# Chapter 16
# Ethics and Empirical Psychology – Critical Remarks to Empirically Informed Ethics

**Antti Kauppinen**

## 16.1 Introduction

The question of whether ethics should be empirically informed has a rhetorical ring to it—how could it be better to be uninformed? Exciting developments in a number of disciplines studying human beings, from psychology and cognitive science to biology, offer hope that ethics, too, could make steady progress were it to hitch its wagons to the train of science. So it is no surprise that some want to erase what they see as outdated and old-fashioned disciplinary boundaries, and no bigger surprise that others react by reaffirming traditional methodologies or by retreating to the grand journals of old. My instinct is on the side of caution in this debate, but I will refrain from grand pronouncements. Disciplinary border skirmishes seem to invite the greatest sin in writing—being boring. In contrast, particular arguments that aim to make concrete progress with existing questions by exploiting a novel methodology can be stimulating even when they go wrong.

So what I will do in this paper is discuss six attempts to draw on psychological discoveries in metaethics and normative ethics. I will focus on psychology, since it is the branch of science that seems to be most closely relevant to ethics. The line between the two disciplines is also particularly porous, which is indicated by the fact that psychology was among the last sciences to gain independence from philosophy. For reasons of space and coherence, I cannot engage much with work inspired by other disciplines, although I believe at least some of the lessons learned from psychology will generalize.

As a general background, I will sketch two opposing philosophical outlooks—one might almost call them philosophical *temperaments*. It is important not to caricature these positions. Moral philosophers have never claimed that empirical facts play no role in ethics. Ancient and Early Modern ethicists and moralists

A. Kauppinen (✉)
Department of Philosophy, Trinity College Dublin, Dublin, Ireland
e-mail: kauppina@tcd.ie

M. Christen et al. (eds.), *Empirically Informed Ethics: Morality between Facts and Norms*,
Library of Ethics and Applied Philosophy 32, DOI 10.1007/978-3-319-01369-5_16,

certainly did not shy away from a variety of empirical claims, and though Hume and Kant in very different ways argued for principled limits of what empirical knowledge can do, they did also draw on a particular understanding of human nature in their ethical works. It is true, however, that in the twentieth century, as the human sciences developed their own empirical methods, philosophers did come to focus on questions that could not be settled by empirical research. I will call the view of that emerged *Armchair Traditionalism* and sum it up in two main theses:

1. In *metaethics*, empirical facts are only relevant for causal explanations of particular moral judgments and the capacity to make moral judgments.
2. In *normative ethics*, empirical facts are only relevant for deriving judgments about particular cases from non-empirical principles and for practical recommendations.

Roughly, then, psychology, social sciences, and biology can tell us why and how people make moral judgments, but not what those judgments are or what if anything makes them true. They can also supply material for minor premises in ethical arguments—it is perhaps *a priori* true that creatures capable of pleasure and pain deserve moral consideration, but whether fetuses are sensate creatures is an empirical question. And insofar as ethics is practical, it needs to issue recommendations that are actually useful to people, which means they depend not only on moral facts but also facts about people. For example, even if utilitarianism is the true moral theory, it is going to depend on facts about human beings what decision procedure they should employ to best approximate actions that maximize utility (see e.g. Railton 1984). There is no doubt that if we are interested in promoting moral behavior and moral thinking, or in designing environments that foster moral development and engagement, we need to look to empirical psychology (for concrete suggestions, see e.g. Chap. 7 by Tanner and Christen, this volume; Chap. 13 by Narvaez and Lapsley, this volume). But that is it: the role of empirical facts is *marginal*, not *essential* or *fundamental* to ethical inquiry.

In making the case against armchair ethics, John Doris and Stephen Stich say:

> It is not possible to step far into the ethics literature without stubbing one's toe on empirical claims. The thought that moral philosophy can proceed unencumbered by facts seems to us an unlikely one: There are just too many places where answers to important ethical questions require—and have very often presupposed—answers to empirical questions. (Doris and Stich 2005, 115)

On one interpretation, this claim is not as such incompatible with Armchair Traditionalism. After all, the latter does allow for empirical answers to play a role in causal explanations and derivative judgments, which are responses to "important ethical questions." But Doris and Stich have in mind something more. They think that empirical evidence can settle or at least contribute to resolving metaethical debates and weigh directly against normative theories, such as virtue ethics. This is often because existing metaethical and normative theories make unnoticed and unsupported empirical presuppositions.

I will call the type of view that rejects Armchair Traditionalism in this way *Ethical Empiricism*, distinguishing between bold and modest versions of it as follows:

1. Metaethics
   (a) Bold Metaethical Empiricism: questions about the nature of moral judgment or facts can be answered via empirical study.
   (b) Modest Metaethical Empiricism: empirical results are an important source of evidence about the nature of moral judgment or facts.
2. Normative ethics
   (a) Bold Normative Ethical Empiricism: normative ethical questions are empirical questions.
   (b) Modest Normative Ethical Empiricism: empirical results are an important source of evidence about non-derivative moral truths and/or the empirical presuppositions of normative theories.

Both bold and modest versions of Ethical Empiricist theses reject Armchair Traditionalism. An increasing number of moral philosophers, including contributors to this volume, appear to subscribe to Ethical Empiricism at least in its modest forms. This is not surprising, given the general popularity of methodological naturalism in philosophy, and the initial plausibility of the theses. Yet to properly evaluate Ethical Empiricism, we need to look at concrete arguments and see whether they support the methodological claims.

So without further ado, I will begin with some psychological arguments in metaethics, and then examine the use of psychology in normative ethics. I will be making reference to various papers in this volume, but my discussion will range more widely. My conclusions will of necessity be tentative. Even if no sound argument supporting Ethical Empiricism can be found among the existing efforts I consider (and there are many I have no space to address here), there could always be a different one. The field of empirically informed ethics is still young. But it may be that we can draw some general morals from looking at why the existing proposals fail (or succeed).

## 16.2 Empirically Informed Metaethics?

Metaethics asks questions about the nature and status of moral thought and talk: Does it purport to represent moral facts or not—that is, are moral judgments cognitive or non-cognitive states? Are there moral facts, and if so, what kind of facts are they? How, if at all, do we acquire moral knowledge? Are moral demands the demands of reason? What does it take to be a moral agent, or a morally responsible agent? These questions are semantic, ontological, epistemological, and broadly metaphysical or conceptual. Some seem clearly out of reach of empirical science—surely no experiment could settle whether the norms of practical reason and morality coincide. But it is less obvious whether armchair methods suffice for others.

One of the core questions of metaethics, in particular the branch that I like to call philosophical moral psychology, is whether moral thoughts purport to represent the way things are, or whether they are directly action-guiding non-cognitive states, or perhaps some sort of hybrid of cognitive and non-cognitive states.[1] Answers to this question are highly significant for other metaethical issues, such as the nature of moral agency, the function of moral language, and the possibility of moral knowledge. Since this question concerns a crucial feature of moral thought, it is a good test case for the potential relevance of psychological discoveries.

How do we go about answering the question? Consider the traditional armchair argument for non-cognitivism. According to it, when we reflect on moral practice and the distinctive point of moral thinking and language, we discover *a priori* that an intimate link to motivation is essential to moral judgment, since otherwise morality wouldn't be action-guiding in the way it is. This view, which comes in many varieties, is known as *moral judgment internalism*. The next step on the argument is that when we reflect on the nature of psychological states, we learn *a priori* that a mind-to-world direction of fit (tendency of content of the state to match our evidence of the way things are) is essential to belief, and that motivation or action-guiding requires a world-to-mind direction of fit (tendency for the state to move us to change the way things are to match its content) (see e.g. Smith 1987). So, we have an *a priori* argument to the effect that moral judgments cannot be (ordinary) beliefs, and hence consist in some type of non-cognitive or hybrid state. Counterarguments have the same structure—for example, if amoralists, people who make moral judgments without being moved by them, are conceptually possible, the moral judgment internalist premise of the non-cognitivist argument is *a priori* false (and *moral judgment externalism* is true).

This armchair debate has persisted for decades without consensus resolution, although arguably significant advance has been made. The same, of course, could be said about any number of major philosophical debates, so this is not a specific reason to reject the armchair method in moral psychology. But it does provide some motivation to look for an additional source of evidence. I will examine two different attempts to use empirical evidence in resolving the dispute.

### *16.2.1 From Surveys of Ordinary People to Conceptual Truths*

Proponents of the philosophical movement known as experimental philosophy have taken to the streets (or classrooms) to present people with philosophically

[1] When philosophers talk about non-cognitive states, they mean thoughts that do not purport to represent the way things are, and hence cannot be true or false. Paradigmatic examples are desires and affective states. Psychologists often use the term 'cognition' more broadly.

interesting scenarios and elicited judgments (often called 'intuitions') about them. When this method is applied to the case of moral judgment, the argument goes in something like this way, using an argument for internalism as an example:

1. Some philosophical debates concern the extension of ordinary people's concepts, such as the concept of moral judgment.
2. Ordinary people's responses to thought experiments reveal/provide evidence about the extension of the folk concept of moral judgment.
3. The majority of ordinary people's responses are as predicted by moral judgment internalism.
4. Hence, (bold) moral judgment internalism is  true / (modest) there is empirical evidence in favor of moral judgment internalism.

For the purposes of assessing the methodology, it does not matter whether Premise 3 is true (the actual survey results conflict with each other). Let us assume, for the sake of argument, that it is the case. Some philosophers would reject Premise 1, and insist that the philosophical debate is about the *nature* or *essence* of moral judgment, which has nothing to do with our everyday *concept* of moral judgment. Perhaps nothing worth calling an intuition plays a role in philosophical methodology (Williamson 2007; Cappelen 2012). Others argue that even if *intuitions* are crucial to philosophical methodology because, for example, they are a source of evidence about modal facts, an intuition isn't the same thing as a response to a survey. Rather, an intuition is perhaps something like an intellectual appearance or seeming (Bealer 2000; Huemer 2001), or a belief or at least an attraction to assent to a proposition that results from mere adequate understanding (Audi 2004; Sosa 2007). Perhaps, as classical rationalists argued, we can have rational insight into the real essences of things. Surveys plausibly do not tap into intuitions in this sense—there is no telling if people's answers are based on intellectual appearances rather than something else altogether (Bengson 2013).

It may well be that these lines of response are more plausible in some philosophical debates than in others. In any case, I will grant that conceptual analysis does have at least some important role in philosophical theorizing, including in metaethical discussion. This means that Premise 2 is crucial for assessing experimental philosophy. At first sight, it seems obvious that ordinary people's responses to scenarios—for example, confident labeling of a subject's mental state as a moral judgment—is evidence about their concept, given that our grasp of the concept MORAL JUDGMENT to some extent guides the way we categorize things. So should philosophers set fire on their armchairs and run out to check whether people think a person who says that stealing is wrong but is not even slightly motivated to refrain from stealing really makes a moral judgment?

Not so fast. To begin with, consider that different people respond differently, yet seem to share the *same* concept, since they apparently *disagree* about its application. If that is the case, there must be a gap between what the shared folk concept (which is the object of philosophical interest) applies to and the way individual users of the concept classify things (which may in itself be of psychological or sociological interest). There are many mutually compatible explanations for the existence of the gap between the folk concept and the folk's actual classifications. First, as Kripke (1981) emphasized, concepts are *normative*, not descriptive: our concept of addition tells us how we should respond to a calculation task, not how we actually do or are disposed to respond. Sometimes we make mistakes by our own lights—fail to be guided correctly by our own concepts. This may be systematic in some cases, with the result that a majority of people classify things incorrectly. Such tricky or borderline cases are often the most interesting philosophically. Second and related, different people have different levels of *competence* with a concept—some might apply it correctly to paradigm cases, but fare poorly when it comes to the harder ones. The result is that some people's responses may reflect their own shortcomings rather than the folk concept, while others will be more reliable judges.

Third, people's responses might be guided by *non-semantic* considerations. Take the 1868 *Desmond* case discussed by Nadelhoffer (2006). A group of Fenian activists tried to blow up a prison wall to free some comrades, but only succeeded in killing civilians nearby. Clearly, this latter effect was not intended, and probably not even foreseen by the Fenians. Yet when caught, the jury convicted them of murder, which implies intentionality. Nadelhoffer's plausible explanation, supported by his own survey results, is that the jury's willingness to blame the terrorists biased their judgment, leading them to attribute intentionality where none was present—where the folk concept of intentional action doesn't apply.

Fourth, *loose talk* is ubiquitous in non-philosophical contexts. In loose talk, people apply a concept to referents that may fulfill some of the criteria of application but lack some necessary features. For example, people may say "I knew it!" when they've made a lucky guess that turns out to have been correct. Here their belief meets one of the criteria for the application of "knows" (truth) but lacks a necessary condition (non-accidental justification). The same goes arguably for people's willingness to classify a robot that can respond differentially to colors as "seeing" a color (Sytsma and Machery 2010). The robot's circuitry is sensitive to light reflectance (a criterion of seeing) even though it lacks experience with a phenomenal character or the ability to know something (other potentially necessary conditions of seeing), so when people are not particularly interested in speaking literally, and when others can be expected to grasp this, they may well loosely describe the robot as 'seeing red' to convey that it can respond differentially to redness. This is no different from saying that a baby alarm *hears* the baby cry or that the iPad *knows* when its battery is low. We cannot in any of these cases draw conclusions about the concept of seeing or hearing or knowing.[2]

All these caveats mean that ordinary people's responses to cases provide weak evidence about their concepts. The evidence provided by dispassionate armchair reflection or open-minded dialogue will often be stronger (Kauppinen 2007).

[2] Note that I do not claim that the term 'seeing' is ambiguous between an informational and a phenomenal reading. Sytsma and Machery (2010) consider the ambiguity hypothesis, which they regard as *ad hoc* in the absence of an explanation of why the folk would use a different sense than philosophers do, and reject on the basis of their data. The hypothesis that the folk speak more loosely than philosophers do has a high prior probability, so it isn't *ad hoc*.

Whether there is any point in running a survey will depend on the comparative odds of mistakes being made in the armchair or on the streets, which may vary case by case. Here, of course, experimental study can provide evidence one way or another—not about people's concepts, but about how they come to make judgments about certain issues.[3] This type of psychological study will not itself either answer or provide evidence for philosophical questions, but may in principle help identify which responses are good sources of evidence about concepts. As such, it can play a potentially useful auxiliary role in explaining away discrepancies between the folk and philosophers, for example, or even in aetiological debunking of intuitions (see below, Sect. 3.2)—although when it comes to verdicts that have gained broad acceptance among philosophers, a psychologist has a heavy burden of proof to show that they do not reflect conceptual competence.

On the whole, the likelihood that surveys provide useful evidence of folk concepts is low. The odds are that either the outcome is easily anticipated from the armchair, or one or another distorting factor intervenes to produce results that merit no weight in conceptual analysis. Thus, even if this kind of experimental method has some place in the philosophical toolkit, it will be marginal.

### *16.2.2 From Best Explanation of Data to the Nature of Moral Judgment*

A very different experimental approach to metaethics takes its departure from the thought that moral judgment is a natural kind—the sort of thing whose nature or essence can be discovered *a posteriori* by looking at what actually happens in people's minds when they make moral judgments. I will focus on Jesse Prinz's (2007a, b) version of this kind of argument. As I construe it, it involves an inference to the best explanation of observations:

1. Moral judgment is a natural kind whose nature can be found by examining what happens in actual paradigm cases.
2. Psychological and neuroimaging data show, among other things, that manipulating emotions changes moral judgment, emotional activation coincides with moral judgment, and emotional deficits lead to deficits in moral judgment.
3. The best explanation of the data is that moral judgments consist in emotions, which are the best fit for the natural kind that constitutes moral judgment.
4. Hence, moral judgments consist in emotions.

Prinz is clearly committed to something like the first premise, given that he says that the way to avoid the 'impasse' resulting from conflicting intuitions is "turning to psychology and neuroscience, which give us techniques for investigating what goes on in the mind when people are actually engaged in moral evaluation"

[3] This more modest goal is sometimes emphasized by Joshua Knobe (e.g. Knobe 2007).

(Prinz 2009, 702). Is this true? That depends in part on what natural kinds are. There are many ways to think about them. According to one prominent view, deriving from Kripke (1980) and Putnam's (1975) work in the philosophy of language, natural kinds are roughly speaking *a posteriori discoverable microstructural essences*. Water is a paradigm case here: since it turns out, *a posteriori*, that the *actual* watery stuff around us is $H_2O$, water is *necessarily* $H_2O$. Roughly, water is *that* stuff there in the rivers and lakes and rain (the term 'water' is a rigid designator); anything that is not that very substance, however similar in superficial properties, isn't water. Hence, the XYZ on Putnam's Twin Earth isn't water.

Another well-known contender is Richard Boyd's view. According to Boyd, when we look for a definition of a natural kind K, we're looking for those commonalities in the causal profiles of the things we classify as Ks that explain our explanatory and inductive success with respect to our term for K (Boyd 2010, 215). Such projectable patterns are *homeostatic property clusters*—sets of properties that reliably co-occur in virtue of some law-like connection, either because the presence of some properties favors the presences of others or because some underlying mechanism favors co-presence (Boyd 1999). According to Boyd's 'accommodationist' semantics, natural kind terms refer to the property clusters that causally regulate their use, even if people have false beliefs about their nature (so that alchemists, for example, succeed in talking about mercury).

There is good reason to think that moral judgments do not form a natural kind in the microstructural sense. We just do not think of moral judgments as psychological states like *that* (pointing to some paradigmatic case of moral judgment), so that nothing that is constituted by the same pattern of brain activation or mental states is a moral judgment. Rather, moral judgment seems to be a *functional* kind: any psychological state that plays a certain functional role is a moral judgment, however it is realized in the mind and brain. In this respect, moral judgments are more like chairs than like water: even if all actual chairs happened to be made of plastic, being made of plastic would be an accidental property of chairs. What makes something a chair is that it's an artifact with a certain practical function. Similarly, even if Twin Earthers have a very different kind of brain and mind from ours, as long as they make judgments that are categorical (apply to agents regardless of their desires or interests), presumptively universalizable (apply to all non-morally similar cases), have felt intersubjective authority, and are somehow linked to non-self-interested sanctioning behavior, to take a few relatively uncontroversial marks of moral judgment, they do make moral judgments.[4] An indication of this is that it is possible for us to *disagree* with them about moral matters, which would not be the case if they were incapable of moral thoughts.

Here is another way to make the case that the concept of moral judgment is not a natural kind concept. This line of argument does not assume that essence must be microstructural, or that MORAL JUDGMENT is necessarily a functional concept (I will use small caps to indicate I'm talking about a concept). Supposed it turned out that

[4] There is now some controversy about this; see Sinnott-Armstrong (2008b), Sinnott-Armstrong & Wheatley (2012) for an argument in favour of disunity of moral judgment.

the psychological states we actually identify as moral judgments only motivate people by way of a desire to look good in the eyes of others. Gunnar Björnsson and Ragnar Francén Olinder (2013), on whose work I draw here, dub this the Cynical Hypothesis. Would its truth mean that internalism is false—or that what we thought were moral judgments were not moral judgments after all? That depends on what kind of concept MORAL JUDGMENT is. Björnsson and Francén suggest it is parallel to TIGER. Take Kripke's (1980) example of the putative conceptual truth that TIGERS ARE MAMMALS. What if it turned out that all animals we actually identify as tigers, or at least the paradigmatic 'tigers', are reptiles? According to Kripke, it would not follow that there are no tigers. Rather, it would turn out that we were wrong about the nature of tigers. Our concept of a tiger is a concept of an animal like *those* (demonstrating paradigm examples of the animal we actually identify as a tiger), whatever kind of animal it turns out to be. (Perhaps tigers need not even be animals.) Björnsson and Francén claim the same goes for moral judgments. If it turns out the Cynical Hypothesis is true, it is not that we don't make any moral judgments, but that we mistook a common correlation between judgment and motivation as a conceptual truth. As they say:

> The cynical hypothesis concerns the actual states of mind that we paradigmatically think of as moral opinions, and it allows that they have almost all the characteristics we normally ascribe to them. They are still categorical, based on familiar moral considerations (e.g. wellbeing, autonomy and respect for rights), often in competition with our prudential considerations, invoked to settle practical issues, and expressed to condemn behaviour near and far. Moreover, people are still affected by moral considerations, some more than others. What is different is just that moral opinions affect action less directly than most of us think. (Björnsson and Francén Olinder 2013, 8)

The other option is that if the Cynical Hypothesis is true, no one makes moral judgments. This parallels the case of WITCH. It is evidently possible for paradigmatic 'witches' or all people we identify as witches to fail to be witches. Why so? Because having supernatural powers as a result of an alliance with an evil it is part of our concept of a witch, and no one has such powers.[5] Why is it part of WITCH? The appealing answer Björnsson and Francén suggest is roughly that there is a certain interest of ours that the concept serves (or served). This is plausibly not just the purpose for which the concept was introduced, but, let us say, the purpose that sustains its use. Having supernatural powers is essential to being a witch, because the point of talking about witches is to identify those with supernatural powers as a result of an alliance with evil. If it turns out no one actually identified as a witch has magical powers, it is not that we were wrong about witches, but that there are no witches at all.

The key question, then, is what interest our concept of moral judgment serves. Would there be a point in attributing people moral judgments if the Cynical

[5] An anonymous referee pointed out that people who self-identify as witches do not think being a witch involves having supernatural powers. Alas, I do not think that believing that one is a witch gives one any special conceptual insight. Indeed, thinking that you are a witch without thinking that you have supernatural powers shows a rather poor grasp of the concept of a witch.

Hypothesis turned out to be true? The internalist will respond: no, it *is* an essential part of the point of talking about moral judgment to distinguish between people who are motivated by what they think is right, as opposed to people who are motivated only by what others think about them. Consider this: why would we introduce in our language an expression for "Martina thinks that X is morally wrong?". Maybe Martina engages in punishing behavior for X, where X involves harming a third party, for example. But why—only because she would be otherwise punished or thought badly of by third parties, or because she thinks X is wrong? The internalist may note that we talk about social norms in the former case. Social norms, after all, overlap with moral norms, and can play the roles that Björnsson and Francén list in the quotation above. They can be categorical (as Philippa Foot (1972) noted, even the norms of etiquette are), promote autonomy, compete with prudential considerations, and so on. For the internalist, the *crucial* difference between moral judgment and socially normative judgment is precisely that the former motivates without regard for and sometimes against what others think. Externalists, too, think that moral judgments motivate by way of something like a desire to do the right thing, and not the (cynical) desire to look good in the eyes of others. So if the Cynical Hypothesis is true and it turns out that states of mind actually identified as a moral judgment only motivate by way of desire to please others, it is not that we were wrong about the nature of moral judgment, but that there are no moral judgments.[6] This, of course, would be a startling discovery, but about human beings rather than about moral judgment.

I do not think this issue can be definitively settled here. All I want to say is that the internalist rejoinder is plausible, and it if is true, it is not an empirical possibility that moral judgments fail to motivate—the empirical possibility is merely that what we actually identify as paradigm cases of moral judgment are not such. We cannot get at the nature of moral judgments by looking at states actually *believed* to be moral judgments, since it may turn out that they are not moral judgments after all. What settles this is an *a priori* investigation into the point of using the relevant concepts. What Björnsson and Francén successfully establish is that *if* that inquiry goes one way, MORAL JUDGMENT is a natural kind concept, and the truth of internalism turns out to be an empirical question. However, I believe that reflection on the point of using the concept supports the opposite conclusion in this case.[7]

What about the Boydian conception of natural kinds? It does appear to be the case that moral thoughts can play a role in explanation and prediction—for

[6] Consider also a Supercynical Hypothesis: not only the states of mind we actually identify as moral judgments not intrinsically motivating, but they are also not in fact based on considerations like rights and well-being, but only what agents unconsciously take to be in their self-interest. Would we still feel the pressure to say that there are moral judgments, but we are wrong about their nature? Why not, if moral judgment is a natural kind whose nature we can identify *a posteriori*?

[7] Mark Alfano pointed out that there is a further possibility I do not consider in the text: reforming our concept as a result of an empirical discovery. I agree that this is a significant option. It might make more sense to modify our concept rather than stop using it, if the world does not cooperate, especially if there is another natural kind in the Boydian sense in the vicinity. Whether this is the case for philosophically interesting concepts remains to be seen.

example, people tend to do what they genuinely think they ought to do, and people tend to think an action is wrong when it involves hurting people they care about. If that's all it takes to form a natural kind, then surely moral judgment is one. One way to see Prinz's argument is as making the case that this natural kind is *constituted* by another natural kind, namely sentiments of approbation and disapprobation. This would explain the empirical observations about emotion, as well as at least many of the other regularities we observe anyway, such as a defeasible link to motivation and tendency for negative judgment when innocent people are harmed, given that both are features of emotional responses. So there is some support for Prinz's constitution claim.

This argument relies crucially on the assumption that the empirical observations (and conceptual platitudes) are *best explained* by taking emotions of approbation and disapprobation to constitute moral judgment. It is thus open to challenge that there is an even better explanation available. I have elsewhere proposed that there is a better candidate: moral *intuition* (Kauppinen forthcoming). As noted above, there is controversy about the nature of intuitions in general, but there is much to be said in favor of thinking of intuitions as *intellectual appearances*: spontaneous and compelling non-doxastic seemings that result from merely thinking about (as opposed to perceiving or remembering) something (see e.g. Huemer 2001). What I have argued is that emotional manifestations of moral sentiments can also constitute intellectual appearances in this sense: when we merely think about taking advantage of someone's disability or disrespecting a national hero, we may have a spontaneous and compelling emotional experience that manifests our disapprobation and presents the action as morally wrong. Such sentimental intuitions can both cause and justify belief (just in the same defeasible way as other intellectual or perceptual appearances do) and motivate us to act. I emphasize that not all moral judgments are based on intuitions: we may also engage in reasoning or simply be disposed to apply rules. This is important, because on my picture, unlike on Prinz's, it is possible (and indeed common) for people to make moral judgments without having emotional responses.

If it is indeed possible to judge without emotion, radical sentimentalist views of Prinz's type are wrong. The crucial test cases here are people with emotional deficits. The most discussed case is that of *psychopaths*. Prinz argues that they can have moral thoughts only *deferentially*, by reference to what other, emotionally typical people regard as right or wrong (e.g. this volume, Chap. 6, p. 101). Yet it is easy enough to imagine a psychopath, or some other emotionally deficient character, making a non-deferential moral judgment and thinking, for example, that everyone else is making a moral mistake. And we ourselves seem to make entirely unsentimental judgments much of the time—although we should take this data point with a grain of salt, given the limits of introspection. Further, *a priori* support comes from considering the conceptual possibility of amoralists, subjects who make moral judgments without any motivation. Insofar as amoralists are possible, there is little reason to think that judgments are constituted by inherently motivating states like the emotions. So, once we distinguish between moral appearances (intuitions) and beliefs (judgments), the best explanation of both the empirical data and conceptual platitudes is that moral intuitions rather than judgments are sentimental in nature. Premise 3 of the Prinz-style argument is thus false.

So, in short, given that moral judgments do not form a natural kind in the Kripkean sense (MORAL JUDGMENT isn't a natural kind concept), we cannot investigate their nature by observing 'what happens in the head' in the actual paradigm cases. Even if there are natural kinds in the property cluster sense associated with moral judgment, we need to engage in *a priori* reflection to figure out whether they constitute moral judgment or some other associated state. In this kind of reflection we draw on conceptual connections that are *not* discovered *a posteriori*, for example on views about the connection between moral judgment and motivation. Since such key features of moral thoughts are assumed rather than discovered in this empirically informed inquiry, its metaethical scope and significance are limited.

## 16.3 Empirically Informed Normative Ethics?

As a reminder, these are the Ethical Empiricist theses about normative ethics I want to look at next:

Bold version: normative ethical questions are empirical questions.
Modest version: empirical results are an important source of evidence about nonderivative moral truths and/or the empirical presuppositions of normative theories.

Whatever the status of metaethics, both bold and modest ethical empiricists face the challenge of justifying the move from an 'is' to an 'ought.' This is something that has been attempted in a number of ways. In this section, I will examine one bold and three modest attempts to make use of psychological evidence in normative ethics.

### *16.3.1 Via Reduction to Normative Conclusions*

A radical way of closing the is-ought gap is proposed by Prinz (2007b). As a radical naturalist, he believes that all facts are natural, so "moral facts are natural facts, if they are facts at all" (Prinz 2007b, 3). We can derive moral conclusions from facts whose truth can (at least in principle) be empirically established. To his credit, Prinz lays his cards on the table and gives a very clear account of how he believes this can be done. His example features a character called Smith, whose obligation to give to charity, Prinz claims, is entailed by a set of non-moral premises. Here is his argument (Prinz 2007b, 5):

1. Smith has an obligation to give to charity if 'Smith ought to give to charity' is true.
2. 'Smith ought to give to charity' is true, if the word 'ought' expresses a concept that applies to Smith's relationship to giving to charity.

3. The word ‘ought’ expresses a prescriptive sentiment.
4. Smith has a prescriptive sentiment towards giving to charity.
5. Thus, the sentence ‘Smith ought to give to charity’ is true. (2, 3, 4)
6. Thus, Smith has an obligation to give to charity. (1, 5)

The first two premises are surely uncontroversial (provided 2 is read charitably), regardless of what theory of truth is correct, and so is the step from 5 to 6. Premise 4 is a factual stipulation. That leaves Premise 3. Whether it is true is a metaethical question, which I’ve already argued cannot be settled by empirical study. If that is the case, it’s already sufficient to render the derivation non-empirical (while still preserving its status as an inference from an is to an ought). But suppose Premise 3 is true. Does the conclusion then follow? No, because 5 does not follow from 3 and 4.

Why is this the case? Well, if the word ‘ought’ expresses a prescriptive sentiment, it is surely the *speaker*’*s* prescriptive sentiment. If I say you ought to clean your room, I am expressing, at most, my own sentiment in favor of your cleaning the room. Maybe you do not share that sentiment. No matter. On Prinz’s semantics, according to which concepts are psychological entities such as sentiments, my utterance of “You ought to clean your room” still expresses a concept that applies to your cleaning your room. By parallel reasoning, in Prinz’s example, it does not matter to the truth of “Smith ought to give to charity” whether *Smith* has a prescriptive sentiment towards giving to charity. Premise 4 is irrelevant.

But whose prescriptive sentiment, then, makes Premise 5 true, if we grant Prinz the rest of his premises? That is a tricky question. Consider first a semantic relativist variant, 5′:

5′. Thus, the sentence ‘Smith ought to give to charity’ is *true-for-S*.

To reach *that* conclusion, premise 4 would have to be

4′. S has a prescriptive sentiment towards giving to charity.

(Here S may or may not be identical with Smith.) As a relativist, Prinz might be sympathetic to this move. To be sure, it is not clear whether we can make sense of relative truth, though valiant efforts have been made (e.g. MacFarlane 2005). But let us suppose we can. Have we then accomplished the goal of deriving an ought from an is? No, because 6 doesn’t follow from 5′ together with 1. 1, the uncontroversial disquotational principle, appeals to *unrelativized* truth. But it does not follow from the *truth-for-S* of “Smith ought to give to charity” that Smith ought to give to charity. After all, whether Smith ought to give to charity is not a perspective-relative fact. Also, given different sentiments on part of some $S_2$, “It is not the case the Smith ought to give to charity” could be true-for-$S_2$, so that applying disquotation would give rise to (ontological) contradiction—it being the case both that Smith ought and ought not give to charity.

Premise 6 would, to be sure, follow from the original 5 and 1. But what would make the original 5 non-relatively true, assuming for the sake of argument that ‘ought’ expresses a prescriptive sentiment? The only plausible candidate is that it is *correct* or *appropriate* to have a prescriptive sentiment towards giving to charity. But that is not an empirical fact (to assume otherwise would be to beg the question—the argument is precisely meant to establish that normative facts are empirical). Instead, it is itself a *normative* fact. So, in short, Prinz's argument is either invalid (because Premise 5 doesn't follow from 3 to 4, and if 4 and 5 are replaced by 4′ and 5′, the conclusion does not follow), or involves an 'ought' premise. I do not think there is any way to fix the argument. Bold versions of normative ethical empiricism have little hope of success. But that leaves a number of modest theses that might be viable.

### *16.3.2 Via Aetiological Debunking to Normative Conclusions*

A very different kind of normative ethical empiricist argument has received a lot of attention in recent years. It aims to show that key non-consequentialist beliefs are best explained as the result of emotional reactions, and that their aetiology renders them untrustworthy. Given that we should not base our normative theories on or accommodate untrustworthy beliefs, this shows that we should reject nonconsequentialist ethics. The general form of the argument is the following:

**Aetiological Debunking Argument**

1. Empirical investigation shows that belief that p results from process X.
2. Process X does not confer justification to/undermines the justification of beliefs it gives rise to.
3. Hence, empirical investigation undermines the justification for belief that p.

As a starting point, everyone but the most hardcore skeptic agrees that some causal processes that result in beliefs are justification-conferring or transmitting. For example, competent logical deduction transmits justification from belief in premises to belief in conclusion. But many of our beliefs do not result from any kind of reasoning. Perceptual beliefs are one paradigm case of such *non-inferential beliefs*. Some say that their justification is exclusively a matter of *coherence*, their fit together with the rest of our beliefs. But pure coherentism seems to sell perceptual beliefs short. Surely their justification has something to do with their causal history as well. Indeed, it seems that perceptual beliefs can be justified in spite of clashing with our prior beliefs. In the absence of a reason to doubt, if I see my Head of Department peel off his skin and reveal the shiny robotic machinery underneath, I should revise a lot of my beliefs rather than reject the poorly cohering perception.[8]

[8] Granted, in extreme cases like this there generally is a reason to doubt and check the initial appearance, as Markus Christen pointed out to me. Nevertheless, perceptions do start out with initial credibility independent of coherence.

To stick with the case of perception, why are (some) non-inferential perceptual beliefs justified? I will focus on just two influential schools of thought. According to one *externalist* view, non-inferential beliefs are justified when they result from a causal process that *reliably tracks the truth*, even if the believer is unaware of this (Goldman 1979; Nozick 1981). According to a recently popular *internalist* view I will call epistemic liberalism, non-inferential beliefs are justified when they are based on *appearances there is no sufficient reason to doubt* (Pryor 2000; Bengson 2010). Internalists often hold that justification has to do with epistemic praise- or blameworthiness, and that there is no reason to blame someone who believes things to be the way they seem to be, if he or she has no reason to doubt the appearances. These two views of justification give rise to different criteria for evaluating processes that result in non-inferential beliefs: they fail to confer justification if they do not reliably track the truth or if they do not involve appearances beyond reasonable doubt.

The specific aetiological debunking argument made by Joshua Greene (2008) and Peter Singer (2005) has this form:

*The A Posteriori Argument for Consequentialism*

1. Empirical investigation shows that nonconsequentialist moral intuitions* are proximately caused by emotional reactions.
2. Emotional reactions do not confer justification to the beliefs they give rise to.
3. So, empirical investigation undermines the justification of nonconsequentialist moral intuitions*.
4. Nonconsequentialist moral theory rests crucially on nonconsequentialist intuitions*.
5. So, nonconsequentialist moral theory is unsupported by evidence.

(In this argument, 'intuitions' are taken to be spontaneous, non-inferential beliefs rather than intellectual appearances. Since precision is important here, I'll use 'intuition*' to refer to such beliefs to distinguish them from intuitions proper.) To begin with Premise 1, in the background of Greene and his colleagues' argument is a general *Dual Process Model* of the mind. Roughly speaking, the model distinguishes between System 1—automatic, uncontrolled, fast, associative, and often affective processes functioning below the level of consciousness—and System 2, which is conscious, slow, effortful, and capable of reasoning (for a general picture, see Sloman 1996; Kahneman 2011; see also Chap. 7 by Tanner and Christen, this volume). The key empirical data suggest that nonconsequentialist judgments selectively involve the activation of areas of the brain associated with emotion, involve faster reaction times, and go missing in subjects who suffer from emotional defects (Greene et al. 2001, 2009). Consequentialist judgments, in contrast, appear to engage System 2 reasoning. These results and interpretations have been challenged. For example, McGuire et al. (2009) argue that there is no difference between consequentialist and nonconsequentialist responses in reaction times, and Klein (2011) argues that the fMRI evidence does not in fact suggest selective emotional activation in nonconsequentialist responses. And finally, perhaps most decisively, Kahane et al. (2012) find that in cases in which the nonconsequentialist response is counterintuitive (for example, it calls for speaking the truth to the murderer at the door), it is nonconsequentialist responses that take more conscious effort, suggesting that what engages System 2 is overriding intuitions, not consequentialist rationality.

There is thus plenty of reason to doubt the empirical premise of the A Posteriori Argument for Consequentialism. But suppose there is some truth in it—that emotional responses play a different role in accounting for nonconsequentialist beliefs than consequentialist ones, at least in the Trolley Cases. Premise 2 then becomes crucial. Why does not the fact that thinking about being pushed off a bridge or thinking about pushing someone off a bridge in order to save more people feels bad provide some justification for believing that it is morally wrong? Although some of the things that Greene says suggest that the problem is that it is the mere fact that emotions are involved undermines justification, his considered position is that emotions are *responsive to morally irrelevant factors* (and therefore, presumably, fail to track moral truth). This, of course, breaks down to two claims: emotions are responsive to factors x, y, and z, say, and x, y, and z are morally irrelevant. The first claim is clearly empirical. The second claim, however, is not empirical, as critics like Selim Berker (2009), have pointed out. Its truth must be established the same way as the truth of any other moral claim, perhaps involving appeal to substantive (and controversial) moral intuitions.

But Greene is surely right in responding that while this is true, the scientific data still does important work in the normative argument (Greene manuscript, 9). It may, after all, be a surprising discovery that our beliefs track features x, y, and z. We may, on reflection, agree that x, y, and z are morally irrelevant. In Greene's case, the factor he sees as crucial to explaining people's responses is the *use of personal force*. As he notes, it is not question-begging for a consequentialist to take this to be morally irrelevant: "Whether your normative proclivities are consequentialist, deontological, or otherwise, it's hard for you to argue that personal force is morally relevant." (Greene manuscript, 17) It is thus very plausible that psychological processes that track the use of personal force do not track moral truth, and the beliefs that are their outputs lack justification in the externalist sense. (Insofar as a subject is *aware* of what underlies her responses, she presumably lacks justification in the internalist sense as well.)

In support of Premise 2, Greene (manuscript) further argues that it is likely that emotions will be responsive to irrelevant factors, especially in novel situations. The distal explanation of why we have particular affective responses is that they have been, on the whole, fitness-enhancing in the course of human evaluation. It pays off, as a rule, for us to be afraid of big things moving fast toward us, since most such things were (and are) dangerous. But this response will sometimes misfire, especially in evolutionarily novel situations (the subway train will not leap off its track to pounce on us). Similarly, the Greene/Singer hypothesis is that evolution has favored the development of negative emotions to using up close and personal violence. Such innate aversion is fitness-enhancing for some reason (presumably it reduces interpersonal conflict). Violence (or assistance) at a distance, however, was not an issue during the era of human evolutionary adaptation. Consequently, our automatic, 'point-and-shoot' moral emotions are likely to misfire in modern, complex, or unusual situations—to fail to respond to morally relevant factors.

This is an impressive line of argument. If the aetiology of beliefs is relevant to their justificatory status, then surely empirical study of the aetiology can in principle reveal that they lack justification. But I do want to raise three concerns with Greene's case: not all emotions are created equal; intuitions aren't so easily done away with; and what counts at the end of the day is not whether particular individuals are justified but whether justification is available for nonconsequentialist beliefs. Before I go into these, however, I want to register some doubts about an approach that has gained popularity recently. According to this type of response, emotional intuitions can be reliably truth-tracking in just the same way as *expert intuitions** in general (see Chap. 7 by Tanner and Christen, this volume; Chap. 11, by Musschenga, this volume; Chap. 13 by Narvaez and Lapsley, this volume; Allman and Woodward 2008). Expert intuitions* are, roughly, spontaneous judgments that result from automatic, System 1 processes that respond to environmental cues that the subject is not consciously aware of, but are nevertheless reliable. Paradigmatic examples are quick situational assessments by chess masters and experienced nurses or firemen: without knowing just why, the fireman feels that the building is about to collapse and reacts to save himself at just the right time. If moral intuitions* of at least some people were of this type, there would be no reason to suspect them.

Alas, contrary to optimists, they cannot be. As an authoritative recent overview (Kahneman and Klein 2009) argues, there are two conditions for the development of intuitive expertise or implicit learning. First, the environment must exhibit regularities that the associative System 1 can latch onto. This may or may not be the case for morality in general, but surely will not be for outlandish philosophical thought experiments. Most importantly, however, training System 1 requires "prolonged practice and feedback that is both rapid and unequivocal" (Kahneman and Klein 2009: 524). A nurse who diagnoses and treats a baby will typically be able to check whether the baby's condition is improving (temperature returning to normal etc.), and thus gets feedback on the correctness of the diagnosis. There is nothing analogous to this in the case of moral judgment. Even if there is a recurring type of moral problem, there's no rapid and unequivocal indication that a subject is judgment is on the right track. If you judge that abortion is wrong even if it is not and act on your belief, there is no negative feedback that results simply from your having made a moral mistake. (The only reliable negative feedback you will get for acting on a moral judgment is from people who disagree with you, but that is not an indication that you are wrong.) So we cannot train our intuitive system to respond to moral truths in the same way we can train it to respond to truths about good chess moves or ill infants. The expertise defense of moral intuitions* is unsuccessful.[9]

[9] To be sure, I do not mean to deny that there can be moral expertise in some meaningful sense—some people are better at articulating principles, more consistent, better informed about pertinent non-moral facts, and so on. Perhaps it is even the case that their judgments should be privileged in reflective equilibrium, as Musschenga argues (Chap. 11 this volume). But nonconsequentialists cannot defend intuitions* on these grounds.

If the expertise defense will not work, how can nonconsequentialists respond to Greene's challenge? To begin with the first option I mentioned, Greene stakes a bold claim about nonconsequentialist intuitions*: "All of the factors that push us away from consequentialism will, once brought into the light, turn out to be things that we will all regard as morally irrelevant." (manuscript, 21) So when we trace down the aetiology of any nonconsequentialist intuition*, we always hit an affective reaction that is caused by a factor that is, on reflection, morally irrelevant. However, it is one thing to say that some morally relevant emotions are triggered by simulating the use of personal force or some other morally irrelevant factor, and another to say that *all* are. For example, it is extremely plausible that we have a negative emotional response, such as resentment, to being used as a mere means by someone else, as well as a weaker sympathetic response to imagining ourselves in such a position. Such reactions are also almost certainly fitness-enhancing, at least in the personal case—they motivate retaliation and decrease the likelihood of being exploited in the future. Being used as a mere means, in turn, is not uncontroversially a morally irrelevant factor—to claim otherwise is to beg the question in favor of consequentialism. This means that at least some emotions are responses to factors that are plausibly morally relevant. Note also that there is a long tradition of sentimentalist ethics arguing that such reactions need not be rooted in an egocentric perspective, but can also be felt from what Hume called the 'Common Point of View' and Adam Smith called the impartial spectator's perspective. I argue elsewhere that precisely such impartially empathetic emotional responses constitute canonical moral appearances or intuitions (Kauppinen forthcoming).

Sentiments felt from the Common Point of View are far from the kind of automatic gut reactions that Greene discusses. They are not or need not be quick and unreflective, evolutionary fitness-enhancing, or responsive to features that are uncontroversially morally irrelevant. So insofar as nonconsequentialist moral judgments are based on *that* kind of emotional intuition, there is no obvious reason to think they lack justification. From this perspective, Greene's problem is that he works with a palette that is too narrow: it is either reasoning or gut reaction, and nothing in between.

Of course, it remains to be shown that at least some nonconsequentialist judgments result from the better kind of emotional response. The current data does not settle the issue even concerning the Trolley Cases. Although people are more likely to condemn the agent who pushes a fat man down (where there is both personal force and use as a means) than an agent who drops the fat man through a trapdoor (where there is use as a means but no personal force), they are nevertheless more likely to condemn the latter than an agent in the standard Switch cases (where there is neither personal force nor use as a means) (see Greene et al. 2009). So *use as a means* has an effect independently of personal force. Indeed, one possible explanation for why the use of personal force plays a role may be that it raises the *salience* of the use as mere means (cf. Chap. 6 by Prinz, this volume, p. 106). Moreover, many philosophers report the intuition that the trapdoor drop is wrong, as well as intuitions about other more fine-grained scenarios. These are unlikely to be mere gut reactions, since they are reflectively stable. But they may well be the good kind of sentimental intuitions I talk about. We may not be able to assess their *reliability* in a non-circular fashion (we will have to assume that using someone as a mere means is wrong, for example), but we can at least say that they're *moral appearances* we have been given no reason to doubt.[10]

The second problem is that reliance on intuitions may be unavoidable. Greene insists that in the psychological sense of "intuition" (by which he means judgment resulting from unconscious, automatic process), "Consequentialism can do just fine without intuitions" (manuscript, 20). But this seems inconsistent with Greene's own acknowledgement that the source of evidence for the moral irrelevance of the use of personal force is "substantive moral intuitions" (manuscript, 7), unless of course the substantive moral intuitions* are not intuitions in the psychological sense. But consequentialism does seem to rely on precisely the same sort of intuitions* (in the psychological sense) as nonconsequentialism. For example, we judge that in Trolley Cases, the "body count" is not morally irrelevant (for consequentialists, it is the only relevant feature). But why? Is it not also an evolved emotional reaction to prefer fewer deaths to more deaths? Surely it is. But if point-and-shoot emotions are unreliable for principled reasons, then so is the core utilitarian intuition*. If the positive response to maximizing is what I have called the good kind of emotional intuition—which I think is likely—then it does have justificatory force, but so do at least some nonconsequentialist intuitions. There is no dialectical advantage here for consequentialism.

The third and final point is that for some purposes, crucially including the choice of which normative theory to accept, the justificatory status of particular individual beliefs does not matter. Those who accept Premise 4 of the A Posteriori Argument for Consequentialism may grant that most people's nonconsequentialist beliefs are based on knee-jerk reactions that undermine their justification, while insisting that genuine intuitive propositional justification is *available* for nonconsequentialist beliefs. That is all that is needed to justify nonconsequentialist theory. Some Kantian nonconsequentialists reject the premise altogether (e.g. Wood 2011). If there is rational justification available for nonconsequentialist beliefs, it again does not matter if *most people* believe the right thing for the wrong reasons. Suppose, for a parallel, that most people believed the Earth is round because a holy book written thousands of years ago happened to say so, without any scientific evidence. That would hardly be relevant to whether *I* or the scientific community in general should accept or reject that the Earth is round. Similarly, Premise 5 does not follow even if people in general lack justification for nonconsequentialist beliefs.

In short, although it is in principle possible that empirical evidence concerning aetiology would undermine the justification of some moral beliefs, the path is far from straightforward. Merely showing that some judgments are intuitive does not

[10] I argue elsewhere that we do have a non-question-begging way of evaluating whether certain kinds of intuitions are trustworthy. This involves appealing to the practical function of making moral judgments, roughly making peaceful social relations possible without a Hobbesian sovereign ruling by force, and noting that being guided by intuitions felt from the Common Point of View is reliably conducive to that goal.

suffice, and for some crucial purposes, such as choice between moral theories, it does not even matter whether most people are justified in believing one way or another.

### *16.3.3 Via Ethical Conservatism to Normative Conclusions*

Shaun Nichols, Mark Timmons, and Theresa Lopez develop a novel modest ethical empiricist argument developed in their contribution to this volume. They argue, first, that many of our central ethical commitments cannot be rationally justified, but result from "a-rational and a-reliable emotional processes" (this volume, Chap. 9, p. 160). But some of such commitments nevertheless have normative authority, which presumably entails that the subjects are justified in believing in their contents. This seems to be the structure of their argument:

1. Entrenched ethical commitments have normative authority in spite of resulting from non-rational and non-truth-tracking emotional processes (Ethical Conservatism)
2. Empirical study can identify which commitments are entrenched.
3. Hence, empirical study can identify which ethical commitments have normative authority.

If empirical study can establish which commitments have normative authority, it surely has more the marginal significance for ethics. So this is an interesting new line of argument.

For a commitment to be *entrenched* is for it to be non-inferential and the result of natural human emotional reactions (or at least resonate with such reactions). It seems plausible that empirical study can indeed establish which commitments are entrenched in this sense, as Premise 2 says, and do so better than armchair reflection. Nichols, Timmons, and Lopez provide an example of how to do it with their studies of outcome-dependent blame, which suggest that even if intention and reasons for action are held fixed, people regard an agent as more blameworthy if the outcome is bad, as long as the agent has been negligent. For my purposes, the details and the soundness of this argument do not matter.

The definition of an entrenched commitment appeals to natural human emotional reactions. I take it that 'natural' here means being part of the normal human biological makeup. A number of contributors to this volume argue, in line with much recent biological research (e.g. de Waal 1996), that some morally relevant emotions are indeed natural in this sense. For example, Van Schaik et al. (Chap. 4, this volume) note that humans, unlike other primates, engage in prosocial behaviors not only reactively—in response to need, proximity, or the presence of an audience—but also proactively, as seen in our tendency to cooperate and share in economic games. Why? Crudely, as the kind of foragers we are, we have to cooperate with each other to survive. As cooperative breeders, we have a tendency to respond to need and conform to expectations; as cooperative hunters, we also have a tendency to match rewards with contributions and build a reputation as reliable reciprocators. Van Schaik et al. hypothesize that these four psychological elements—sympathy, wish to conform, sense of fairness, and concern with reputation—are "the major components of human moral psychology, upon which our reflective morality is built" (this volume, Chap. 4, p. 77). They suggest that moral emotions are "the subjective side of the evolved proximate regulators of human cooperation" (p. 77); see also Naves de Brito (Chap. 3, this volume). They are likely to emerge early and cross-culturally, and will be to an extent independent of conscious control.

As Jesse Prinz (Chap. 6, this volume) points out, even if morally relevant emotions are natural in this sense, it does not mean that our capacity to make moral judgments or tendency to adopt certain moral rules is an evolutionary adaptation. After all, other species that have similar responses and behaviors (see Chap. 5 by Brosnan, this volume) plausibly do not make moral judgments. Prinz's suggestion is that the human capacity to make moral judgments is an evolutionary byproduct of putting together capacities that are adaptations for other purposes, including imitation and capacity for abstract thought in addition to prosocial and reactive emotions. Support for this hypothesis can also be found in neuroscience, if, as Prehn and Heekeren (Chap. 8, this volume) argue, "the "moral brain" can be broken up into several modules whose functions originally have nothing to do with morality (emotion, social cognition, cognitive control, etc.)." (p. 156)

Biological considerations thus support the hypothesis that some moral commitments are entrenched, and indeed provide clues about which commitments are likely to be such. I am not going to take issue with the psychological part of Nichols, Timmons, and Lopez's Chap. 9 regarding which commitments are entrenched. The important question concerns the *epistemic standing* of entrenched commitments. Precisely what does normative authority mean in this context, and why should entrenched commitments have it? To begin with the former, the parallels that Nichols, Timmons, and Lopez draw between entrenched commitments and other beliefs suggest that they think there is *no reason to suspend* beliefs that have normative authority. This may or may not mean that the beliefs are *justified*—perhaps there are reasons not to suspend beliefs that are independent of their justification. Unfortunately, the epistemic part of the paper is extremely sketchy, so it is not possible to determine what the exact view is. In any case, at the end of the paper, Nichols, Timmons, and Lopez offer a further suggestion: some commitments may be entrenched yet biased, in which case they lack normative authority. They argue that bias can be exposed by seeing "whether people withdraw their judgments under full information" (p. 173).

Why should we not suspend entrenched commitments, even if they are not truth-tracking, and even if we know this? Nichols, Timmons, and Lopez offer two suggestions. The first appeals to the *undesirable consequences* of suspending entrenched commitments: "If we give up all of the ethical judgments that critically depend on our a-rational and a-reliable processes, then we might well be left with an ethical world view more barren than almost anyone is willing to accept." (p. 160). This appears to suggest a *pragmatic and non-epistemic* reason for maintaining entrenched commitments: they are not epistemically justified, but if we give them up, we are left with a barely recognizable ethical outlook, which is a bad thing (at least from our current perspective). How dramatic the change would be depends on how important entrenched commitments are to our actual ethical outlook. In any case, from the perspective of an ethical theorist, the pragmatic argument is extremely weak. If the truth is that most or all of our current ethical beliefs are unjustified, then that is the truth, however unpleasant and hard to accept it is. Error theorists in metaethics are in fact quite happy to accept this, and have argued that evolutionary influences on our moral judgments do warrant such global moral skepticism (Joyce 2006).

The second suggestion that Nichols, Timmons, and Lopez make draws on an analogy with aesthetics. They claim that "Finding out that one's aesthetic tastes (and related judgments) in music are grounded in a-rational and a-reliable mechanisms is not itself a good reason for rejecting those tastes and related judgments" (footnote 2, p. 161). If ethical judgments are relevantly similar, the same goes for them. But there is much reason to doubt this. The reason why ungrounded judgments about music, for example, are relatively immune to rejection is either that there is no fact of the matter or that the facts are relative to individual subjects' tastes (in which case taste-based judgments are automatically truth-tracking and hence justified). I will not rehearse familiar arguments against moral nihilism or relativism here (see e.g. Shafer-Landau 2003). Suffice it to say that there is not much point in normative inquiry of any sort, empirically informed or not, if there are no objective facts of the matter. And why would a commitment have normative authority if any contrary judgment would be just as justified? Normative authority is precisely what ungrounded aesthetic judgments lack—for example, I have no reason to resist acquiring a new taste in ice creams, since liking pistachio would be just as unproblematic as liking chocolate.

So far there is little reason to regard entrenched commitments as prima facie justified or authoritative. Indeed, there is some positive reason to doubt this. Suppose it turns out to be an entrenched commitment, at least for some people, that homosexuality is morally wrong. This is not implausible, and certainly not impossible. Should we then regard belief in the wrongness of homosexuality as prima facie justified, or authoritative for those who hold it? I do not think so. Ethical conservatism threatens to become conservative ethics. Further, the natural emotional reactions underlying entrenched commitments can *conflict*. As Van Schaik et al. point out, sympathy for someone's suffering can conflict with the sense of fairness. Perhaps the person is starving because he did not bother to go on a hunt when everyone else did. If caring and justice are equally entrenched, which side has normative authority in the case of conflict? If it is both, how do we decide between the claims?

So my first problem with ethical conservatism is that entrenched commitments do not, as such, seem to merit normative authority. The second issue is that it is not clear why we should think of entrenched or other emotionally driven commitments as *unreliable* or *non-truth-tracking* in the first place. (Nichols, Timmons, and Lopez use the word 'a-reliable', but there's nothing else it could mean.) As I argued in the earlier sections, there are other ways of privileging certain emotional responses in ethics. It may well be that informed, impartially sympathetic emotions track moral truth, either because moral truths simply are truths about how we would respond, were we to be impartially sympathetic and informed, or because they just happen to tap into mind-independent moral facts. It is, for example, morally wrong to rape a child or knowingly sell a faulty product. Most of us have a non-accidental negative emotional response to raping a child or knowingly selling a faulty product. These responses, then, appear to track at least some moral truths. To establish that they are reliable, we would naturally need to tell much more of a story of how they not only accidentally coincide with moral facts. I will not attempt to do so here. In any case, my bet is that when we have fuller story of which ethical emotions are trustworthy, their being *entrenched* will turn out to play no role in it. Thus, even if empirical research can establish which commitments are entrenched, that discovery will not provide evidence for or against normative views.

### *16.3.4 Via Psychological Unfeasibility to Normative Conclusions*

The final kind of normative argument based on empirical psychology that I want to consider is relatively old. It takes its point of departure from the thought that ethics is for human beings, and thus has to take into account human cognitive and motivational limitations. Moral ideals and demands have to be *psychologically feasible* for the kind of beings we are. This constraint on moral theories is closely related to the old thesis that 'ought implies can'—it cannot be the case that morality requires people to do things they are unable to do, because it would be wrong to blame them for failing to do the impossible. There are deep questions concerning these constraints—What exactly does it mean that someone is psychologically unable to do something? Are there normative demands that do not imply an ought or blame for failure? —but I will assume here that they are along the right lines. This opens up a different kind of potential role for empirical psychology. Since it is an empirical question what human abilities are like, scientific psychology can in principle lead to new normative insights.

The best-known recent argument along these lines is the situationist attack on virtue ethics, in particular its focus on becoming a certain kind of person with certain character traits. Its structure is basically as follows:

1. Virtue ethics tells people to cultivate robust character traits.
2. Most people's behavior varies in response to contextual factors, including very minor ones.
3. Behavioral variance in response to minor contextual factors is inconsistent with the common existence of robust character traits.
4. So, empirical evidence shows robust character traits are, at best, rare/the existence of robust character traits is not empirically supported. (2, 3)
5. An ideal that most people cannot live up to is not psychologically feasible.
6. So, the virtue ethical ideal is not psychologically feasible. (1, 4, 5)

7. A moral theory whose ideal is not psychologically feasible should be rejected. (The Feasibility Constraint)
8. Hence, virtue ethics should be rejected. (6, 7)

In premise 1, robust character traits are "dispositions that lead to trait-relevant behavior across a wide variety of trait-relevant situations" (Doris and Stich 2005: 119) or "relatively long-term stable disposition[s] to act in distinctive ways" (Harman 1999: 317). For example, honesty is a disposition to be truthful and forthcoming in a wide variety of situations in which there might be something to be gained by deception. The perhaps counterintuitive Premise 2 is supported by a large number of social psychological studies that have found, among other things, that people's helping behavior systematically varies due to contextual factors like mood, hurry, and the presence of others, and that a large majority of subjects are willing to hurt others under minor social pressure (for thorough overviews, see Doris 2002; Alfano 2013). Premise 3 draws on the idea that if people had robust character traits, their behavior, especially in such morally relevant cases, would vary from person to person, depending on how virtuous they were. But in fact it seems that it is the situational features rather than people's dispositions that seem to account for manifest behaviors. The remaining steps draw out the conclusions: at most few people seem to have robust character traits. There are, at best, fragmentary character traits like "office-party-temperance" (Doris 2002) that are nothing like virtues. So the virtue ethical ideal is unfeasible and should not be adopted.

In response, virtue ethicists have typically attacked Premises 3 and 5 instead of rejecting the Feasibility Constraint. The first line of defense begins with the rejection of the understanding of character traits that underlies the situationist attack. Character traits are not dispositions to *act*, it says. Rather, they are in the first instance dispositions to perceive, feel, and reason in certain ways, and consequently, perhaps, to act. There is a gap between manifest behavior and character traits (see e.g. Sreenivasan 2002). Perhaps, as Julia Annas (2011) maintains, they are akin to *skills*. This complicates the task of showing the non-existence of traits, since mere behavioral evidence is not sufficient. So, for all the current evidence shows, people may after all have robust character traits. A weakness of this response is that if people's perceptions of reasons do not make a difference to how they act, there is not much reason to focus on them in ethical theorizing. Nor do those who take this line of response typically provide positive empirical evidence for the existence of character traits (although see Russell 2009).

The other main line of response is to grant that virtue is rare, but nevertheless an attainable or at least practically useful ideal (e.g. Appiah 2008, 47)—in my terms, to deny Premise 5, the notion that a psychological ideal few can live up to is psychologically unfeasible in the relevant sense. In their rejoinder, Doris and Stich say that "if virtue is expected to be rare, it is not obvious what role virtue theory could have in a (generally applicable) programme of moral education." (Doris and Stich 2005, 120) This is a weak objection for many reasons. First, it assumes that it is an important standard for assessing normative theories is whether they serve practical didactic aims. This surely need not be the key aspiration of any normative theorist. Second, rarity and difficulty of attaining an ideal do not in any obvious way render it didactically obsolete. Suppose it's very rare to anyone to play guitar as well as Mark Knopfler (as it is). Does that mean it is a bad idea to try to play like Knopfler, when you are practicing to become a better player? Hardly.

Mark Alfano (2013) has a different objection to the virtue-as-an-ideal response. He notes that among the hard core of virtue ethics are claims about the explanatory and predictive power of character traits, as well as what he calls egalitarianism (almost anyone can reliably act in accordance with virtue) and cross-situational consistency in response to reasons. If virtue is hard and rare, Alfano says, "the virtues are loose cogs in our motivational machinery, reliably licensing neither the explanation nor the prediction of behavior" (Alfano 2013: 63). This rejoinder illustrates the common mistake of treating virtue as an all-or-nothing property. It is, however, much more natural to think of virtue as a matter of degree. We can be more or less honest or chaste—that is to say, roughly, we may be more or less sensitive to reasons for truth-telling or abstinence.[11] The truth of positive virtue attribution will depend on the context (a chaste French politician does not cheat on his mistress), much as the truth of other utterances containing scalar adjectives (such as 'tall') does.

The empirical evidence certainly suggests that we may possess such traits to a lower degree than we like to think, so that most of us perhaps cannot, in most contexts, truthfully be described as brave or just, period. But that is to say we are *to some degree* brave or just, so that our behavior may be to some extent be explained and predicted by reference to bravery or justice. Almost everyone can become *more* virtuous, and the more they approach the ideal of the *phronimos*, the more the attribution of virtue traits will explain and predict their behavior. That is how thinking of virtue as gradable reconciles the virtue-as-an-ideal line with explanatory/predictive power and egalitarianism.

Edouard Machery (2010) has recently developed the situationist critique further. As he sees it, the real problem is that virtue ethical ideals presuppose *unified agency*. By this he means that…

> …the psychological causes that are meant to constitute our character and the kind of person we are (our values, desires, norms, emotions, etc.) have a specific causal structure: They (or at least many of them) are unified. That is, they are causally influenced by a common cause or they causally influence one another. (Machery 2010, 225)

[11] Following a lead from Robert Adams (2006), who in turn draws on the old distinction between imperfect and perfect duties, Alfano notes that some 'low-fidelity' virtues, such as generosity, require one to be responsive to some occasions in which giving is called for, while other 'high-fidelity' virtues, such as chastity or justice, require a high degree of consistency—to possess them one has to respond suitably nearly every time. I do not think this is the same dimension I am talking about. The degree of virtuousness is not identical with frequency of acting on a certain kind of reason. You do not have to be very chaste to refrain from sleeping with someone other than your partner 100 % of the time, because the reason to do so is strong. (Insofar as chastity is a virtue, the degree to which it is possessed is manifest in the subtle ways one interacts with attractive non-partners.) Hence, even a low degree of chastity explains and predicts full faithfulness in deed. At the other end, even the most perfectly generous person will not give on every occasion, as the contrary demands of justice, friendship, and other virtues intervene, and the strength of her reasons to give diminishes the less she has to give or more she deprivation she herself suffers.

Machery then argues that human agency is not unified in this sense. He draws on Dual Process Models and research on implicit biases, which suggests that people's conscious values often come apart from their automatic responses. But why is the potential, and indeed frequent disunity between System 1 and System 2 processes a problem for virtue ethics? The reason Machery gives is that "we have no direct control over some psychological causes—namely over the automatic systems—suggesting that it might be difficult to bring them in step with the other states and dispositions that are meant to constitute character." (Machery 2010, 227) But this lack of control, surely, does not come as a surprise to the virtue ethicist. Aristotle, after all, is explicit that acquiring virtue is slow work and significantly subject to moral luck when it comes to having the right sort of temperament, teachers, and environment. What is more, this still looks like a version of the difficulty challenge. So even if Machery is right about the disunity of agency, that does not seem to pose a new problem for the virtue ethicist.

This substantial response leaves Machery's methodological challenge intact, however. He argues that "the proper response to the situationist threat involves examining the empirical literature on agency in detail. There is no easy way for moral philosophers out of a laborious study of human behavior." (Machery 2010, 227) So any defense of virtue ethics must be empirically informed to be credible. To be sure, insofar as we accept Ought Implies Can or Feasibility Constraint, it is hard to deny that empirical facts about human agency potentially undermine character-based ethics. But I still want to reject Machery's methodological thesis. I believe the burden of proof here is on the critic who denies the commonsense view of character that virtue ethics relies on. That is, it is not that the virtue ethicist has to dig through empirical literature to show that courage or kindness is possible (even if rare). Recall the point I made above: the core empirical assumption is not that some or many people are perfectly courageous or kind, but that *people are more or less courageous or kind*, and that most of us can improve in these respects. For all the evidence situationists have presented, we still have no good reason to believe this is false.

## 16.4 Conclusion: Building a Better Armchair

I have charted various ways in which empirical psychological results might be or have been claimed to be important to metaethics and normative ethics in ways that go beyond Armchair Traditionalism. I believe that we have not been given any good reason to believe in bold versions of Ethical Empiricism. Neither metaethical nor normative questions are empirical questions, or questions that could be settled by empirical findings. I have also found various Modest Ethical Empiricist arguments wanting. Generally, the empirical evidence does not do the work it is alleged to do, or provides weak support for one view or another only under strong non-empirical assumptions. Too often, empirical information is noise that distracts from the core issues.

Nevertheless, I cannot claim to have vindicated Armchair Traditionalism either. I have left the door open for the possibility that empirical discoveries may help in conceptual analysis (although only indirectly) and that they may help identify what natural kinds constitute moral thoughts (although the actual identification draws crucially on armchair reflection). I have also allowed that normative ethics may yet benefit from understanding the roots of our intuitions and the feasibility of ethical ideals, even if the existing claims are exaggerated. Perhaps the best overall conclusion to draw is that while armchair reflection will and ought to continue to be central to ethical inquiry, findings about what, why, and how we judge may stimulate and even challenge its results at several important junctures.