challenged in a similar fashion. This is problematic. In a sense, the bar for including something in the common ground is not set very high. And it is hard to imagine how it could be raised. For instance, it would not be feasible to restrict one's starting points to claims well supported by empirical science – not many philosophical conclusions can be drawn from such claims alone. It seems that intuitions are not very reliable, and yet there is no escape from relying on them.

This may sound very much like what many proponents of DE are saying, but the similarity is deceptive. What they have in mind is that intuition-states are poor evidence of their contents, but still need to be treated as evidence of their contents. What I have in mind is that intuition-contents can be undermined by other intuition-contents, but still need to be treated as argumentative starting points. The latter is not a serious worry, but the former is. It puts into question the possibility of making progress on issues like the morality of procreation and dozens of other moral issues. I am reluctant to make sweeping statements about impasse in ethics in general, or, a fortiori, philosophy in general. This thesis is, after all, a work of philosophy, and endorsing the global scepticism would make my argument self-refuting. Surely each debate needs to be assessed on its own merit. But I can see neither much hope for satisfactory answers to many of our perennial questions without a serious methodological reform, nor prospects for such reform. Freeing philosophy from the intuition-dogma may not turn out to be that liberatory after all.

BIBLIOGRAPHY

- Ahlenius, H. & Tännsjö, T. (2012). Chinese and Westerners Respond Differently to the Trolley Dilemmas. *Journal of Cognition and Culture, 12,* 195-201.
- Alexander, J. (2012). *Experimental philosophy. An introduction*. Cambridge: Polity Press.
- Alexander, J. & Weinberg, J. (2006). Analytic epistemology and experimental philosophy. *Philosophy Compass, 2/1*, 56–80.
- Alexander, J. & Weinberg, J. (2014). The "unreliability" of epistemic intuitions. In E. Machery & E. O'Neill (Eds.), *Current controversies in experimental philosophy* (pp. 128-145). New York: Routlege.
- Andow, J. (2015). How "intuition" exploded. *Metaphilosophy, 46* (2), 189-212.
- Andow, J. (2017). A partial defence of descriptive evidentialism about intuitions: a reply to Molyneux. *Metaphilosophy, 48* (1-2), 183-195.
- Anscombe, E. (1958). Modern moral philosophy. *Philosophy*, *33* (124), 1-19.
- Audi, R. (1993). Ethical reflectionism. *The Monist*, *76* (3/1), 295–315.
- Awad, D., Dsouza, S., Shariff, A., Rahwan, I., Bonnefon, J.-F. (2020) Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences of the United States of America, 117* (5), 2332-2337.
- Baggini, J. (2006). *The pig that wants to be eaten and ninenty-nine other thought experiments*. London: Granta Books.
- Baz, A. (2016). *The crisis of method in contemporary analytic philosophy*. Oxford: Oxford University Press.
- Bealer, G. (1998). Intuition and the autonomy of philosophy. In M. R. DePaul & W. Ramsey (Eds.), *Rethinking intuition. The psychology of intuition and its role in philosophical inquiry* (pp. 201-239). Lanham: Rowman & Littlefield.
- Bealer, G. (1999). A theory of the a priori. *Philosophical Perspectives*, *13*, 29-55.
- Bedke, M. S. (2008). Ethical intuitions: What they are, what they are not, and how they justify. *American Philosophical Quarterly, 45*(3), 253–269.

- Benatar, D. (2006). *Better never to have been. The harm of coming into existence*. Oxford: Oxford University Press.
- Benatar, D. (2013). Still Better Never to Have Been: A Reply to (More of) My Critics. *The Journal of Ethics, 17* (1/2), 121-151.
- Bengson, J. (2014). How philosophers use intuition and 'intuition'. *Philosophical Studies*, *171*, 555–576.
- Boyd, R. (1988). How to be a moral realist. In G. Sayre-McCord (Ed.), *Essays on moral realism* (pp. 181-228). Ithaca: Cornell University Press.
- Boonin, D. (2003). *A defense of abortion.* Cambridge: Cambridge University Press.
- Brandt, R. B. (1954). *Hopi ethics. A theoretical analysis*. Chicago: University of Chicago Press.
- Brandt, R. B. (1979). *A theory of the good and the right*. Oxford: Oxford University Press.
- Brandt, R. B. (1984). Utilitarianism and moral rights. *Canadian Journal of Philosophy, 14* (1), 1-19.
- Brennan, J. & Magness, P. (2019). *Cracks in the ivory tower. The moral mess of higher education.* Oxford: Oxford University Press.
- Briggs, R. A. (2019). Normative Theories of Rational Choice: Expected Utility. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, URL = <https://plato.stanford.edu/archives/fall2019/entries/rationality-normative-utility/ >.
- Bromberger, S. (1966). Why-questions. In R. G. Colodny (Ed.), *Mind and Cosmos. Essays in contemporary science and philosophy* (pp. 86-111). Pittsburgh: University of Pittsburgh Press.
- Brouwer, L.E.J. (1975). *Collected works, I*, A. Heyting (Ed.), Amsterdam: North-Holland.
- Brown, J. (2011). Thought Experiments, Intuitions and Philosophical Evidence. *Dialectica, 65*, 493-516.
- Brun, G. (2014). Reflective Equilibrium Without Intuitions? *Ethical Theory and Moral Practice, 17*, 237-252.
- Cappelen, H. (2012). *Philosophy without intuitions*. Oxford: Oxford University Press.

- Cappelen, H. (2014a). X-phi without intuitions? In A. R. Booth, D. P. Rowbottom (Eds.), *Intuitions* (pp. 269-286). Oxford: Oxford University Press.
- Cappelen, H. (2014b). Replies to Weatherson, Chalmers, Weinberg, and Bengson. *Philosophical Studies, 171*, 577-600.
- Cappelen, H. (2017). Disagreement in Philosophy: An Optimistic Perspective. In G. D'Oro, S. Overgaard (Eds.), *The Cambridge companion to philosophical methodology* (pp. 56-74). Cambridge: Cambridge University Press.
- Cappelen, H., Deutsch, M. (2018). Review of Avner Baz, The Crisis of method in contemporary analytic philosophy. *Notre Dame Philosophical Reviews*, URL = <https://ndpr.nd.edu/reviews/the-crisis-of-method-in-contemporary-analytic-philosophy/>.
- Chalmers, D. (2014). Intuitions in philosophy: a minimal defense. *Philosophical Studies*, *171*, 535–544.
- Chudnoff, E. (2013). *Intuition.* Oxford: Oxford University Press.
- Chudnoff, E. (2021). *Forming impressions. Expertise in perception and intuition.* Oxford: Oxford University Press.
- Churchland, P. M. (1981). Eliminative Materialism and the Propositional Attitudes. *The Journal of Philosophy, 78* (2), 67-90.
- Churchland, P. S. (2017). Neurophilosophy. In D. Livingstone Smith (Ed.), *How biology shapes philosophy* (pp. 113-148). Cambridge: Cambridge University Press.
- Clark, S. (2021). *Good lives. Autobiography, self-knowledge, narrative, and self-realisation*. Oxford: Oxford University Press.
- Climenhaga, N. (2018). Intuitions are used as evidence in philosophy. *Mind*, *127* (505), 69-104.
- Cooper, R. (2005). Thought experiments. *Metaphilosophy, 36* (3), 328-347.
- Cowley, C. (2011). Moral philosophy and the 'real world'. *Analytic Teaching and Philosophical Practice*, 31 (1), 21-30.
- Craig, W. L. (2008). *Reasonable faith. Christian truth and apologetics*. Wheaton: Crossway.
- Crisp, R. (2002). Sidgwick and the boundaries of intuitionism. In P. Stratton-Lake (Ed.), *Ethical intuitionism: re-evaluations* (pp. 56-75). Oxford: Oxford University Press.

- Crisp. R. (2021). Well-being. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, URL = <https://plato.stanford.edu/entries/well-being/>.
- Cummins, R. (1998). Reflection on Reflective Equilibrium In M. R. DePaul & W. Ramsey (Eds.), *Rethinking intuition. The psychology of intuition and its role in philosophical inquiry* (pp. 113-128). Lanham: Rowman & Littlefield.
- Daly, C. (2015). Introduction and historical overview. In C. Daly (Ed.), *The Palgrave handbook of philosophical methods* (pp. 1-30). New York: Palgrave Macmillan.
- Dancy, J. (2009). *Ethics without principles*. Oxford: Oxford University Press.
- Daniels, N. (2020). Reflective equilibrium. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, URL = <https://plato.stanford.edu/archives/sum2020/entries/reflective-equilibrium/>.
- Davenport, E. (1983). Literature as Thought Experiment (On Aiding and Abetting the Muse). *Philosophy of the Social Sciences, 13* (3), 279-306.
- De Brigard, F. (2010). If you like it, does it matter if it's real? *Philosophical Psychology, 23* (1), 43-57.
- De Cruz, H. (2015). Where Philosophical Intuitions Come From. *Australasian Journal of Philosophy, 93* (2), 233-249.
- DeRose, K. (2017). *The appearance of ignorance. Knowledge, skepticism, and context, volume 2*. Oxford: Oxford University Press.
- Deutsch, M. (2009). Experimental philosophy and the theory of reference. *Mind and Language, 24* (4), 445-466.
- Deutsch, M. (2015). *The myth of the intuitive. Experimental philosophy and philosophical method*. Cambridge: MIT Press.
- Deutsch, M. (2016). Gettier's method. In J. Nado (Ed.), Advances in experimental philosophy and philosophical methodology (pp. 69-98). London: Bloomsbury Academic.
- Deutsch, M. (2020). Conceptual analysis without concepts. *Synthese, 198* (11), 11125-11157.
- Doris, J. (2012). Lack of character. *Personality and moral behavior*. Cambridge: Cambridge University Press.

- Dorr, C. (2010). Review of James Ladyman and Don Ross, with David Spurrett and John Collier, Every Thing Must Go: Metaphysics Naturalized. *Notre Dame Philosophical Reviews*, URL = <https://ndpr.nd.edu/reviews/every-thing-must-go-metaphysics-naturalized/>.
- Downes, S. M. (2021). Evolutionary psychology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, URL = <https://plato.stanford.edu/entries/evolutionary-psychology/>.
- Dretske, F. (1981). *Knowledge and the Flow of Information*, Cambridge, MA: MIT Press.
- Earlenbaugh, J. & Molyneux, B. (2009). Intuitions are inclinations to believe. *Philosophical Studies*, *145*, 89-109.
- Epicurus (1994). *Letter to Menoeceus*: Diogenes Laertius 10.121-135. In B. Inwood & L. P. Gerson, *The Epicurus reader. Selected writings and testimonia* (pp. 28-31). Indianapolis: Hackett.
- Fischer, E., Collins., J. (2015). Rationalism and naturalism in the age of experimental philosophy. In E. Fischer, J. Collins (Eds.), Experimental, rationalism, and naturalism. Rethinking philosophical method (pp. 3-33). New York: Routledge.
- Fischer, E., Curtis, M. (Eds.) (2019). *Methodological advances in experimental philosophy*. London: Bloomsbury Academic.
- Fodor, J. (1998). *Concepts: Where Cognitive Science Went Wrong*, Oxford: Oxford University Press.
- Fodor, J., Garrett, M. (1967). Some syntactic determinants of sentential complexity. *Perception & Psychophysics, 2* (7), 289-296.
- Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, *5*, 5-15.
- Gendler, T. S. (2010). *Thought experiment. On the power and limits of imaginary cases.* New York: Garland.
- Gettier, E. (1963). Is justified true belief knowledge? *Analysis*, *23*(6), 121–123.
- Glymour, C. (1980). *Theory and evidence*. Princeton: Princeton University Press.
- Goff, P. (2017). *Consciousness and fundamental reality*. Oxford: Oxford University Press.

- Goldman, A. (1967). A Causal Theory of Knowing, *The Journal of Philosophy, 64* (12), 357-372.
- Goldman, A. (2007). Philosophical intuitions: Their target, their source, and their epistemic status. *Grazer Philosophische Studien, 74* (1), 1-26.
- Goldman, A. & Pust, J. (1998). Philosophical theory and intuitional evidence. In M. R. DePaul & W. Ramsey (Eds.), *Rethinking intuition. The psychology of intuition and its role in philosophical inquiry* (pp. 179-197). Lanham: Rowman & Littlefield.
- Goodman, N., (1955). *Fact, fiction & forecast*. Cambridge, MA: Harvard University Press.
- Grahek, N. (2007). *Feeling pain and being in pain*. Cambridge, MA: MIT Press.
- Greene, J. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology, vol. 3* (pp. 35-80). Cambridge, MA: MIT Press.
- Greene, J. (2013). *Moral tribes. Emotion, reason and the gap between us and them*. New York: Penguin.
- Griffin, J. (1988). *Well-being: its meaning, measurement, and moral importance.* Oxford: Clarendon.
- Griffin, J. (2008). *On human rights.* Oxford: Oxford University Press.
- Gutting, G. (1998). "Rethinking intuition": A historical and metaphilosophical perspective. In M. R. DePaul & W. Ramsey (Eds.), *Rethinking intuition. The psychology of intuition and its role in philosophical inquiry* (pp. 3-13). Lanham: Rowman & Littlefield.
- Hales, S. D. (2006). *Relativism and the foundations of philosophy*. Cambridge, MA: MIT Press.
- Hampton, J. (2006). Concepts as prototypes. In B.H. Ross (Ed.), *Psychology of learning and motivation. Volume 46* (pp. 79-113). New York: Academic Press.
- Hare, R. M. (1972). The argument from received opinion. In R. M. Hare, *Essays on philosophical method* (pp. 117-135), Berkeley: University of California Press.
- Hoyningen-Huene, P. (2006). Context of discovery versus context of justification and Thomas Kuhn. In J. Schickore & F. Steinle (Eds.), *Revisiting discovery and justification. Historical and philosophical perspective on the context ditinction* (pp. 119-131). Dordrecht: Springer.

- Hewitt, S. (2010). What do our intuitions about the experience machine really tell us about hedonism? *Philosophical Studies, 151* (3), 331-349.
- Hintikka, J. (1999). The emperor's new intuitions. *The Journal of Philosophy, 96* (3), 127-147.
- Horwich, P. (2012). *Wittgenstein's metaphilosophy*. Oxford: Oxford University Press.
- Huemer, M. (2008). *Ethical intuitionism*. New York: Palgrave Macmillan.
- Huemer, M. (2019). *Dialogues on ethical vegetarianism*. New York: Routledge.
- Hume, D. (1748/1999). *An enquiry concerning human understanding*. Edited by T. L. Beauchamp. Oxford: Oxford University Press.
- Ichikawa, J. J. (2014). Who Needs Intuitions? Two Experimentalist Critiques. In A. R. Booth, D. P. Rowbottom (Eds.), *Intuitions* (pp. 232-255). Oxford: Oxford University Press.
- Jackson, F. (1998). *From metaphysics to ethics. A defence of conceptual analysis*. Oxford: Oxford University Press.
- Jeske, D. (1993). Persons, compensation, and utilitarianism. *The Philosophical Review, 102,* (4), 541-575.
- Jeske, D. (2018). *The evil within. Why we need moral philosophy*. Oxford: Oxford University Press.
- Kaczor, C. (2015). *The ethics of abortion. Women's rights, human life, and the question of justice*. New York: Routledge.
- Kagan, S. (2001). Thinking about cases. *Social Philosophy and Policy*, *18* (2), 44-63.
- Kahneman, D. (2013). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.
- Kahneman, D., Tversky, A. (1984). Choices, values, and frames. *American Psychologist, 39* (4), 341-350.
- Kamm, F. (2007). *Intricate ethics. Rights, responsibilities, and permissible harm*. Oxford: Oxford University Press.
- Kauppinen, A. (2014). The rise and fall of experimental philosophy. In J. Knobe, S. Nichols (Eds.), *Experimental philosophy. Volume 2* (pp. 2-30). Oxford: Oxford University Press.
- Kavka, G. S. (1979). The numbers should count. *Philosophical Studies, 36*, 285-294.

- Kekulé, A./ Benfey, O. T. (1890/1958). August Kekulé and the birth of the structural theory of organic chemistry in 1858. *Journal of Chemical Education*, *35* (1), 21-23.
- Knobe, J. (2023). Difference and robustness in the patterns of philosophical intuition across demographic groups. *Review of Philosophy and Psychology, 14* (2), 435-455.
- Knobe, J., Buckwalter, W., Nichols, S., Robbins, P., Sarkissian, H., Sommers, T. (2012). Experimental philosophy. *Annual Review of Psychology, 63*, 81-99.
- Knobe, J., Nichols, S. (2008). The Experimental Philosophy Manifesto. In J. Knobe, S. Nichols (Eds.), *Experimental philosophy.* (pp. 3-16). Oxford: Oxford University Press.
- Knobe. J., Nichols, S. (2017). Experimental philosophy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, URL = <https://plato.stanford.edu/entries/experimental-philosophy/>.
- Kolodny, N. (2017). Help wanted: subordinates. In E. Anderson, *Private government* (pp. 99-107). Princeton: Princeton University Press.
- Kornblith, H. (1998). The role of intuitions in philosophical enquiry: an account with no unnatural ingredients. In M. R. DePaul & W. Ramsey (Eds.), *Rethinking intuition. The psychology of intuition and its role in philosophical inquiry* (pp. 129-141). Lanham: Rowman & Littlefield.
- Korsgaard, C. M. (1996). Personal identity and the unity of agency: A Kantian response to Parfit. In C. M. Korsgaard, *Creating the Kngdom of Ends* (pp. 363-388). Cambridge: Cambridge University Press.
- Korsgaard, C. M. (2008). Realism and constructivism in twentieth-century moral philosophy. In C. M. Korsgaard, *The constitution of agency: essays on practical reason and moral psychology* (pp. 302-326). Oxford: Oxford University Press.
- Korsgaard, C. M. (2014).On having a good. *Philosophy, 89* (3), 405-429.
- Korsgaard, C. M. (2018). *Fellow creatures. Our obligations to the other animals.* Oxford: Oxford University Press.
- Kripke, S. A. (1980). *Naming and necessity.* Oxford: Basil Blackwell.
- Kuhse, H., Singer, P. (1985). *Should the baby live? The problem of handicapped infants.* Oxford: Oxford University Press.

- Ladd, J. (1957). *The structure of a moral code. A philosophical analysis of ethical discourse applied to the ethics of the Navaho Indians.* Cambridge, MA: Harvard University Press.
- Landes, E. (2020). The threat of the intuition-shaped hole. *Inquiry*, 1-26.
- Levin, M. E, Levin, M. R. (1977). Flagpoles, shadows and deductive explanation. *Philosophical Studies, 32* (3), 293-299.
- Lewis, D. (1983). *Philosophical papers. Volume 1*. Oxford: Oxford University Press.
- Lipton, P. (2001). Is explanation a guide to inference? A reply to Wesley C. Salmon. In G. Hon & S. S. Rakover, *Explanation. Theoretical approaches and applications* (pp. 93-120). Dordrecht: Springer.
- Ludwig., K. (2007). The epistemology of thought experiments: first person versus third person approaches. *Midwest Studies in Philosophy,* 31 (1), 128-159.
- Lycan, W. (2002). Explanation and epistemology. In P. K. Moser (Ed.), *The Oxford Handbook of Epistemology* (pp. 408-433). Oxford: Oxford University Press.
- Lycan, W. (2013). On Two Main Themes in Gutting's *What Philosophers Know. The Southern Journal of Philosophy, 51* (1), 112-120.
- Lyons, J. C. (2016). Experiential evidence? *Philosophical Studies, 173*, 1053-1079.
- MacAskill, W. (2022). *What we owe the future.* London: Oneworld.
- Machery, E. (2009). *Doing without concepts*. Oxford: Oxford University Press.
- Machery, E. (2017). *Philosophy within its proper bounds*. Oxford: Oxford University Press.
- Machery, E., Stich, S., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., Sirker, S., Usui, N., Hashimoto, T. (2015). Gettier across cultures. *Nous, 51* (3), 645-664.
- Mackonis, A. (2013). Inference to the best explanation, coherence and other explanatory virtues. *Synthese, 190* (6), 975-995.
- Magidor, O. (2022). Category mistakes. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, URL = <https://plato.stanford.edu/entries/category-mistakes/>.
- Malmgren, A. (2011). Rationalism and the content of intuitive judgements. *Mind, 120* (478), 263-327.
- Margolis, E., Laurence, S. (2023). Concepts. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, URL = <https://plato.stanford.edu/entries/concepts/>.

- McGinn, C. (2012). *Truth by Analysis. Games, Names, and Philosophy*. Oxford: Oxford University Press.
- McMahan, J. (2013). Moral intuition. In H. LaFollette & I. Persson (Eds.), *The Blackwell guide to moral theory* (pp. 103-120). Chichester: Wiley.
- Mikhail, J. (2013). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment.* Cambridge: Cambridge University Press.
- Miller, R. (2000). Without intuitions. *Metaphilosophy,* 31 (3), 231-250.
- Miščević, N. (2021). *Thought experiments*. Dordrecht: Springer.
- Molyneux, B. (2014). New arguments that philosophers don't treat intuitions as evidence. *Metaphilosophy, 45* (3), 441-461.
- Nado, J. (2017). Demythologizing intuition. *Inquiry, 60* (4), 386-402.
- Nagel, T. (1986). *The view from nowhere*. Oxford: Oxford University Press.
- Nagel, T. (2002). *Mortal questions*. Cambridge: Cambridge University Press.
- Nolan., D. (2009). Platitudes and metaphysics. In D. Braddon-Mitchell, R. Nola (Eds.), *Conceptual analysis and philosophical naturalism* (pp. 267-300). Cambridge, MA: MIT Press.
- Norton, J. D. (2004). On thought experiments: is there more to the argument?*Philosophy of science,* 71 (5), 1139-1151.
- Nozick, R. (1969). Newcomb's problem and two principles of choice. In N. Rescher (Ed.), *Essays in honor of Carl G. Hempel* (pp. 114-146). Dordrecht: Reidel.
- Nozick, R. (1979). *Anarchy, state, and utopia*. New York: Basic Books.
- Nozick, R. (1989). *The examined life. Philosophical meditations*. New York: Simon & Schuster.
- O'Neill, O. (2005). The dark side of human rights. *International Affairs, 81* (2), 427-439.
- Otsuka, M. (2000). Scanlon and the claims of the many versus one. *Analysis, 60*, 288-293.
- Otsuka, M. (2006). Saving lives, moral theory, and the claims of individuals. *Philosophy & Public Affairs, 34* (2), 109-135.
- Papineau, D. (2009). The poverty of analysis. *Aristotelian Society Supplementary Volume, 83* (1), 1-30.

- Parfit, D. (1978). Innumerate ethics. *Philosophy & Public Affairs, 7* (4), 285-301.
- Parfit, D. (1987). *Reasons and persons.* Oxford: Oxford University Press.
- Parfit, D. (2011). *On what matters. Volume 1.* Oxford: Oxford University Press.
- Paulo, N. (2020). The unreliable intuitions objection against reflective equilibrium. *The Journal of Ethics*, 24, 333–353.
- Plato (2013). *Republic, Volume I: Books 1-5*. Edited and translated by C. Emlyn-Jones, W. Preddy. Loeb Classical Library 237. Cambridge, MA: Harvard University Press.
- Prinz, J. (2008). Empirical philosophy and experimental philosophy. In J. Knobe, S. Nichols (Eds.), *Experimental philosophy.* (pp. 189-208). Oxford: Oxford University Press.
- Pust, J. (2000). *Intuitions as evidence*. New York: Routledge.
- Pust, J. (2019). Intuition. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, URL = <https://plato.stanford.edu/archives/sum2019/entries/intuition/>.
- Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review, 60* (1), 20-43.
- Quine, W. V. O. (1960/2013). *World and object.* Cambridge, MA: MIT Press.
- Rachels, S. (1998). Is it good to make happy people? *Bioethics, 12* (2), 93-110.
- Rawls, J. (1951). Outline of a decision procedure for ethics. *The Philosophical Review, 60* (2), 177-197.
- Rawls, J. (1971/1999). *A theory of justice*. Cambridge: Harvard University Press.
- Rawls, J. (1974). The independence of moral theory. *Proceedings and Addresses of the American Philosophical Association, 48*, 5-22.
- Raz, J. (1986). *The morality of freedom.* Oxford: Oxford University Press.
- Rey, G. (2023). The analytic/synthetic distinction. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, URL = <https://plato.stanford.edu/entries/analytic-synthetic/>.
- Ridge, M.. McKeever, S. (2021). Moral particularism and moral generalism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, URL = <https://plato.stanford.edu/archives/sum2023/entries/moral-particularism-generalism/ >.

- Rosenhouse, J. (2009). *The Monty Hall problem. The remarkable story of math's most contentious brain teaser*. Oxford: Oxford University Press.
- Ross., W. D. (1930/2007). *The right and the good*. Edited by P. Stratton-Lake. Oxford: Clarendon Press.
- Rowland, R. (2017). Our intuitions about the experience machine. *Journal of Ethics and Social Philosophy, 12* (1), 110-117.
- Russell, B. (1914/2009). *The philosophy of logical atomism*. New York: Routledge.
- Sanders, J. T. (1987). Why the numbers should sometimes count. *Philosophy and Public Affairs, 17* (1), 3-14.
- Scanlon, T. M. (1998). *What we owe to each other*. Cambridge: Harvard University Press.
- Scholz, B. C., Pelletier, F. J., Pullum, G. K., Nefdt, R. (2022). Philosophy of linguistincs. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, URL = <https://plato.stanford.edu/entries/linguistics/>.
- Schwarz, S. (1990). *The moral question of abortion.* Chicago: Loyola University Press.
- Schwitzgebel, E., Cushman, F. (2015). Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition, 141,* 127-137.
- Sencerz, S. (1983). Moral intuitions and justification in ethics. *Philosophical Studies, 50* (1), 77-95.
- Shaw, W. H. (1980). Intuition and moral philosophy. *American Philosophical Quarterly, 17* (2), 127-134.
- Sidgwick, H. (1874/1962). *The methods of ethics. Seventh edition*. London: Palgrave Macmillan.
- Singer, P. (2002). *Animal liberation*. New York: Harper Collins.
- Singer, P. (2005). Ethics and intuition. *The Journal of Ethics, 9*, 331-352.
- Singer, P. (2009). *The life you can save. How to do your part to end world poverty*. New York: Random House.
- Sinnott-Armstrong, W., Young, L., Cushman, F. (2010). Moral intuitions. In J. Doris (Ed.), *The moral psychology handbook* (pp. 46-73). Oxford: Oxford University Press.

- Sosa, E. (2009). A defense of the use of intuitions in philosophy. In D. Murphy & M. Bishop (Eds.), *Stich and his critics* (pp. 101-112). Chichester: Wiley.
- Souder, L. (2003). What are we to think about thought experiments? *Argumentation, 17* (2), 203-217.
- Stich, S. (2010). Philosophy and WEIRD intuition. *Behavioral and Brain Sciences,* 33 (2-3), 110-111.
- Stich, S. & Tobia, K. (2016). Experimental philosophy and the philosophical tradition. In J. Sytsma (Ed.), *A companion to experimental philosophy* (pp. 5-21). Chichester: Wiley.
- Stratton-Lake, P. (2007). Introduction. In: W. D. Ross, *The right and the good* (pp. Ix-l). Oxford: Clarendon Press.
- Stratton-Lake, P. (2020). Intuitionism in ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, URL = <https://plato.stanford.edu/entries/intuitionism-ethics/>.
- Strevens, M. (2019). *Thinking off your feet. How empirical psychology vindicates armchair philosophy*. Cambridge: Harvard University Press.
- Stuart, M. T., Fehige, Y., Brown, J. R. (2018). Thought experiments: state of the art. In M. T. Stuart, Y. Fehige, J. R. Brown (Eds.), *The Routledge companion to thought experiments* (pp. 1-28). New York: Routledge.
- Sumner, L. W. (1986). A review of Michael Tooley, Abortion and infanticide. *Canadian Journal of Philosophy, 16* (3), 527-543.
- Sytsma, J., Bluhm, R., Willemsen, P., Reuter, K. (2019). Causal attributions and corpus analysis. In E. Fischer & M. Curtis (Eds.), *Methodological advances in experimental philosophy* (pp. 209-238). London: Bloomsbury Academic.
- Tännsjö, T. (2002). Why we ought to accept the Repugnant Conclusion. *Utilitas, 14* (3), 339-359.
- Taurek, J. M. (1977). Should the numbers count? *Philosophy and Public Affairs*, 6 (4), 293-316.
- Taurek, J. M. (2020). Reply to Parfit's *Innumerate ethics*. J. McMahan, T. Campbell, J. Goodrich & K. Ramakrishnan (Eds.), *Principles and persons. The legacy of Derek Parfit* (pp. 311-322). Oxford: Oxford University Press.

- Thagard, P. (2010). *The brain and the meaning of life*. Princeton: Princeton University Press.
- Thomas, N. (2021). Mental imagery. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, URL = <https://plato.stanford.edu/entries/mental-imagery/>.
- Thompson, M. (1972). Singular terms and intuitions in Kant's epistemology. *Review of Metaphysics,* 26 (2), 314-343.
- Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, 59 (2), 204-217.
- Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal, 94* (6), 1395-1415.
- Thomson, J. J. (2008). Turning the trolley. *Philosophy and Public Affairs, 36* (4), 359-374.
- Timmermann, J. (2004). How people count, but not their numbers. *Analysis, 64* (2), 106-112.
- Tobia, K. (2015). Philosophical method and intuitions as assumptions. *Metaphilosophy, 46* (4-5), 575-594.
- Tobia, K., Buckwalter, W., Stich., S. (2012). Moral intuitions: Are philosophers experts? *Philosophical Psychology, 26* (5), 629-638.
- Tooley, M. (1972). Abortion and infanticide. *Philosophy & Public Affairs, 2* (1), 37-65.
- Tooley. M. (1983). *Abortion and infanticide.* Oxford: Clarendon Press.
- Turri, J. (2016). Knowledge judments in "Gettier" cases. In J. Sytsma & W. Buckwalter (Eds.), *A companion to experimental philosophy* (pp. 337-348). Chichester: Wiley.
- Tushnet, M., Seidman, L. M. (1986). A comment on Tooley's *Abortion and infanticide*. *Ethics, 96* (2), 350-355.
- Unger, P. (1979). There are no ordinary things. *Synthese, 41* (2), 117-154.
- Unger, P. (1996). *Living high and letting die. Our illusion of innocence*. Oxford: Oxford University Press.
- van Inwagen, P. (1997). Materialism and the psychological-continuity account of personal identity. In J. Tomberlin (Ed.), *Philosophical Perspectives 11. Mind, Causation and World* (pp. 305–19). Wiley.

- Weatherson, B. (2003). What good are counterexamples? *Philosophical Studies, 115* (1), 1-31.
- Weijers, D. (2014). Nozick's experience machine is dead, long live the experience machine! *Philosophical Psychology, 27* (4), 513-535.
- Weinberg, J., Nichols, S., Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics,* 29 (1-2), 429-460.
- Williamson, T. (1998). *Vagueness*. New York: Routledge.
- Williamson, T. (2018). *Doing philosophy. From common curiosity to logical reasoning*. Oxford: Oxford University Press.
- Williamson, T. (2022). *The philosophy of philosophy. Second edition*. Hoboken, NJ: Wiley-Blackwell.
- Wittgenstein, L. (1921/2001). *Tractatus logico-philosophicus*. Translated by D. F. Pears & B. F. McGuinness. New York: Routledge.
- Wright, J. C. (2014). On intuitional stability: The clear, the strong, and the paradigmatic. In J. Knobe, S. Nichols (Eds.), *Experimental philosophy. Volume 2* (pp. 51-74). Oxford: Oxford University Press.
- Wysocki, T. (2017). Arguments over intuitions? *Review of Philosophy and Psychology,* (2), 1-23.