Naturalizing Metaethics

Jesse Prinz

Decades ago, it was suggested that epistemology could be naturalized, meaning, roughly, that it could be treated as an empirically-informed psychological inquiry. In more recent years, there has been a concerted effort to naturalize ethics, with a focus on questions in moral psychology, and occasional normative ethics. Less effort has been put into the naturalization of metaethics: the study of what, if anything, makes moral judgments true. The discussion presents a systematic overview of core questions in metaethics, and argues that each of these can be illuminated by psychological research. These include questions about realism, expressivism, error theory, and relativism. Metaethics is beholden to moral psychology, and moral psychology can be studied empirically. The primary goal is to establish empirical tractability, but, in so doing, the paper also takes a provisional stance on core questions, defending a view that is relativist, subjective, and emotionally grounded.

Keywords
Error theory | Expressivism | Metaethics | Moral realism | Naturalism | Relativism | Sentimentalism

Author

Jesse Prinz
jesse@subcortex.com
City University of New York
New York, NY, U.S.A.

Commentator

Yann Wilhelm
ywilhelm@students.uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

Moral philosophy has taken an empirical turn, with experimental results being brought to bear on core questions in moral psychology (e.g., is altruism motivated by empathy?) and normative ethics (e.g., how plausible are the presuppositions of virtue theory?). Some of the recent empirical work also bears on core questions in metaethics. Metaethical questions are varied, but they broadly concern the foundations of moral judgments. What is the basis of such judgments? What, if anything, could render them true? Here I will argue that these questions can be empirically addressed, and longstanding debates between leading metaethical theories may ultimately be settled experimentally. I will describe empirical results that bear on core metaethical questions. I will not present these results in detail here. My goal is programmatic: I seek to establish the empirical tractability of metaethics. Some of the experiments I describe are exploratory pilot studies, presented in an effort to motivate more research. Even with such preliminary results, we will see that some metaethical theories already enjoy greater empirical support than others. I will argue that the best-supported theory at this stage of inquiry is a form of relativist sentimentalism. Defending this position is subsidiary to my primary goal of advertising the value of empirical methods in metaethical theorizing. There has already been an empirical turn in ethics, but metaethics has been less explicitly targeted by these new approaches.

Talking about "an empirical turn" clearly alludes to another turn in the recent history of

philosophy: the linguistic turn. When philosophers turned their attention to language, there was an effort to recast philosophical problems as linguistic in nature. A new set of technical tools was brought into the field: formal semantics. Logic has been part of philosophy historically, but after the linguistic turn it was perceived to be an essential component of philosophical training. Just as formal semantics increased philosophical precision with the linguistic turn, empirical methods have dramatically augmented our tool chest, and stubborn debates may begin to give way. The empirical turn is as momentous as the linguistic turn, and perhaps even more so. Formal semantics allowed us to articulate differences between theories, and empirical methods provide new opportunities for theory confirmation. Neither turn rendered traditional approaches to philosophy idle, but rather supplemented them. Within metaethics, this supplementation may offer the best hope of settling which competing theories are true.

In calling for a naturalist metaethics, it is important to avoid confusion with two other views. "Naturalism" is sometimes construed as a metaphysical thesis, and also sometimes as a semantic thesis. Metaphysically, "naturalism" refers to the view that everything that exists belongs to the natural world, as opposed to the non-natural, or supernatural world. This is sometimes presented as a synonym for physicalism, which can be defined as the view that the world described by the physical sciences is complete, in that any physical duplicate of this world would be a duplicate simpliciter. The causal closure of the physical world and the success of physical science are taken as evidence for this metaphysical view. Semantic naturalism attempts to reductively analyze concepts from one domain in terms of another, which is considered more likely to be natural in a metaphysical sense. In philosophy of mind, this might involve defining psychological concepts in neural or causal terms, while in ethics it might involve defining moral properties in terms of psychological, logical, or social terms (such as hedonic states, principles of reason, or social contracts). Here I will be concerned with methodogical naturalism, which has recent roots in the work of W.V.O. Quine, who grew skeptical about philosophizing through linguistic analysis, and emphasized the empricial revisability of philosophical claims (1969). Quine drew on the methods of John Dewey, and insisted that "knowledge, mind, and meaning [...] are to be studied in the same empricial spirit that animates nautral science" (1969, p. 26). More succinctly, methodological naturalism can be defined as follows:

> Methodological naturalism $=_{\text{Df}}$ the view that we should study a domain using empirical methods.

This is the kind of naturalism that has long been advocated, but too rarely followed, in the domain of epistemology (Kornblith 1985). Neither metaphysical nor semantic naturalism are equivalent to methodological naturalism. Metaphysical naturalism is a view about what exists, not about how to study it. Indeed, some non-naturalists in this metaphysical sense believe that empirical methods can be used to study non-physical or supernatural entities. Semantic naturalism is a view about how to state theories (viz., in reductionist terms), but practitioners have rarely used empirical science in defense of such theories (consider so-called naturalistic semantics). Methodological naturalism has been deployed in discussions of both first-order ethics (e.g., Brandt 1959; Flanagan 1991; Doris 1998; Greene 2007) and in metaethics (e.g., Railton 1993; Prinz 2007b). As Railton points out, a naturalist methodology could result in a reductionist theory of morality, but it need not (see also Boyd 1988). In principle, science could support traditional intuitionism, which is not naturalistic in either of these other senses.

1.1 Methodological preamble

Philosophy has always been methodologically pluralistic. Some use intuitions to arrive at necessary and sufficient conditions for the application of concepts (e.g., Plato's early dialogues). Some try to systematize and revise a large set of beliefs using reflective equilibrium (e.g., Rawls on justice). Some use transcendental ar-

guments to figure out preconditions for thought and action (e.g., Kant). Some use aphorisms or stories to reveal facts about ourselves or to envision possible alternatives (e.g., Nietzsche and the existentialist tradition). Some propose historical analyses of prevailing institutions and values (e.g., Hobbes, Rousseau, and Foucault). Some disclose hidden social forces that buffer prevailing categories (e.g., Marilyn Frye on gender). Some analyze case studies (e.g., Kuhn), probe the structure of experience (e.g., Husserl), or propose formalizations (e.g., Frege). These and other methods suggest that philosophy is a many-splendored thing, and among its many forms one can also find the deployment of empirical results. Examples include Descartes and James on the emotions, Merleau-Ponty on embodiment, and Wittgenstein on aspect perception. Empirical observations have often guided philosophical inquiry. Locke was inspired by corpuscular physics, Marx took solace in Darwinian biology, and Carnap incorporated ideas from behaviorism.

The term "empirical" is used in different ways. In its broadest application, it refers to observational methods. Observation can include an examination of the world, both inner and outer, with and without special instruments. Even introspection can be regarded as a form of observation, as the etymology of the term suggests, and in this sense, the introspection of intuitions is an observational method. Philosophers who use intuitions in theory-construction can be characterised as doing something empirical in this broad sense. Linguistics has used such intuitions to construct syntactic theories, and few would deny that syntax is an empirical field. But the term "empirical" is also used more narrowly to refer to the use of scientific methods, which involve the design of repeatable observation procedures, and the quantification and mathematical analysis of data. The empirical turn in philosophy has been marked by a dramatic increase in the use of scientific results.

Many philosophers have long held a positive attitude toward science, but the frequent use of scientific results (outside of the philosophies of science) is a recent phenomenon. It became popular in naturalized epistemology, which draws on the psychology of decision-making, and philosophy of mind, which has drawn on psychology, computer science, and artificial intelligence. Over the last decade, empirical methods have also become widely used, and widely contested, in ethics.

The resistance to empirical methods in ethics is often chalked up to the fact that ethics is a normative domain, and empirical methods provide descriptive results. This can only be part of the story, however, as there has been little uptake of empirical methods in metaethics. Metaethics is a descriptive domain; it does not tell us how to act morally, but rather explores the semantic commitments and metaphysical foundations of such claims. I suspect the reason for resistance is less interesting and more sociological. Psychology is a young profession, which grew out of philosophy and physiology but then acquired its own institutional standing in the academy, and it has had to fiercely guard its status as a science by distancing itself from the humanities. Meanwhile, philosophy underwent an analytic turn, which led to a preoccupation with conceptual analysis, and an anxiety about psychologism. On this vision, the field began to model itself on logic or mathematics, which were, in turn, taken to be *a priori* domains. I think this is a fundamental mistake. In many domains, the concepts that matter most are grounded in human usage, not in a transcendental realm like (allegedly) mathematics. The arbiters of conceptual truth include both the inferences we are inclined to draw and our linguistic behavior, both of which can be studied empirically. I will not argue directly against a priori approaches, but will instead make an empirical case, or better yet an invitation, by attempting to illustrate how empirical findings make contact with traditional philosophical questions in metaethics.

One manifestation of the empirical turn has been the rise of experimental philosophy. This term most often refers to the work of philosophers who conduct studies that probe people's intuitions about philosophical thought experiments. Strictly speaking, much of this work is not experimental, since the term "experiment" is often reserved in psychological re-

search for studies in which researchers attempt to manipulate the mental states of their participants—experimental conditions are compared against control conditions. Experimental philosophy often explores standing intuitions, rather than the factors that influence those intuitions (e.g., Mikhail 2002). For example, some trolley studies simply poll opinions about the permissability of certain actions. That is a survey rather than an experiment. One can use survey methods to conduct experiments, however. For example, one could conduct a trolley study in which some vignettes use evocative language in an effort to manipulate participants' emotions. Few experimental philosophy studies do anything like this. Most ask for opinions without manipulating psychological states. Thus, experimental philosophy characteristically examines the *content* of people's concepts and beliefs, rather than the underlying psychological processes. In this sense, experimental philosophy is an extension of conceptual analysis. For those interestested in underlying processes, it can, to this extent, be of limited interest. Some experimental philosophy has also been criticized for failing to meet standards of reliable behavioral reseach (Woolfolk 2013). That said, conceptual questions are often important for philosophical theorizing, and methodological problems with experimental philosophy can be addressed by conducting better and better experiments. Often the first efforts (including much of the work I will describe below) are best regarded as analagous to pilot studies, in need of refinement but successful enough to warrant more careful investigation. In addition, many philosophers draw on (and increasingly conduct) studies that qualify as genuine experiments and meet the standards of good social science. There is a long tradition of philosophers using research published in social science journals to defend philosophical positions. For those who find paradigm cases of experimental philosophy too limiting (because they are based on conceptual intuitions or fail to meet certain standards), there are many other empirical results that can provide illumination. The term "empirical philosophy" can be used as a broader label to cover both opinion polls and experimental manipulations. As I use the term, it refers to the use of scientific results, whether obtained by a philosopher or not, to address philosophical questions. The empirical turn should not be dismissed as philosophy-through-opinion-polls; it is a multi-pronged effort advance philosophical debates by drawing upon observational methods of any kind.

The motivations for the empirical turn are varied, but the most general impetus is the belief that some questions cannot be resolved by more traditional philosophical methods. For example, philosophers interested in the physical basis of consciousness cannot rely on introspection or on an analysis of the concept "conscious." And even those interested in analysis of concepts have worried about the limits of introspection. There are basically three different theories of what concepts are: Platonic entities, emergent features of linguistic practice, or psychological states. None of these can be completely investigated by introspection. Even psychological states can be difficult to introspect, because much mental activity is unconscious, and because introspection may be prone to error and bias. Moreover, even if a philosopher could perfectly introspect her own concepts, she would not know thereby that others shared the same concepts, and this would greatly limit the scope of her theories. Some experimental philosophers have argued that philosophers' intuitions are not shared by laypeople. When philosophers and laypeople do agree on intuitions, there is still no guarantee that these accurately reflect reality. For example, most people (at least in the West) find it intuitively plausible that human action derives from character traits, but some empirical philosophers (most notably Owen Flanagan, John Doris, and Gilbert Harman) have drawn attention to psychological research that challenges this assumption.

Traditional and empirical approaches to philosophy are sometimes placed in opposition, but they can also be regarded as interdependent. On one division of labor, traditional methods are used to pose questions and to devise theories that might answer those questions. Empirical methods can then be used to test these theories. This is an over-simplification, of course, because observa-

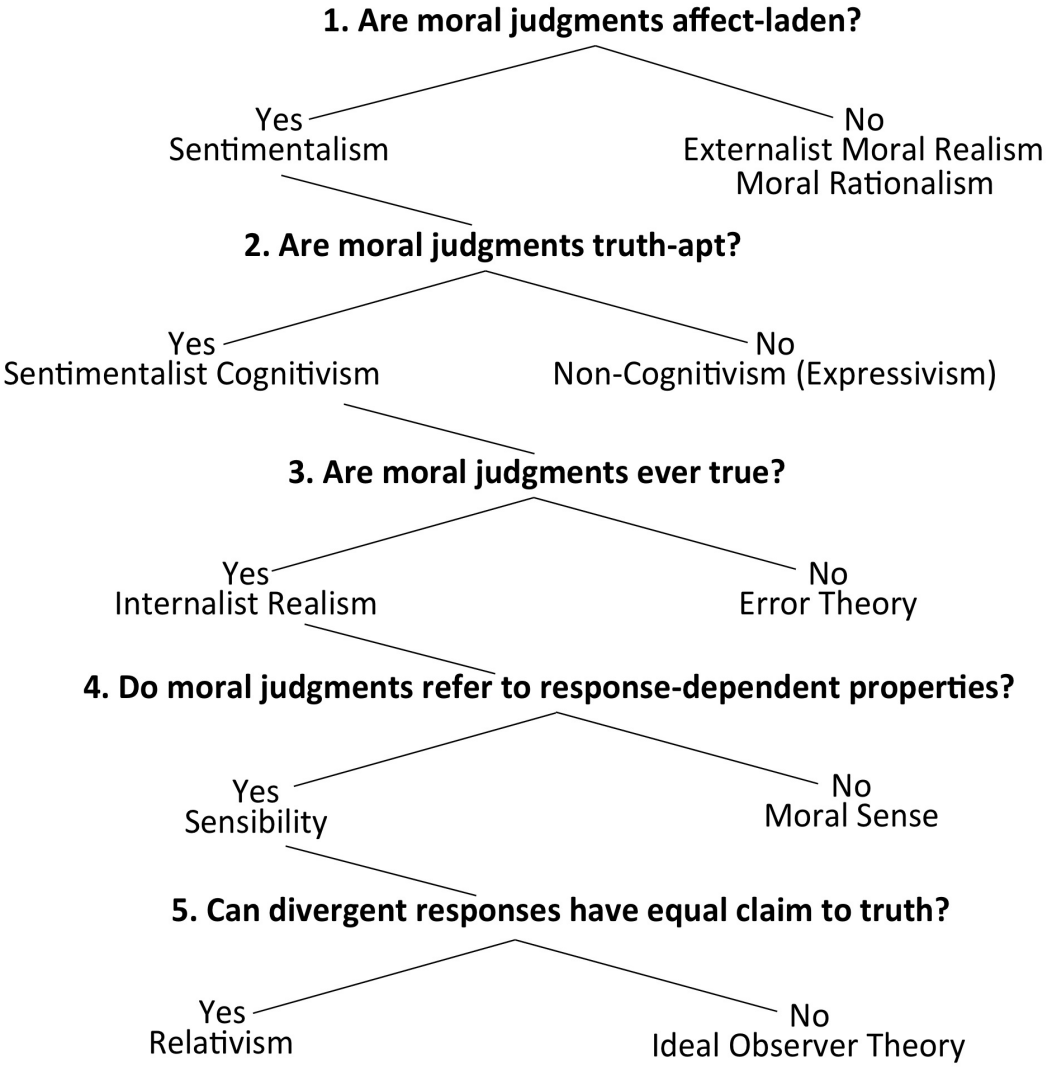

Figure 1: A Metaethics decision-tree

tions can inspire theories, and traditional methods can sometimes refute theories (Gettier cases are a parade example in epistemology), but the proposed division of labor is a decent approximation. Traditional methods have limited testing power because theoretical posits are often difficult to directly observe, and empirical methods have limited power in constructing theories, because theories outstrip evidence. In what follows I will test theories derived from philosophical reflection against the tribunal of empirical evidence.

1.2 A roadmap to metaethics

Let us turn now to the focus of discussion: metaethics. Metaethical questions concern the nature of the moral domain. Metaethicists ask: what kinds of things are we talking about when we make moral judgements? Put differently, metaethics concerns the truthmakers of moral judgements: what kinds of facts, if any, make moral judgements true? That is a metaphysical question but it is normally approached semantically by exploring what we are semantically committed to when we make moral judgements. Metaethics differs from first-order ethics, which concerns the content, derivation, and application of such judgements.

There are many different metaethical theories, and a complete survey here would be impossible. I will focus on major theories that have emerged over the last two hundred and fifty years, with emphasis on proposals that dominated discussion in the twentieth century. To be clear from the outset, my goal is not to consider specific theories that have been ad-

vanced by currently active authors in metaethics. Rather, I will survey broader classes of theories that have been around for some time (decades or centuries) in an effort to establish the relevance of empirical work. An adequate examination of any recent theory would require a narrower focus than I am after here, since each theory makes empirical commitments, if at all, in different places.

To facilitate discussion, I will map out the theories of interest using a decision tree (Figure 1). The tree could easily be arranged differently. Almost any branch could be the starting place, with other nodes occurring higher or lower than they are in this rendition. As we will see in a moment, I begin with a question about "affect" or emotions. This may seem odd to some contemporary metaethicists. Some contemporary metaethicists discuss emotions (such as Alan Gibbard 1990, and Simon Blackburn 1998), but others do not (for example, emotions are discussed less among moral realists). Historically, however, emotions have been a central focus in metaethics. British moralists, who advanced many of the questions that continue to drive the subfield, often begin their analyses with a discussion of moral sentiments. Indeed, the most famous controversy in metaethics before the twentieth century is probably the dispute between British sentimentalism and Kantian rationalism. Even in the twentieth century, some of the most discussed debates concern emotions, such as the debate between emotivists and their opponents. Moreover, the recent empirical turn was triggered, in part, by discoveries linking emotions to moral judgement. So this starting point has considerable historical depth and great relevance to the methodological sea-change that I am interested in here. That said, I don't intend this tree to be anything like a complete map of meta-ethics. One could begin elsewhere and branch out in further directions (I expand the tree leftward, but interesting questions also come up on the right). Though incomplete, the nodes of this tree encompass much of what one might cover in an introduction to metaethics in the Anglophone analytic tradition.

The first question in the metaethics decision tree is, I note, a question about emotions. More precisely, we can ask: are moral judgments affect-laden? The term "affect" is used instead of "emotion" here, because it is broader. I intend the term to cover any conative state, such as a preference, desire, or pro-attitude. For most of this discussion, I will focus on emotions rather than these other affective constructs, because emotions are implicated in the empirical work I will be considering.

The other key term in question 1 is "moral judgments." By "moral judgments" I mean atomic judgments using thin moral concepts, such as *wrong*, *bad*, or *immoral*. The judgment expressed by "Shoplifting is wrong" would be an example. There are many other judgments that arise in moral contexts, including judgments containing thick concepts, such as *cruel* or *unjust*. One can also ask whether these are affect-laden. On one analysis, thick concepts are hybrids that have both a descriptive and an evaluative component, the latter of which may implicate the emotions. For the sake of simplicitly I will ignore that debate here.

Notice that judgements are not sentences but rather the thoughts that sentences express. To propose that such thoughts are affect-laden is to say that each token instance involves an emotion or other conative state. There are different forms of "involvement" that have been discussed in metaethics. On some theories, moral judgments contain conative states as constituent parts. This was the view of Francis Hutcheson, David Hume, and some other British moralists. One might weaken this by saying that moral judgments do not contain emotions, but refer to them. In this vein, John McDowell, David Wiggins, and Alan Gibbard suggest that moral judgments reflect the conviction or norm that it would be warranted to feel certain emotions. Both of these approaches have gone under the heading "sentimentalism," with the prefix "neo-" for the views that say the link between moral judgments and emotions is second-order. Here is a definition:

> Sentimentalism $=_{Df}$ Moral judgments essentially involve affective states, such as emotions, in one of two ways: such states are constituent parts of moral judgments (traditional sentimentalism); or moral judgments are judgments about the appropriateness of such states (neo-sentimentalism).

Those who deny that moral judgments are affect-laden fall into different categories, but two of the most important metaethical theories of this kind are externalist moral realism and (some forms of) rationalism. Moral realists say that there are moral facts, which is to say that some states of affairs are truly right or wrong (cf. Sayre-McCord 1988). Externalist moral realists add a further requirement, namely mind-independence:

> Externalism moral realism $=_{Df}$ There are moral facts and these obtain independently of our recognition of them.

If moral facts are mind-independent, it also follows that we can come to know them without being moved by them. Like scientific facts, we can know that they obtain without feeling any way towards them. Cornell realists, some intuitionists, and many divine command theorists fall into this category. Moral rationalism is a view about how the epistemology or normative status of moral truths:

> Moral rationalism $=_{Df}$ Moral truths can be discovered and justified though a purely rational decision procedure.

Kant (1797) is traditionally read this way, though he also claimed that moral judgments involve moral feelings.

The remainder of my metaethics decision tree concerns those who think that moral judgments are affect-laden. Among those who say that moral judgments essentially involve conative states, there is a divide between those who think that moral judgments are nevertheless truth-apt and those who deny this. This is the second division of the tree. A judgment is truth-apt if it is the kind of thing that can be evaluated as true or false. Some affect-laden judgments may turn out to have a merely expressive function. If I say, "Disco sucks!" I may not be attempting to represent a fact, but merely expressing how I feel. Expressivists follow this analogy:

> Expressivism $=_{Df}$ Moral assertions express mere feelings or non-assertoric attitudes, and do not purport to convey facts.

Charles Stevenson and A. J. Ayer are credited with devising the emotivist theory of morality, which is the simplest theory of this kind. A more sophisticated variant has been developed by Simon Blackburn, who proposes that moral judgments aspire for quasi-truth, but not truth, and thus an ontologically neutral stand-in—which can explain why moral judgments have an assertoric form. Alan Gibbard says that moral judgments do not directly express feelings, as emotivists claim, but rather express the acceptance of norms according to which feelings such as anger and guilt would be appropriate. All these theories have been broadly classified as expressivist.

Those who say that moral judgments are affect-laden and truth-apt need not deny that moral judgments are expressive, but they insist that they more than express feelings; they assert facts. If so, moral judgments can be true or false. Subjectivism falls into this camp:

> Subjectivism $=_{Df}$ the truth of the judgment that something is morally good or bad depends on the feelings or other subjective states of someone who makes that judgment.

For instance, one might propose that "killing is wrong" means "I disapprove of killing." That judgment is true, if the speaker disapproves of killing, and false otherwise. As we will see, there are also more sophisticated forms of subjectivism. Subjectivists are internalist moral realists: they believe in moral facts, but they deny that those facts obtain independently of our attitudes.

The term "cognitivism" has been used to refer to any view on which moral judgments are truth-apt, which is to say they can be assessed for truth. Expressivists are non-cognitivists, and both subjectivists and external moral realists are cognitivists. One could also have a cognitivist theory and nevertheless insist that all moral judgments are false. This would be an error theory.

> Error theory $=_{Df}$ Moral judgments are truth-apt, but they are never true.

The most famous error theory comes from J. L. Mackie. Mackie argues that moral judgments are incoherent. On the one hand, they presuppose that moral facts are objective, which is to say mind-independent. On the other hand, moral judgments presuppose that moral facts are action guiding, and that suggests that they directly motivate us. This suggests that moral judgments must be affect-laden, or otherwise dependent on our subjective states. Since nothing can be both objective and subjective, moral judgments can never be true. Opponents of the error theory deny this and insist that some moral judgments are true. They are, in this sense, moral realists. Moral realists who also claim that moral judgments are affect-laden must take Mackie's challenge head on, showing that truth is compatible with being action-guiding.

Such sentimentalist realists face an immediate question. They can accept Mackie's claim that moral judgments represent objective properties, and find some way to circumvent the incoherence, or they could say that moral judgments refer to properties that are subjective, or response-dependent. The first option might seem untenable, since it accepts that moral judgments are both objective and subjective, an apparent contradiction. But the contradiction can be mitigated by distinguishing between sense and reference. One might say that moral concepts have affect-laden senses—that is, we grasp them by means of feeling—and objective referents. Consider, for example, Kant's aesthetics, according to which beauty consists in a kind of purposeful purposelessness that causes a free-play of the understanding, which results in aesthetic pleasure. A work may have purposeful purposelessness without our recognizing that this is so, but when we recognize it, we feel a certain way. Within ethics, Francis Hutcheson may have held a view that was objectivist and subjectivist in just this way. He suggests that moral facts are established by divine command, but God has furnished us with a moral sense, and that sense works by means of the emotions; when we see objectively bad actions, we feel disapprobation. This has been called a moral sense theory, because it treats our moral passions as a kind of sensory capacity that picks up on real moral facts.

In contrast to this view, one might argue that moral facts are not objective, as Mackie has maintained, but rather are dependent on our responses. This need not imply that moral judgments are mere expressions of feeling; one might say instead that moral judgments refer to response-dependent properties. The idea of response-dependent properties derives from John Locke's notion of secondary qualities. Primary qualities, such as shape, for Locke, exist independently of being perceived, whereas secondary qualities consist in the power that certain things have to cause responses in us. Colors, for Locke, are not out there in the world, but consist in the fact that objects cause certain sensations in us. The moral analogue of this view has been called the sensibility theory, and its adherents include John McDowell, David Wiggins, and David McNaughton. They resist the causal language found in Locke's theory of colors, but say something close:

> Sensibility theory $=_{Df}$ moral properties are those that warrant moral emotions.

Strictly speaking, the sensibility theory is a form of subjectivism, since it says that moral judgments refer to subjective properties (the property of warranting moral emotions), but the notion of warrant allows these theorists to avoid a pitfall or simple subjectivism. For a simple subjectivist it makes no sense to wonder whether something that I disapprove of is really wrong, but for the sensibility theories I can en-

tertain such doubt because I can wonder whether an event really warrants what I happen to feel. The notion of warrant here is not unproblematic, and it often goes unanalyzed. There is one notable exception, however, and that is the ideal observer theory (Firth 1952; Brandt 1959):

> Ideal observer theory $=_{Df}$ The morally good or bad is that which an observer would regard as good or bad under ideal circumstances.

Such circumstances might involve acquiring the status of a moral sage (or consulting a moral sage), as on some virtue theoretic theories, or an ideal version of myself (Smith 1994). Ideal observer theorists are committed to response-dependence; they say that responses determine moral truth, and they further require that those responses come from certain kinds of epistemic agents.

Ideal observer theories offer a negative answer to the final question in the metaethics decision tree. They specify conditions of ideal observation in order to find an authoritative set of responses among a diversity of moral opinions. The hope is that one set of judgments will emerge as epistemically superior to all others; on this view, all moral judges converge under ideal conditions. Here, moral truths work out to be universal. This, of course, is a controversial claim. Suppose we define ideal observers as those who are free from bias, aware of pertinent non-moral facts, and reasoning carefully. It could turn out that, two such observers could still disagree on moral matters. This prognosis leads toward the view that there is no way to arrive at moral consensus. Those who think that moral judgments are rendered true by a judge's response but deny consensus under optimal epistemic conditions end up saying that moral judgments are relative. This view can be stated as follows:

> Metaethical relativism $=_{Df}$ Two judgments expressed using tokens of the same word types, and grasped by tokens of the same mental attitude types can have different truth-values if they are made by different observers.

I will now try to show that each question on the decision tree can be empirically illuminated. Some of the empirical results that I will present come from unpublished, exploratory studies. My goal here is not a detailed documentation of scientific findings, but rather to establish, by means of example, ways in which empirical methods might be brought to bear on the foregoing questions. The hope is that the studies described here might be taken up by others and improved upon.

2 Empirical resolutions to metaethical debates

2.1 Sentimentalism vs. rationalism and externalism

Let's begin with the first question on the metaethics decision tree: are moral judgments affect-laden? No question in ethics has received more empirical attention than this. Dozens of studies have attempted to determine whether emotions play a central role in morality, and the evidence has consistently shown that they do. Let me begin with an unpublished study of my own and then offer a brief review of the empirical literature.

To begin with, let's consider folk intuitions. Do ordinary people use emotions as evidence when attributing moral judgments? To test this, I conducted a simple vignette study, which pitted emotions against verbal testimony. A group of college undergraduates taking an introductory-level philosophy class responded to the following probe:

Fred belongs to a fraternity and his brothers in the fraternity sometimes smoke marijuana. Fred insists that he thinks it's morally acceptable to smoke marijuana. He says, "You guys are not doing anything wrong when you smoke." But Fred also feels disgusted with his frat brothers when he sees them smoking. One day, to prove that he thinks smoking is okay, he smokes marijuana himself. Afterwards, he feels incredibly ashamed about smoking the drug.

Which of the following seems more likely:

1. Fred says he morally approves of marijuana smoking, but in reality he thinks it is morally wrong.
2. Fred feels badly about smoking marijuana, but in reality he thinks it is morally acceptable.

In my small sample, 68.4% chose answer 1, suggesting that the majority of them take emotions as evidence for moral values, even when that directly contradicts self-report. This suggests that many people take emotions to be a sufficient evidence for attribution moral attitudes. An even more dramatic result was obtained when another twenty participants assessed this scenario:

Frank belongs to a fraternity and his brothers in the fraternity sometimes smoke marijuana. Frank insists that their actions are morally unacceptable. He says, "You guys are doing something wrong when you smoke." But Frank does not feel any anger or disgust when he sees his frat brothers smoking. One day, when they are not around, he smokes marijuana himself. Afterwards, he doesn't feel any shame about smoking the drug.

Which of the following seems more likely:

1. Frank says he morally opposes marijuana smoking, but in reality he thinks it is morally acceptable.
2. Frank doesn't feel badly about smoking marijuana, but in reality he thinks it is morally wrong.

Here, 89.5% of participants chose response 1, indicating that they take emotions to be necessary for the attribution of moral attitudes. Absent the right feelings, verbal testimony is treated as an unreliable indicator of a person's values.

This study has at least four serious limitations: people may not trust self-reports; the results were far from unanimous; it fails to distinguish evidence for moral attitudes and essence of moral attitudes; and folk beliefs about moral judgments may be wrong. To get around these limitations we must move beyond experimental philosophy, and look for more direct evidence that emotions actually are sufficient and necessary for moral judgments. But the study is still revealing, because it shows that emotions are used as evidence in moral attribution. Most participants make attributions that fall in line with sentimentalism.

To show that emotions actually do contribute to moral cognition, we can look at three kinds of evidence: cognitive neuroscience, behavioral psychology, and pathology. In each domain, sentimentalism finds support. There have now been dozens of neuroimaging studies on moral judgment tasks, and every one of them, to my knowledge, shows an increase in activation in brain structures associated with emotion, when moral decisions are compared to non-moral decisions. Key stuctures include the posterior cingulate, temporal pole, orbitofrontal cortex, and ventromedial prefrontal cortex. There are only two groups of studies that even appear to depart from this pattern. Joshua Greene et al. (2001) report that emotions play more of a role in deontological judgments than in consequentialist judgments, but their data show that, as compared to non-moral judgments, emotions are involved in both (see their Figure 1). Moreover, Greene et al. use moral dilemmas in which the common denominator is saving lives—they manipulate the nature of the harm necessary in order to save five people in danger. Thus, each moral judgment condition presumably elicits the judgment that it would be good to help people in need. This positive moral judgment may be emotionally grounded, but the neuroimaging method subtracts away this emotional information, because it is present in each scenario, and imaging results of this kind report only contrasts between different conditions. Thus, a major dimension of moral emotions may be systematically concealed by the method. The other study that fails to show an increase in emotional responses during moral judgment is one condition in a series of imagining experiments performed by Jana Borg et al. (2006). But, in that condition, a moral scenario is compared to a scenario about an encroaching fire that threatens one's property, and it is un-

surprising that moral judgments produce less of an emotional response than a case of personal loss.

Brain science resoundingly links moral judgment to emotion, but the method is correlational. Moral rationalists and externalists could concede that moral judgments excite emotional responses, while denying that these are the basis of moral judgment. Imagine the following view: we use reason to arrive at moral judgments, but morality matters to us, so when we arrive at those judgments emotions normally kick in. By analogy, reason might be used to determine that certain life activities (smoking, high fat diets, sleep deprivation) are harmful, and, upon drawing that reason-based conclusion, we tend to experience corresponding emotions, such as anxiety when contemplating lighting a cigarette. Neuroimaging results showing responses to cigarettes might confirm this, showing emotion areas active when cigarettes are seen, but that wouldn't refute a rationalist theory of how we arrive at the judgment that cigarettes are dangerous.

To adjudicate between the thesis that emotions are constitutive of moral judgments, as opposed to mere consequences, we need behavioral evidence. Numerous studies now establish a causal link between emotion and moral judgment. When emotions are induced, they influence how good or bad things seem. Induction methods have been widely varied: hypnosis, dirt, film clips, autobiographical recall, and smells. In one recent study, Kendal Eskine, Natalie Kacinik, and I induced bitterness by giving people a bitter beverage and found that moral judgments became more severe (Eskine et al. 2011). In other recent studies Angelika Seidel and I use sound clips to induce specific emotions, and we have shown that different emotions have different moral effects: anger induces more stringent wrongness judgments about crimes against persons; disgust induces greater stringency on crimes against nature (such as cannibalism); and happiness induces stronger judgments that it is both good and compulsory to help the needy (Seidel & Prinz 2013a, 2013b). There is also evidence that we feel different emotions when judging our own actions than when judging others. When another person commits a crime against nature, we tend to feel disgust, but when we perform an act deemed by others to be unnatural, the most common response seems to be shame. Conversely, when others commit crimes against persons, we feel angry, but guilt is the natural response when we perform such acts ourselves. To test this hypothesis, I conducted a forced-choice study in which a group of college undergraduates had to pick guilt or shame in response to mildly "unnatural" acts ("Suppose your roommate catches you masturbating"), and mildly harmful acts ("Suppose you take something from someone and never return it"). 80% chose shame for the first case, and over 90% picked guilt for the second.

Such findings demonstrate that different emotions play different roles. I mentioned three distinctions that are currently receiving empirical attention: the split between positive and negative emotions (praise and blame), between two kinds of blame (crimes against nature and crimes against persons), and between self- and other-directed blame. The self/other distinction may be particularly important because it helps us see how moral emotions differ from their non-moral analogues. Anger (or at least irritation) and disgust can both occur in non-moral contexts, but they take on a moral cast, I submit, when paired with dispositions to feel guilt and shame, respectively. If I find eating insects physically revolting, I will experience disgust when I see others eat insects, and disgust when I inadvertently eat them myself. But if I found insect eating immoral, it would not be disgust that I experience in the first-person case, but shame. This feeling of shame would motivate me to make amends for my actions, or to conceal my wrongdoing from others, not simply to repel the unwanted food from my body. The self-directed emotions round out the punitive cast of our moral attitudes. We see morally bad acts as not just worth aggressing against, but as worthy of apology. This need not be a second-order belief. Rather it is implicit in the fact that moralized behaviors carry emotional dispositions toward self

and other that together promote a punitive attitude: a disposition to issue and submit to punishment.

Putting this together, I propose that standing moral values (the values that a given individual has for an extended period of time) consist in dispositions to feel the self- and other-directed emotions that I have been discussing. Such an emotional disposition can be called a sentiment. On any given occasion when a standing value becomes active in thought—i.e., when a moral judgment is made—these dispositions result, all else being equal, in an emotional state. The emotion that is felt depends on who is doing what to whom. For example, if I recall a situation in which I hurt someone's feelings, I will have a feeling of guilt regarding that event, because a person was harmed and I was the culprit. This feeling of guilt toward an event constitutes my judgment that the action was wrong, and I gain introspective access to this judgment by feeling guilt well up inside me. If this is right, then emotions are not merely effects of moral judgments, but essential components of them.

Against this picture, one might object that emotions are merely a heuristic that can be used in certain circumstances, but not strictly necessary for making moral judgment. Following the analogy mentioned before, anxiety might be used as a heuristic when deciding whether to smoke, but the judgment that smoking is dangerous does not depend on fear, and was initially arrived at by the light of reason.

To establish that emotions are not merely helpful heuristics, one must see what happens when emotions are reduced or eliminated. To look into this, Eskine (2011) gave people the bitter taste manipulation and then warned them not to let the feelings caused by that beverage interfere with the moral judgments. In this condition, he found that moral judgments were considerably less severe than a control condition, suggesting that, when we ignore emotions, it is harder to see things as wrong. The finding indicates, in other words, that moral judgments subside when emotions are absent. The study cannot confirm this strong claim, however, because people cannot suppress emotions completely. More powerful evidence comes from the clinical populations who suffer from emotional deficits. For example, psychopaths, who suffer from deficit in guilt and other negative emotions, notoriously fail to appreciate what is wrong with their actions (Hare 1993). Similarly, people with Huntington's disease, which impairs disgust, show high incidence of paraphelias, suggesting that they cease to see deviant sexual behavior as wrong (Schmidt & Bonelli 2008). Kramer (1993, p. 278) argues that anti-depressants can flatten affect in a way that results in a "loss of moral sensibility." There is also a positive relationship between alexithymia and Machiavellianism, suggesting that a reduction in emotional competence may act in ways that are more instrumental than moral (Wastell & Booth 2003). For better or worse, there is no clinical condition in which all emotions are absent and behavioral function remains, but these findings suggest that selective or global dampening of the emotions leads to corresponding deficits in moral judgment. That is, people with diminished emotions seem to be insensitive to corresponding parts of the moral domain, suggesting that they may not be forming moral judgments.

The evidence summarized here suggests that emotions arise when we make moral judgments, that emotions are consulted when reporting such judgments, and that moral judgments are impaired when emotions are unavailable. Some of this evidence is preliminary, but, for present purposes, let's assume that the findings hold up to further and more stringent testing. By inference to the best explanation, such findings suggest that emotions are components of moral judgments. The idea is that, when people say something is morally bad, the thought they are expressing on that occasion consists of a negative emotion directed towards the thing judged bad. Emotions, on this view, function like predicates in thought. That is what traditionally sentimentalists, such as Hume, seem to have maintained. Hume thought ideas—the components of thoughts—were stored copies of impressions, and the idea of moral badness consisted in a stored copy of the impression of disapprobation.

Traditional sentimentalism, which says that emotions (or sentiments) are actually components of moral judgments, differs conspicuously from neo-sentimentalism. Neo-sentimentalists theories say that moral judgments are judgments about the appropriateness of emotions. These theories do not straightforwardly predict that emotions come on line when we make moral judgments, nor that a reduction in emotions should interfere with our ability to moralize. Instead, they predict that people will think about emotions when they make moral judgments. Correlatively, they also predict that people with limited metacognitive abilities will lose their ability to make moral judgments; this is not the case (Nichols 2008). Thus, given the current state of evidence, traditional sentimentalism outperforms neo-sentimentalism empirically. Traditional sentimentalism predicts a robust pattern of empirical findings.

Rationalists and externalist moral realists might baulk at this point and say that the empirical evidence lacks the adequate modal strength to support sentimentalism. The evidence shows that emotions are often consulted when making moral judgments, but this leaves open the possibility that we might also make moral judgments dispassionately under circumstances that have not yet been empirically explored. So stated, this objection is just an expression of faith. It suffers from both conceptual and empirical weaknesses. Conceptually, opponents of sentimentalism must say what moral judgments are, such that they can be had dispassionately. What thought is a dispassionate person conveying, when she says, "Killing the innocent is morally bad?" Any attempt to give a reductive answer will be vulnerable to open-question worries. No descriptive substitute for the phrase "morally bad" leaves us with a sentence that is conceptually synonymous with the third.

Arguably, the open-question argument does not threaten sentimentalism. Let's distinguish two kinds of open questions. First, given a certain attitude towards killing, one can still wonder whether killing really is morally bad. Second, given a certain attitude toward killing, one can wonder whether one is thereby regarding it as morally bad. Reductive theories of value leave both questions open. If I form the attitude that killing cannot be willed as a universal law, I can still wonder both whether killing is bad and whether I am judging that it is bad. Sentimentalism leaves the first question open, but not the second. When experiencing outrage at killing, it seems impossible to wonder I am regarding killing as bad. I can of course wonder whether killing really is as bad as it seems. Such doubts can arise because I may not know the true source of the emotion I am feeling. Perhaps my outrage comes from some extraneous source (such as a bitter beverage), for example. But this open question does not threaten the thesis that moral judgments are constituted by sentiments. The only open question that poses such a threat would be one about what my attitude is, not one about whether my attitude is true. The fact that some sentiments are experienced as condemnatory effectively closes the question about whether someone experiencing those sentiments is adopting a moral stance. By analogy, imagine tasting a wine and wondering whether it really is delicious, while experiencing gustatory pleasure. We can have this thought (a thought about truth), because we can't be sure where the pleasure came from (was it the wine or the company?). But we can't experience gustatory pleasure and wonder whether we are, at that moment, finding the experience delicious. Thus, gustatory pleasure is plausible a component of deliciousness judgments.

The foregoing may look like a conceptual argument for sentimentalism. But it can also be construed as an empirical claim. The argument hangs on the premise that people experiencing outrage take themselves to be making moral judgments. This can be empirically tested. Indeed, all the evidence about people consulting their emotions when making moral judgments stands as evidential support. Merely making someone mad results in more negative moral attitudes. This can be interpreted as showing that, when people are angry, there is no question for them about whether they are holding something in negative moral regard. Conversely, it would be easy to show that people do not ne-

cessarily draw this inference when they form the judgment that something cannot be willed as a universal law. Opponents of sentimentalism owe us a positive account of evaluative thoughts that avoids open-question worries as successfully as sentimentalist accounts.

Opponents of sentimentalism might try to bypass this demand by offering a non-reductive account of moral judgments, treating thin moral concepts as primitives. That possibility, which was attractive to Moore, looks unmotivated given the empirical evidence for an emotional foundation. Every study suggests that emotions arise when we make moral judgments. All evidence also suggests that when emotions are eliminated, judgments subside as well. This does not prove that we can make moral judgments without emotions, but, by induction, it provides evidence. Some have argued that extant evidence is ambiguous about whether emotions are essental components of moral judgments or mere accompaniments, but I have suggested here that the former may provide a better explanation (and certainly better predictions) of the total pattern of data (Huebner et al. 2009; Waldmann et al. 2012). Until opponents of sentimentalism can identify some clear cases of moral judgments without emotions, they will be on the losing side of the debate. At the moment, there is no empirical evidence that this ever happens.

Notice too, that it would be relatively uninteresting to show that, under as-yet-unidentified and highly unusual conditions, people can make what look like moral judgments in the absence of emotions. The sentimentalist will reply that the vast majority of ordinary moral judgments are emotionally based. If moral vocabulary is occasionally used dispassionately, sentimentalists can ask whether the thoughts expressed on such occasions are of the same kind that we find, in study after study, in the usual cases. Upon finding a class of dispassionate judgments, one might do best to posit an ambiguity in the category. The sentimentalist can content herself with the project of providing a metaethics for garden-variety moral judgments, while leaving open the possibility that there may be psychological exotica, which conform to the theories of their opponents. At the moment, there is no empirical evidence for such exotica.

More modestly, the empirically-minded sentimentalist might welcome an attempt to find evidence for opposing views. Little effort has been put into this task, though empirical claims for emotion-free moralizing are occasionally advanced. The most publicized example is Koenigs et al.'s (2007) study, which shows intact consequentialist judgments in patients who suffer from ventromedial prefrontal brain injuries, which are thought to impair emotion. But this description is misleading. As the authors note, ventromedial patients are highly emotional, and their most notorious symptom is that they are insensitive to costs when seeking rewards. Presumably, reward- seeking is an affectively grounded behavior. The fact that these patients make normal consequentialist judgments does not entail that they rely on reason alone, but rather on their positive emotions. Since these emotions cannot be easily regulated by negative feedback in ventromedial patients, they tend to be more consequentialist than healthy populations—that is, they are more willing to push a heavy man in a trolley's path in order to save five.

Will better empirical evidence for rationalism or externalist moral realism be forthcoming? I doubt it. Rationalists hold that we can arrive at moral judgments through reasoning. Unlike some sentimentalists, I think reasoning is important to morality. It is likely that we use reasoning to extrapolate from basic values to novel cases. But it is unlikely that we could use reasoning to derive basic moral values. Philosophers have tried to do this for centuries with no consensus behind any view. This might be described as a strong empirical argument by induction: thousands of smart, trained moral experts have failed to identify a line of reasoning that is widely recognized as providing adequate rational support for basic moral propositions. Moreover, when moral debates arise, there is little evidence that reasoning is efficacious on its own. Instead, societal transformations in values seem to arrive with political upheavals, economic revolutions, and generational change. Attitudes towards slavery changed with the indus-

trial revolution, women's suffrage came with a world war, and increase in support for gay rights correlates with the dissolution of traditional social roles and economic transformations that have made procreation more costly than abstinence. I don't mean to imply that there are no rational arguments for these liberation movements. Rather, I am suggesting that those arguments take hold only when social conditions are right. It is noteworthy, for example, that scientific racism appeared very late in the history of slavery, suggesting that slavery was not simply based on false beliefs about racial inequality. In fact many societies have enslaved their own people, and many proponents of scientific racism have been against slavery. Rather, advocacy of slavery seems to reflect a set of basic moral values that changed in recent history: values that say social standing can be determined by the lottery of birth. With industrialization, models of labor based on the idea of self-determination took hold, and the idea that birth should determine social standing began to wane. Of course, it hasn't disappeared altogether, but it has been tempered by the the emergence of a new norm. Before industrialization, the idea that human beings are born equal and free might have seemed manifestly false, and thus it could have played no effective role in any argument against slavery. With industrialization, this premise gained appeal, and became the foundation of compelling arguments. Arguments are not inert, but they are only as good as the premises on which they are based, and the plausibility of those premises may depend on factors other than reasonging. It is possible that reasons have little role in driving basic values. And if so, then the recent broadening moral umbrella is not the result of a rational inference to the conclusion that our basic values cover more cases than we thought, but rather an irrational shift in basic values.

A realist might concede that such considerations threaten rationalism, but vie instead for a kind of intuitionist perspective, according to which basic moral truths are simply obvious. To me, this looks like a magical moral epistemology —one wonders what moral facts could be such that our moral sense could simply lock on to them. It is also open to a damaging empirical objection. Phenomenologically, it is true that moral intuitions often seem immediate and unbidden, but this can be readily explained on a sentimentalist account. Emotions are conditioned (by training or evolution) to arise automatically and often intensely when certain actions, such as torturing babies, are considered. This gives an impression of immediacy without postulating any special contact with moral reality. Moreover, these intuitions vary from group to group. For example, there is empirical evidence that liberals and conservatives have divergent basic values (Graham et al. 2009). The presence of such foundational intuitions can be explained demographically, and their lack of convergence casts doubt on the existence of a moral faculty that reveals universal moral truths. In other words, intuitionism is vulnerable to a debunking argument. Social science coupled with sentimentalism provides a good explanation of deeply-held intuitions, so there is no need to suppose that these intuitions reflect anything deeper.

This point about moral variation, to which I will return, also counts against some forms of externalist moral realism. Advocates of that position sometimes suggest that objective moral facts can be established by identifying the external factors that best explain human moral behavior or judgments. If moral behavior and judgments vary from group to group, however, it is unlikely that we will find an external common denominator underlying these practices. Such a search also seems unnecessary given that we already have good explanations of moral behavior and judgments in terms of socially-conditioned sentiments.

None of these arguments are the nail in the coffin for externalist realist or rationalist theories. They merely illustrate the relevance of empirical results. The findings mentioned here must be explained. It is my contention that sentimentalism provides the best explanation of the findings I have reviewed, but further arguments and evidence could tip the balance in another direction.

2.2 Cognitivism vs. non-cognitivism

Let's move on to the second question on the metaethics decision tree: Are moral judgments truth-apt? As positioned on the tree, this is a

question that arises for sentimentalists, raised pressingly by the conclusion that moral judgments have a basis in the emotions. It is that conclusion that seems to put truth-aptness in jeopardy, since emotions have not traditionally been regarded as having accuracy conditions in the way regarded as allowing for truth. But, it should be noted that the question of truth-aptness could also be raised independently of sentimentalism. There are non-sentimentalist theories that deny truth-aptness (for example, one might say that moral judgments are imperative, while denying that they need be passionate), and there are non-sentimentalist theories that accept truth-aptness (the vast majority fall in this category). To keep things as neutral as possible, I will begin by asking whether there is any empirical evidence that moral judgments are non-cognitive, whether or not they are affect-laden.

The posing of this question is itself a degree of philosophical progress, because non-cognitivists too rarely reflect on the predictions of their view. Indeed, the most obvious empirical prediction fails resoundingly. If moral judgments do not aim at truth, we might expect them to have a non-declarative syntactic. For example, we might expect them to take the form of imperatives or exclamations. But they do not. In every language that I know of, moral judgments are expressed using declarative sentences, which should stand as a profound embarrassment to the theory. Granted, non-cognitivists sometimes propose elaborate logics to accommodate this fact, but it is surprising that they should have to do so. One would expect the surface grammar to reflect the non-cognitive form.

To push things further, one might look for more subtle linguistic evidence in favor of non-cognitivism. For example, some non-cognitivists assume that moral utterances have the illocutionary force of directives, such as orders, requests, or demands. Directives often occur in speech contexts that contain words that play a role in persuasion, such as "come!", "let's", or "we encourage you…" To empricially test this kind of non-cognitivism, Olasov (2011) ingeniously used this technique for sociolinguistics, called corpus analysis. He used a set of such linguistic elements that correlate with directive speech, such as those just mentioned, and he searched corpora of spoken and written texts for co-variance between these elements and moral terms. He calls the directive elements "suasion markers," and the correlations between these and other linguistic items a "suasion score." Non-cognitivism seems to predict a high suasion score, given the postulated directive function of moral judgments. This prediction fails. Not only is there no positive correlation between moral vocabulary and suasion markers, there is actually a negative correlation, which approaches significance. This negative relationship was observed in seventeen out of nineteen different categories of text that he examined. These results are preliminary—a first foray into empirical ethics—but they provide compelling evidence that moral discourse is not directive in nature.

Non-cognitivism entails that moral discourse does not aim to refer to facts in the world. This carries another linguistic prediction that can be readily tested. Certainly adverbs are used to indicate a focus on how things are in the world. These include "really," "truly," and "actually." These words have other uses ("really" can be a term of emphasis), but they often play a role in emphasizing the factive nature of the modified phrase. Therefore, if non-cognitivism were true one might expect these words to rarely be used as modifiers for moral terms. To test this, I used Google search engine to search for and note the frequency of three phrases: "really immoral," "truly immoral," and "actually immoral." To do this, I needed a baseline, and chose to compare "immoral" to a word widely believed to designate a objective feature of the world. I chose "triangular," a classic primary quality, on a Lockean scheme. The results are as follows (as of March, 2013):

"really triangular": 6,500 hits
"really immoral": 10,600 hits
"truly triangular": 4,920 hits
"truly immoral": 32,000 hits
"actually triangular": 21,600 hits
"actually immoral": 61,600 hits

Clearly, the adverbs that indicate a real-world focus are used more frequently for moral terms

than for terms designating objective physical features—over six times as common in the case of "truly." I also tried the phrases "in truth," "truthfully," and "in actual fact":

"truthfully triangular": 6 hits
"truthfully immoral": 44 hits
"in truth triangular": 46 hits
"in truth immoral": 1,350 hits
"in actual fact triangular": 2 hits
"in actual fact immoral": 133 hits

These truth-tracking phrases modify "immoral" between 7 and 166 times more frequently than they modify "triangular." Moreover, these differentials are misleadingly small because the baserate for "immoral" is far lower than "triangular" (6,910,000 hits as compared to 11,600,000). This was just an exploratory study, but there is a simple implication. Non-cognitivism makes linguistic predictions, and when those are tested, they do not seem to pan out. Non-cognitivists owe us evidence, or they must deny that their theory makes predictions, in which case it would cease to be falsifiable.

In response, non-cognitivists might claim that there is one crucial line of evidence in favor of the view, and it's a line of evidence that we have already seen. In the previous section, I surveyed studies suggesting that morality is affect-laden. At the start of this section, I said the non-cognitivism is orthogonal to affect-ladeness, but some non-cognitivists would vehemently disagree with this. They would say that non-cognitivism *follows from* affect-ladeness. Emotions are traditionally regarded as feelings, and feelings are not traditionally believed to be representations of anything. If the thought that killing innocents is wrong is really a bad feeling about killing, then why think this thought has any truth conditions? Does a feeling of indigestion or irritation really refer?

This move might have been compelling in the early part of the twentieth century, but the last fifty years of emotion research have emphasized the intentionality of affect. Some philosophers have adopted cognitive theories of the emotions, according to which emotions are identical to judgments. Elsewhere I have argued against such theories, in favor of the view that emotions are bodily feelings (Prinz 2004), but contemporary feeling theorists still insist that emotions aim to refer. Feeling sad, for example, can be understood as a downtrodden bodily state that represents loss. To say that the feeling represents loss is to say that it has the function of arising in response to losses, and hence carries the information that there has been a loss to a person who experiences it. In a like manner, pain may indicate tissue damage and fatigue may indicate energy depletion, even though pain and fatigue are bodily feelings. None of these feelings are arbitrary. They prepare an organism to cope with specific conditions or events. Emotions qua feelings are in the business of keeping us abreast about how we are faring. Each emotion has a different significance, and any one of them can misfire. I might be sad when there is no loss, or frightened when there is no threat. Such emotions would qualify as erroneous.

If emotions are in the business of representing, then there is no difficulty supposing that moral judgments are truth-apt. When we sincerely assert that, "Killing innocents is bad," we express a negative feeling towards killing, and that feeling functions as a kind of visceral predicate. It attributes a property to killing (I will have more to say about this property below). In this sense, moral discourse may be much like other forms of emotional discourse. If we say that some food is icky, we express a feeling, while also attributing a property. For example, the feeling of ickiness might represent the property of noxiousness, or perhaps something more subjective, such as the property of causing nausea in the speaker. Someone who calls something "icky" need not know what property that feeling represents, but most language users probably recognize that in using this term we are attempting to say something about whatever it is that elicits the feeling. By analogy to "icky," moral assertions can be understood as both expressive and predicative. It is a mistake, based on overly simplistic theories of emotions, to assume that feelings cannot play a semantic function. Once we see that feelings can represent properties and function as predicates,

non-cognitivism no longer looks like a serious option.

2.3 Realism vs. the error theory

It is one thing to say that moral assertions aim to represent and quite another to say that they succeed in doing so. It is possible that when we say that an action is immoral, we aim to ascribe a property to it, but we do not succeed in doing so. This is precisely what defenders of the error theory have claimed. So, even if the forgoing case for cognitivism succeeds, we must now descend the decision tree and ask whether moral judgments are ever true.

The error theory, which states that moral judgments are truth-apt but always false, was first promulgated by J. L. Mackie (1977). Mackie's argument begins with the premise that moral predicates aim to represent properties with two important features. The first is objectivity: moral properties are supposed to be the kinds of things that can obtain independent of our beliefs, desires, inclinations, and preferences. The second is action-guidingness: moral properties are supposed to be the kinds of things that compel us to act when we recognize them. Mackie's second premise is that these two features are difficult to reconcile. Objective properties are usually the kinds of things about which we can be indifferent. Mackie uses the term "queer" to describe properties that are both objective and action-guiding, and he also suggests that such queer properties would require an odd epistemology. For these reasons, he thinks we shouldn't postulate objective action-guiding properties. But, Mackie thinks that moral concepts commit to the existence of such properties, and, thus, that moral judgments posit properties that don't exist. Therefore, moral judgments are systematically false.

In recent years, the error theory has become popular among evolutionary ethicists (Ruse 1991; Joyce 2006). Mackie's theory leaves us with a puzzle. Why do people make moral judgments if they are incoherent? Evolutionary ethicists purport to have an answer. They say that morality is an illusion that has been naturally selected because it confers a survival advantage. For example, if we believe that cheating others is objectively bad and that belief is action-guiding, then we will hold others accountable when they cheat, and we will resist cheating even when it might seem advantageous to do so. This reduces the likelihood of free riders and leads to an evolutionarily stable strategy—one that can foster cooperation and collective works. Evolutionary ethicists also typically endorse sentimentalism, suggesting that moral emotions have evolved to motivate such things as punishment and altruism. Mackie himself is not explicit about the role of emotions in his view, which makes it unclear what he means when he says that we perceive the discovery of alleged moral facts to be action-guiding. The link between judgments and emotions, emphasized by evolutionists, provides one answer.

The evolutionary addendum to Mackie's argument may look like an empirical reason for siding with the error theory. Natural selection is a well-confirmed process, emotions have some basis in evolution, and evolutionary models confirm that emotionally-grounded moral instincts would be adaptive. But there are empirical reasons for doubting the evolutionary story, and for doubting the key premises in Mackie's argument. Consequently, I think the case for the error theory fails.

The evidence for an evolved moral sense is underwhelming. A thorough critique cannot be undertaken here, but let me offer two broad reasons for doubt (for more discussion, see Prinz 2007a). First, there is little evidence for a moral sense in closely related species. Recall that moral judgments are underwritten by emotions such as anger, disgust, guilt, and shame. There is no evidence that the last three of these emotions exist in chimpanzees, and the anger they exhibit might better be described as reactive aggression, because there is little reason to believe chimps form robust tendencies to be angry about third party offences when they are not directly involved. Evolutionists point out that chimps engage in reciprocal altruism, and other forms or prosocial behavior, but these behaviors may not depend on any moral judgments. Indeed, psychopaths engage in reciprocal altruism (Widom 1976), and chimps often be-

have in ways that seem psychopathic; they can be extremely violent (Wrangham 2004) and indifferent to each others welfare (Silk et al. 2005).

Evolutionary ethicists might concede this and argue that morality evolved in the human species after we split from other primates. But this position is vulnerable to a second objection: there is good reason to think that morality in humans is learned. Moral judgments derive from emotions that originate outside the moral domain, such as disgust, which is first applied to noxious agents and later expended to the social domain, through conditioning (Prinz 2007a). Even guilt and shame may be learned byproducts of non-moral emotions: shame is related to embarrassment and guilt may be a blend of sadness and anxiety brought on by violating a social norm (Prinz 2005). These emotions and their range of application depend on extensive conditioning in childhood. Moral variation across cultures is considerable, as we will see, and shared moral values can be attributed to widespread constraints on building a stable society (for example, stable societies must prohibit wanton murder within the in-group). Moreover, there is no poverty-of-the-stimulus argument for morality; children receive ample "negative data" in the form of punishment, and they directly imitate values in their communities. As I argue in greater detail elsewhere, arguments for innate moral norms have been unconvincing (Prinz 2007a). This suggests that morality is learned, not evolved.

If morality is acquired through learning, then one cannot bolster Mackie's argument by assuming that morality is the product of evolution. This alone does not undermine the error theory, however. Error theorists might abandon the evolutionary approach and try to explain systematic error by appeal to a learning story. There is some evidence that people tend to treat certain rules as universally binding, regardless of operative conventions. When asked whether it would be okay to hit a classmate if the teacher granted permission, children tend to say "no." Turiel (1983, Ch. 7) who made this discovery, denies that such objectivist leanings are innate. Rather, he thinks children learn to distinguish moral and conventional rules. Some subsequent authors have argued that the learning in question involves emotional conditionism (Blair 1995; Nichols 2004). Moral rules are aquired through the inculcation of emotions such as anger, guilt, and shame. There are strong negative feelings associated with hitting that don't disappear when children imagine the teacher saying it is okay to hit. Violating social conventions may lead to other emotions, such as embarrassment, but these are mitigated when we move from one social setting to another. For example, wearing a hat at the dinner table might be frowned on in some circumstances, but not when wearing a birthday hat at a birthday party. The idea that moral rules are learned by emotional conditioning could also explain their motivational impact; emotions impel us to act, so emotionally grounded rules seem to carry practical demands. This analysis would explain both features emphasized by Mackie—action-guidingness and objectivity—without assuming that moral rules actually are objective. Thus, the error theory could get off the ground without assuming that morality is a product of evolution.

On closer scrutiny, however, this argument is not strong enough to rescue the error theory. It conflates objectivity with authority independence. It is true that children think hitting is wrong even when it is permitted, but that does not mean they think moral truths exist independently of subjective responses. Many of our subjective responses seem independent of what authorities happen to say—our preferences for food and music, for example. But we don't necessarily infer that these things are objective. So it is a further empirical question whether objectivity is an essential feature of how we understand moral properties.

This brings us to the heart of Mackie's argument. Should we grant his first premise that moral assertions entail objectivity? Empirically, the answer is a bit messy. When polled, many people assume that morality is objective, but many reject this assumption (Nichols 2004; Goodwin & Darley 2008). In survey studies, there is a nearly even split between objectivists and their opponents. Strikingly, belief in ob-

jectivity correlates with religiosity. Goodwin and Darley report that religious beliefs were the strongest predictor of objectivity that they were able to find. This suggests that beleif in objectivity is not an essential part of moral competence, but is, rather, an explicitly learned add-on that most often comes with religious education. The authors also found that belief in objectivity goes down in cases of moral issues about which there is considerable public debate, such as abortion. This might be interpreted as showing, again, that objectivity is not a conceptual truth about the moral domain, but rather a negotiable add on, which can be abandoned in light of counter-evidence. Faith in objectivity goes up with certain relgious beliefs (e.g., divine command theory), and goes down when confronted with the fact that decent, intelligent people have very different moral convictions. In Quine's terms, moral objectivism, when it is found, may be collateral information rather than an analytic truth—a belief about morality that we are willing to revise.

To test this hypothesis, I conducted a survey study in which I compared a moral predicate (*immoral*) to two natural kind terms (*beetle* and *tuberculosis*), which paradigmatically aim to designate objective properties, and to two terms that are often said to represent secondary qualities (*red* and *humorous*). If natural kind terms have a presumption of objectivity, then any threat to that presumption should lead people to conclude that those terms don't refer. Things are a little trickier with terms such as *red* and *humorous*: many people believe that they designate objective properties, but are willing to give up this assumption when presented with countervailing evidence. When told that there is no unifying essence to humor, people do not conclude that nothing is funny; they conclude that humorousness is a property that depends on our responses. In other words, objectivity is not analytically entailed by *humorous* or *red*. It is collateral information. My study was designed to see if *immoral* followed this same pattern.

A group of college undergraduates read the following vignette for the *immoral* case, with comparable vignettes for the other terms:

> Suppose scientists discover that there are two kinds of things that people call immoral. Would it be better to say: (a) The term "immoral" is misleading, and it might be better to replace it with two terms corresponding to the two kinds of cases.
> Or
> (b) The fact that there are different cases is interesting, but doesn't affect the word. The fact that we react the same way to these two things is sufficient for saying they are both members of the same category; they are both immoral.

When given these options, 75% chose option (b) for *immoral*, resisting the first option which is tantamount to an error theory. Exactly as many chose option (b) for *red*, and a few more picked (b) for *humorous* (90%). In contrast, (a) was the dominant answer for the natural kind terms, *tuberculosis* and *beetles* (55% and 65% respectively). This suggests that people do not treat moral terms the way that they treat natural kind terms. Even if many people happen to think that morality is objective (as the studies by Nichols 2004, and Goodwin & Darley 2008, suggest), they are willing to give up on this belief without abandoning their moral concepts. They are willing to treat those concepts as response-dependent.

I think these results can be best interpreted as follows. Moral concepts are neutral about moral objectivity. People can acquire these concepts without any beliefs about what kinds of properties they designate. This neutrality begets a kind of resistance to error. If there are no objective moral properties, then it wouldn't follow that moral judgments fail to refer; it would mean only that they refer to response-dependent properties. Thus, it is all but guaranteed that some moral judgments will come out true, and to this extent the evidence favors moral realism (defined as the view that there are truthmakers for some moral judgments). Mackie mistakes a popular but dispensable belief about morality for an analytic truth. His error theory rests on an error. In fact, his argument for the error theory may rest on two

mistakes, the second of which we will come to presently. Of course, this is just one study, and other interpretations may be available, but it provides some evidence against Mackie's conceptual claim and shows how empirical findings might be used to explore whether moralizers are, as he suggests, committed to objectivism. Extant empirical evidence suggests otherwise.

2.4 Sensibility vs. moral sense

The survey study just described suggests that one can possess moral concepts without knowing whether moral judgments refer to properties that are objective. The survey also brings out the possibility that people are willing to accept the conclusion that moral truth depends on our responses. But the survey does not settle whether a response-dependent theory is true. This is the next question on the decision tree. As we have seen, Mackie thinks action-guidingness and objectivity are incompatible. This may suggest that he sees no room for a theory that combines moral objectivity with the view that moral judgments have motivational pull. This, however, is Mackie's second mistake. The hypothesis that morality has an emotional basis reveals a way out of Mackie's argument for incompatibility. Emotions are action-guiding in that they motivate us to act. But some emotions may also represent objective features of the world. Fear, for example, may represent danger, and danger may be an objective property. Emotions can represent objective properties in a motivating way: they simultaneously pick up on information while compelling us to respond adaptively. The fact that fear is action-guiding does not rule out the possibility that it is designed by evolution to track objective threats. Likewise, disgust is action-guiding but it may register real sources of contamination.

This brings us back to "icky." This emotionally-expressive term may refer to something objective, like contamination, or to something subjective, such as the tendency to cause feelings of nausea. We can ask whether ickiness is objective or subjective, even if we grant that the word "icky" is expressive. Expressive terms can have objective referents. Likewise, we can ask this question about moral terms. This question frames a historical debate between Francis Hutcheson, who may have believed that our moral sentiments track objective moral truths, and David Hume, who suggests that morality depends on human responses. The claim that moral judgments track objective properties is called the moral sense theory. It seems to have been defended by Francis Hutcheson in the eighteenth century. It may even have been Kant's considered view, since he had an objective procedure for arriving at moral truth, but also insisted that every moral judgment is associated with a moral feeling. The moral sense view finds an analogue in contemporary authors who combine external standards of moral truth with motivationally charged moral psychologies (e.g., Campbell 2007; Copp 2001; see also Railton 2009, who makes a modest move in that direction). The alternative view, which says that moral judgments refer to response-dependent properties, has been called the sensibility theory (McDowell 1985; Wiggins 1987). We can now ask whether there is any way to decide between these options empirically.

I think there is some reason to favor sensibility over moral sense. For the moral sense theory to be true, there would have to be a candidate objective property to which our moral concepts could refer. Unfortunately, I cannot undertake a review of modern moral sense theories here, but I will offer, instead, a more general line of empirically-informed resistance. Moral rules are emotionally conditioned, and communities condition people to avoid a wide range of different behaviors. Within a given society, the range of things that we learn to condemn is remarkably varied. Examples include physical harm, theft, unfair distributions, neglect, disrespect, selfishness, self-destruction, insults, harassment, privacy invasions, indecent exposure, and sex with the wrong partners (children, animals, relatives, people who are married to other people). One might think that all of these wrongs have a common underlying essence. For example, one might propose that each involves a form of harm. But this is simply not true. Empirical evidence shows that people condemn actions that have no victims, such as

consensual sex between adult siblings and eating the bodies of people who die in accidents (Murphy et al. 2000). Furthermore, harm itself is a subjective construct. It cannot be reduced to something like physical injury. Privacy violations are regarded as a kind of harm, even though they don't hurt or threaten health, whereas manual labor is not considered a harm, but it threatens the body more than, say, theft. Similar problems arise if we try to define moral wrongs in terms of autonomy violations. Mandatory education violates autonomy, but it is considered good, and consensual incest is an expression of autonomy, but is considered bad.

Realists would no doubt resist some of these claims, but theirs is an uphill battle. On the face of it, morality lacks a common denominator. Empirical surveys of human values suggest that moral rules are a potpourri, which can be extended and contracted in any number of ways, with no fixed ingredients. Or rather, the common denominator is not a property shared by the things we condemn, but rather by the condemning itself. Moral sense theorists liken morality to perception, and, in so doing, they imply that there is an external feature of the world that our moral sentiments pick up on. But there is little reason to believe this. Unlike perception, there is massive variation in what we moralize, and there is a perfectly good explanation for this: the content of morality is determined by social conditioning rather than by the mind-independent world. Morality is not something we get by simply observing.

The foregoing is offered as an empirical challenge to moral sense theories, not a decisive refutation. Too often philosophers stick with examples of moral norms that clearly concern harm or violations of autonomy. This inflates optimism about a unifying essence. If one uses empirical methods to discover the full range of things that people actually moralize (such as victimless harms), the task of finding a unified essence looks much harder. Moral sense theorists might reply that this diversity is illusory. They might say, for example, that people would stop condemning victimless crimes on reflection. That claim is amenable to empirical testing, and so far the tests provide little support. For example, Murphy et al. (2000) presented people with cases of incest and cannibalism where it was extremely salient that no one was harmed. They invited people to revise knee jerk moral intuitions and rule that, on reflection, these victimless actions are permissible. A piddling 20% revised accordingly, but 80% stuck to their original view. Moral sense theories seem to place their bets on the 20%. The challenge is to explain why the stubborn and considered opinions of the majority are performance errors of some kind.

Given the diversity of things about which people moralize, I think the sensibility theory is more promising than the moral sense theory. Wrongness is projected, not perceived. The property of being wrong is the property of causing negative sentiments, not a response-independent property that those sentiments are designed to detect. This conclusion follows from an inference to the best explanation. Empirically it looks as if there is no common essence to the things that we find morally wrong—a finding that is difficult to explain on the moral sense model, but easy to explain on the assumption that wrongness is response dependent. By analogy, imagine that we catalogue the things that make people laugh, and find that they lack a shared essence. This would imply that laughter does not pick up on an objective property. The things that we find funny are unified by the very fact that we are amused by them. Likewise for the things we find immoral: disapprobation carves the moral landscape.

2.5 Relativism vs. ideal observers

I have just been arguing that moral truth is response-dependent. Moral judgments can be true, but their truth depends on our sentiments. Something is immoral if it causes anger, disgust, guilt, and shame in us. But now we can ask, who does "us" refer to here? Whose sentiments determine moral truth? This brings us to the final question in the metaethics decision tree. Can divergent responses have equal claim to truth?

Empirical evidence strongly suggests that moral sentiments vary, both within and across

cultures. Within a culture, the clearest divisions are between political orientations. Liberals and conservatives have interminable debates, even when they are exposed to the same science and education. Research suggests that these debates come down to fundamental differences in moral values. Conservatives are much more likely than liberals to emphasize purity, authority, and preservation of the in-group in justifying their moral norms (Haidt 2007). These things are foundational for conservatives and largely irrelevant to liberals.

Across cultures, differences are even greater. Everything that we condemn is accepted somewhere else (such as slavery and torture), and things that have been condemned by other cultures (such as women's suffrage) have been embraced by us. There are cultures whose moral outlooks are dominated by considerations that we tend to downplay in the post-industrial West (sanctity and honor, for example), and ideals that are central to our moral outlook appear to be modern inventions (rights and the idea of human equality).

Descriptively, then, people do not seem to have the same moral values, within or across cultures. There is divergence in our sentiments. Some of this divergence might diminish if we filtered out cases where people were reasoning badly or on poor evidence, but there is ample evidence that disagreements remain among people who reason carefully and draw on the same factual knowledge. Indeed, if we filter for good reasoning, divergence might increase rather than decrease: consider professional normative ethicists, who are experts at reasoning but nevertheless arrive at varied and novel moral perspectives that neither converge with each other nor with the communities to which they belong.

I think such descriptive moral relativism provides support for metaethical moral relativism. This would be a terrible inference on its own, as every metaethics textbook points out, but the inference gains plausibility if bolstered by a premise I argued for above: moral truth is dependent on our responses. If responses vary, even under favorable epistemic conditions, and responses determine truth, then the truth of a moral judgment can vary depending on whose values are being expressed.

The ethical universalist can resist this conclusion by offering an antidote to moral variation. The most natural strategy would be to defend universality by developing an ideal observer theory, and to argue that, under ideal epistemic conditions (which might include external factors as well as being an epistemically ideal agent), judges would arrive at the same set of moral values. This strikes me as woefully unlikely. Once we grant that sentimentalism is true, and that our sentiments track response-dependent properties, it's not clear how to settle on which observer is ideal. Two people who have the same factual knowledge may have different sentiments as a result of differences in temperament (Lovett et al. 2012), reward sensitivity (Moore et al. 2011), gender (Fumagalli et al. 2010), class (Côté et al. 2013), and age (Truett 1993). Whose sentiments are right? Moreover, the standard traits associated with ideal observation may be problematic in the moral domain. Should we consult someone who is disinterested when we know, empirically, that distance from a situation can lead to moral indifference? Should we consult someone who has not been conditioned by a particular culture when we know that innate sentiments are unlikely to deliver moral attitudes? Should we consult someone who attends to every detail of a case, when we know that framing, vivid description, and concreteness can alter moral judgments? These problems strike me as insuperable. There are no clear criteria for ideal observation and no reason to believe that careful observers would converge.

In posing this challenge, I am inviting ideal observer theorists to look at empirical findings and propose epistemic standards that would overcome the sources of variation mentioned here. Some ideal observer theories try to be empirically responsive in this way. For example, Smith (1994) advances the hypothesis that ideal rational agents would converge, but he also realizes that some readers might be reluctant to share his optimistic outlook. To quell these doubts he makes three empirical observations (p. 188): there is considerable moral con-

vergence already (he cites the existence of thick concepts as evidence: we all think brutality is bad and honesty is good); there has been moral progress (he cites slavery, among other examples); and entrenched disagreements often reflect faulty rationality, such as religious beliefs. Here, I think further empirical scrutiny would weaken Smith's case. Divergence is rampant, and people disagree on the scope of thick concepts (is torture brutal? is espionage dishonest?). Cases of (what we consider to be) moral progress are, I've noted, often driven by economic upheavals and other irrational factors, with reasoning playing a post-hoc role. Finally, disagreements remain after bad reasoning and religiosity are controlled for; the examples mentioned, in formulating the challenge include things such as temperament and framing effects. I think empirical evidence provides little reason to expect that rational and informed observers would deliver consistent verdicts.

In light of such worries, universalists might abandon the ideal observer theory and offer instead a procedural approach to consensus, arguing that people would and should converge if they arrived at their sentiments in the right way. For example, many people might agree that it is good to arrive at decisions democratically, taking multiple sentiments into consideration, and we might sentimentally endorse the outcome of democratically-resolved moral disputes. Though I cannot make the case here, I suspect the problems with such a procedural approach outweigh its prospects. Democratic decision-making does not result in moral consensus; it can even polarize. When such procedures increase consensus it is often through power and prestige rather than sentimental convergence. Our faith in democratic procedures may also be an expression of moral relativism rather than a solution. Democratic procedures are an historical anomaly, which emerged in the modern period with the rise of capitalism, and they have often been used to oppress minorities and to impose the values of the many over the few. Perhaps such procedures are an improvement over totalitarian forms of decision-making, but they do not remedy relativism. Indeed, as societies move towards consensus-building procedures, they may actually promote variation, leading to an endless proliferation of values and an ever widening gulf between those who cherish diversity and those who reside in more traditional societies. From a social science perspective, the prospects for a universal morality look grim.

Once the case for relativism is established, the question arises: relative to what? Are moral judgments relative to value systems? Are those systems individuated at the scale of cultures and subcultures or do they vary across individuals? Little empirical work has been done to address this question, but let me end with a suggestion about how to proceed. When examining the semantics of natural kind terms, philosophers have sometimes appealed to a linguistic division of labor (Putnam 1975). We defer to experts and thereby license them to adjudicate the boundaries between natural kinds. Now we can ask, is there such a thing as moral expertise? Do we appeal implicitly or explicitly to moral experts? Would we change our moral judgments if the designated members of our community told us we were morally mistaken? We don't know the answers to such questions, because moral expertise has not been intensively studied. I suspect there will be considerable individual differences, with members of more traditional societies showing more willingness to defer. But I also suspect that deference in the moral domain will be less prevalent than for natural kinds; we are more inclined to take ourselves as having morally authoritative insight. What is most clear, however, is whether the scope of the relativity depends ultimately on how we use moral concepts and terms; and this is something that can be investigated empirically. Naturalizing relativism will require the marriage of cultural anthropology and sociolinguistics. From the armchair, it is tempting to think there is a single true morality; introspective reflection tends towards solipsism.

3 Conclusion

Throughout this discussion, we have worked our way down a metaethics decision tree. I have

made a case for a relativist cognitivist sentimentalist sensibility theory. Admittedly, each of my arguments is only a first pass, and much more could be said for and against these positions. Many of the empirical findings that I have described are preliminary. My main goal here is not to make a decisive case for any position in metaethics. Rather, I am pleading for the relevance of empirical methods in doing this traditionally philosophical work. Moral philosophy is undergoing a process of naturalization. This has been felt most strongly in normative theory (e.g., the debate about the status of character in virtue ethics) and moral psychology (e.g., questions about how deontological and consequentialist judgments are made). I hope to have shown that the empirical work also bears directly on metaethical questions—questions about what, if anything, is the source of moral truth.

Empirical work cannot replace philosophical toil. We need philosophy to pose questions and identify possible theories. Experimental design is itself a kind of philosophical reasoning, and it takes considerable argumentation to move from data to theory. Naturalization is supplementation, not usurpation. But it is not just supplementation. The empirical arsenal may just be our best hope for adjudicating philosophical debates. Reflection can delineate the logical space, but we need observation to locate ourselves therein. Philosophers have always relied on observation, in some sense, but scientific methods allow us to observe processes that are unconscious, inchoate, or distant in space and time. Empirical studies can test the content, prevalence, and malleability of intuitions, and they can also tell us where our intuitions come from—a question of central metaethical concern. We should embrace any tools that help us resolve the questions that we are employed to answer. A century ago, there was a linguistic turn, and philosophers began to treat traditional philosophical problems as amenable to semantic analysis. Around the same time, the boundary between philosophy and psychology was still blurred, and journals such as *Mind* published articles that we might now classify as psychological. Such crossovers fell out of fashion, however, and it has taken a century to get back to this incipient moment. With the linguistic turn, Anglophone philosophers became convinced that we should all learn logic because it would help us make progress. Logic did help, and it did not undermine philosophy. Now, we can encourage all philosophers to learn about methods and results used in the relevant social and physical sciences. The payoff of this naturalistic turn may be vastly greater than the linguistic turn. Science, not formal logic, is positioned to tell us whether morality is a human construction.

Acknowledgments

This discussion has benefited immeasurably from the feedback of anonymous referees and from Ying-Tung Lin, Jessica McCormack, Thomas Metzinger, and Jennifer Windt. I am grateful for their close reading and helpful suggestions.

References

Blackburn, S. (1998). *Ruling passions.* Oxford, UK: Oxford University Press.

Blair, R. J. R. (1995). A cognitive developmental approach to morality: Investigating the psychopath. *Cognition*, *57* (1), 1-29.

Borg, J., Hynes, C., van Horn, J., Grafton, S. & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: An fMRI investigation. *Journal of Cognitive Neuroscience*, *18* (5), 803-817.

Boyd, R. (1988). *Essays on moral realism.* Ithaca, NY: Cornell University Press.

Brandt, R. (1959). *Ethical theory: The problems of normative and critical ethics.* Englewood Cliffs, NJ: Prentice-Hall.

Campbell, R. (2007). What is moral judgment? *Journal of Philosophy*, *104* (7), 321-349.

Copp, D. (2001). Realist-expressivism: A neglected option for moral realism. *Social Philosophy and Policy*, *18* (2), 1-43. 10.1017/S0265052500002880

Côté, S., Piff, P. K. & Willer, R. (2013). For whom do the ends justify the means? Social class and utilitarian moral judgment. *Journal Of Personality And Social Psychology*, *104* (3), 490-503. 10.1037/a0030931

Doris, J. M. (1998). Persons, situations, and virtue ethics. *Noûs*, *32* (4), 504-530.

Eskine, K. J. (2011). *From perceptual symbols to abstraction and back again: The bitter truth about morality.* New York, NY: Doctoral dissertation, Department of Psychology, City University of New York.

Eskine, J. K., Kacinik, A. N. & Prinz, J. J. (2011). A bad taste in the mouth: Gustatory disgust influences moral judgment. *Psychological Science*, *22* (3), 295-299. 10.1177/0956797611398497

Firth, R. (1952). Ethical absolutism and the ideal observer. *Philosophy and Phenomenological Research*, *12* (3), 317-345.

Flanagan, O. (1991). *Varieties of moral personality: Ethics and psychological realism.* Cambridge, MA: Harvard University Press.

Fumagalli, M. M., Ferrucci, R. R., Mameli, F. F., Marceglia, S. S., Mrakic-Sposta, S. S., Zago, S. S. & Priori, A. A. (2010). Gender-related differences in moral judgments. *Cognitive Processing*, *11* (3), 219-226. 10.1007/s10339-009-0335-2

Gibbard, A. (1990). *Wise choices, apt feelings.* Cambridge, MA: Harvard University Press.

Goodwin, G. P. & Darley, J. M. (2008). The psychology of meta-ethics: Exploring objectivism. *Cognition*, *106* (3), 1339-1366. 10.1016/j.cognition.2007.06.007

Graham, J., Haidt, J. & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96* (5), 1029-1046.

Greene, J. D. (2007). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.) *Moral psychology, Vol. 3: The neuroscience of morality, emotion, disease, and development.* Cambridge, MA: MIT Press.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M. & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293* (5537), 2105-2108. 10.1126/science.1062872

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316* (5827), 998-1002. 10.1126/science.1137651

Hare, R. D. (1993). *Without conscience: The disturbing world of the psychopaths among us.* New York, NY: Pocket Books.

Huebner, B., Dwyer, S. & Hauser, M. (2009). The role of emotion in moral psychology. *Trends in Cognitive Sciences*, *13* (1), 1-6. 10.1016/j.tics.2008.09.006

Joyce, R. (2006). *The evolution of morality.* Cambridge, MA: MIT Press.

Kant, I. (1797). *The metaphysics of morals.* M. J. Gregor (Trans.). Cambridge, UK: Cambridge University Press.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M. & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature*, *446*, 908-911. 10.1038/nature05631

Kornblith, H. (Ed.) (1985). *Naturalizing epistemology.* Cambridge, MA: MIT Press.

Kramer, P. D. (1993). *Listening to Prozac.* New York, NY: Viking.

Lovett, B. J., Jordan, A. H. & Wiltermuth, S. S. (2012). Individual differences in the moralization of everyday life. *Ethics & Behavior*, *22* (4), 248-257. 10.1080/10508422.2012.659132

Mackie, J. L. (1977). *Ethics: Inventing right and wrong.* London, UK: Penguin.

McDowell, J. (1985). *Morality and objectivity.* London, UK: Routledge & Kegan Paul.

Mikhail, J. (2002). Aspects of the theory of moral cognition: Investigating intuitive knowledge of the prohibition of intentional battery and the principle of double effect. *Economics Research Paper*, *762385*. http://ssrn.com/abstract=762385

Moore, A. B., Stevens, J. & Conway, A. A. (2011). Individual differences in sensitivity to reward and punishment predict moral judgment. *Personality And Individual Differences*, *50*, 621-625. 10.1016/j.paid.2010.12.006

Murphy, S., Haidt, J. & Björklund, F. (2000). Moral dumbfounding: When intuition finds no reason. Unpublished manuscript, Department of philosophy, University of Virginia.

Nichols, S. (2004). *Sentimental rules: On the natural foundations of moral judgment.* New York, NY: Oxford University Press.

——— (2008). Sentimentalism naturalized. In W. Sinnott-Armstrong (Ed.) *Moral psychology: The cognitive science of morality: Intuition and diversity* (pp. 255-274). Cambridge, MA: MIT Press.

Olasov, I. (2011). Register variation and the moral cognitivism debate. Unpublished manuscript, City University of New York, Graduate Cente.

Prinz, J. J. (2004). *Gut reactions: A perceptual theory of emotion.* New York, NY: Oxford University Press.

——— (2005). Imitation and moral development. In S. Hurley & N. Chater (Eds.) *Perspectives on imitation: From cognitive neuroscience to social science.* Cambridge, MA: MIT Press.

——— (2007a). Is morality innate? In W. Sinnott-Armstrong (Ed.) *Moral psychology, vol 1: Evolution of morals* (pp. 367-406). Cambridge, MA: MIT Press.

——— (2007b). *The emotional construction of morals.* Oxford, UK: Oxford University Press.

Putnam, H. (1975). The meaning of "meaning". *Mind, Language and Reality: Philosophical Papers, Volume 2* (pp. 215-271). Cambridge, UK: Cambridge University Press.

Quine, W. V.O. (1969). *Ontological relativity and other essays.* New York, NY: Columbia University Press.

Railton, P. (1993). What the non-cognitivist helps us to see the naturalist must help us to explain. In J. Haldane and C. Wright (Eds.), *Reality, Representation and Projection* (pp. 279-297). Oxford, UK: Oxford University Press.

——— (2009). Internalism for externalists. *Philosophical Issues*, *19* (1), 166-181. 10.1111/j.1533-6077.2009.00165.x

Ruse, M. (1991). A companion to ethics. In P. Singer (Ed.) *A companion to ethics* (pp. 500-510). Oxford, UK: Blackwell.

Sayre-McCord, G. (Ed.) (1988). *Essays on moral realism.* Ithaca, NY: Cornell University Press.

Schmidt, E. Z. & Bonelli, R. M. (2008). Sexuality in Huntington's disease. *Wiener Medizinische Wochenschrift*, *158* (3-4), 84-90. 10.1007/s10354-007-0477-8.

Seidel, A. & Prinz, J. J. (2013a). Sound morality: Irritating and icky noises amplify divergent moral domains. *Cognition*, *127* (1), 1-5. 10.1016/j.cognition.2012.11.004

——— (2013b). Mad and glad: Musically induced emotions have divergent moral impact. *Motivation and Emotion*, *37* (3), 629-637. 10.1007/s11031-012-9320-7

Silk, J. B., Brosnan, S. F., Vonk, J., Henrich, J., Povinelli, D. J., Richardson, A. F., Lambeth, S. P., Mascaro, J. & Schapiro, S. J. (2005). Chimpanzees are indifferent to the welfare of other group members. *Nature*, *435*, 1357-1359. 10.1038/nature04243

Smith, M. (1994). *The moral problem.* Oxford, UK: Blackwell.

Truett, K. R. (1993). Age differences in conservatism. *Personality and Individual Differences*, *14* (3), 405-411. 10.1016/0191-8869(93)90309-Q

Turiel, E. (1983). *The development of social knowledge: Morality and convention.* Cambridge, UK: Cambridge University Press.

Waldmann, M. R., Nagel, J. & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak & R. G. Morrison (Eds.) *The Oxford handbook of thinking and reasoning.* Oxford, UK: Oxford University Press.

Wastell, C. & Booth, A. (2003). Machiavellianism: An alexithymic perspective. *Journal of Social and Clinical Psychology*, *22* (6), 730-744. 10.1521/jscp.22.6.730.22931

Widom, C. S. (1976). Interpersonal conflict and cooperation in psychopaths. *Journal of Abnormal Psychology*, *85* (3), 330-334. 10.1037/0021-843X.85.3.330

Wiggins, D. (1987). A sensible subjectivism. In D. Wiggins (Ed.) *Needs, values, truth: Essays in the philosophy of value* (pp. 185-214). Oxford, UK: Blackwell.

Woolfolk, R. L. (2013). Experimental philosophy: A methodological critique. *Metaphilosophy*, *44* (1-2), 79-87. 10.1111/meta.12016

Wrangham, R. (2004). Killer species. *Daedalus*, *133*, 25-35.

Conceptualizing Metaethics

A Commentary on Prinz

Yann Wilhelm

In this commentary on Prinz's "Naturalizing Metaethics" I shall first look briefly at his methodological assumptions. I will argue that Prinz's approach is more radical and less conciliatory between analytical and empirical approaches than it seems from his own description. In the second part of my commentary, I shall look at one possible objection to Prinz's sentimentalism: the evidence he presents does not provide the needed modal strength for sentimentalism. I shall present two example of this objection, and argue that Prinz's own depiction doesn't adequately represent it. I shall then use the helpful distinction offered by Jon Tresan between *de dicto*- and *de re*-internalism to analyze underlying problems in the objection. I will present another way of reacting to it, which I think fits nicely with Prinz's naturalized methodology. In the last part, I shall look at his critique of non-cognitivism. Prinz argues that non-cognitivism makes certain linguistic predictions that turn out to be wrong: if non-cognitivism were true we would expect our moral language to reflect this. I will argue that there are many forms of non-cognitivism that predict this surface grammar. The key idea is that non-cognitivism entails a pragmatic theory of moral language. I then offer a speculative explanation about why the moral language has its surface form. This speculation, I argue, has at least the same amount of plausibility as cognitivist theories. Furthermore, this possible explanation is open to empirical investigation. I agree with Prinz that, ultimately, metaethical theories should be tested against empirical evidence. Prinz presents conceptual and empirical work as mutually enhancing enterprises. My commentary is, I hope, a small contribution highlighting the conceptual side of the coin.

Keywords
Cognitivism | De dicto-internalism | De re-internalism | Metaethics | Methodological naturalism | Motivational internalism | Non-cognitivism | Sentimentalism

Commentator

Yann Wilhelm
ywilhelm@students.uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Target Author

Jesse Prinz
jesse@subcortex.com
City University of New York
New York, NY, U.S.A.

1 Metaethics under empirical scrutiny

Prinz proposes to naturalize metaethics. Metaethics is traditionally regarded as a second-order discourse about ethics. Where normative ethics asks what is good and what is bad, what we should or shouldn't do, metaethics asks the question of what morality is itself (DeLapp 2011). Its subject is the ontology of moral properties, the semantics of moral discourse, the epistemic foundation of moral judgments and the psychology of moral opinions. These different aspects are highly interrelated—answers in one area influence questions asked in others.

There are many different ways to tackle the question of what morality itself actually is. Prinz characterizes metaethics as being concerned with the foundations of moral judgments (Prinz this collection, p. 1). This is his starting point, which shapes his decision tree. He acknowledges that one could arrange the tree in different ways, depending on which aspect one wants to pull into focus.

Prinz's primary goal is to show that every question in the decision tree is empirically tractable (this collection, p. 1). This is his *methodological naturalism* (p. 2).[1] He argues that we

1 He contrasts this with *metaphysical naturalism* and *semantic naturalism*. The former says that everything there is belongs to the natural world. The latter tries to reduce concepts from various domains in terms that are more likely to be naturalized in the metaphysical sense.

Wilhelm, Y. (2015). Conceptualizing Metaethics - A Commentary on Prinz.
In T. Metzinger & J. M. Windt (Eds). *Open MIND:* 30(C). Frankfurt am Main: MIND Group. doi: 10.15502/9783958570689

should study the domain of metaethics empirically. He wants to test "[...] theories derived from philosophical reflection against the tribunal of empirical evidence" (Prinz this collection, p. 5).

Metaethics, according to him, is not the sole matter of armchair reflection. This seems natural when we characterize metaethics as the question of what morality itself is. But that goes against the view that metaethics—or philosophy in general—is not concerned with what actually is the case, but with what *must* be the case. What are the *necessary* conditions of morality? On this view, metaethics is concerned with statements that hold *a priori*. Most of the time this means deriving knowledge from reflection upon the meaning of our concepts. This method of *conceptual analysis* had been at the core of philosophy since the *analytic turn* (Prinz this collection, p. 3).

Against this turn Prinz sets the *empirical turn* (this collection, p. 3). He describes this development as an enrichment of the philosopher's tool box. Where conceptual considerations help us to formulate theories and flesh out the differences between different views, empirical methods confirm the theories derived from this work. The former pose questions and formulates possible answers; the latter test those answers. Prinz emphasizes that empirical and traditional approaches are not opposed to one another (this collection, p. 5). Rather, they complement each other. They're more like opposing points on a continuum of methods for exploring the world.

It is important to see that this view is not as conciliatory between traditional analytic philosophy and empirical philosophy as it might seem. It does not leave room for *a priori* armchair reflection. In fact, Prinz even regards conceptual analysis as an empirical task: "[A]rmchair conceptual analysis can be characterized as an introspective memory retrieval process. As such, it can be regarded as a form of observation" (2008, p. 191).

When Prinz speaks of "traditional methods", he does not include conceptual analysis as an *a priori* enterprise. Rather, he is referring to various tools, for example formal semantics or logic, which help us articulate theories. They are tools for exploring the natural world, from which we gain knowledge only through experience. Prinz is a radical empiricist at heart.

An empirical scientist could ask: "What differentiates this from my own work?" For she, too, reflects upon different theories, how they relate to each other, formulates questions, and so on. This is an important part of scientific, empirical work. I think Prinz would agree. An important upshot of his naturalized philosophy is that there are no clear-cut borders between philosophy and psychology (Prinz 2008, pp. 204–206). They are different disciplines not because of their different subject areas or methods but for pragmatic reasons. They are different *academic* disciplines, shaped by sociological and historical processes. The borders between the different disciplines become blurred in the empirical turn. According to Prinz, this is a good thing.

I think this the real strength of Prinz's approach. Arguably many disciplines are divided largely by pragmatical differences, like education and academic organization. Instead of demarcating different approaches, instead of drawing sharp lines between them, Prinz proposes that we unite them in the search for explanations of the natural world.

Prinz's target article is a very good example of this approach. Here I want to make a few remarks in the spirit of Prinz's own methodology. In the next section I will focus on a specific objection against Prinz's answers to the first question in the decision tree. I think that it can clarify some consequences of his methodological naturalism for metaethics.

2 Internalism and modal strength

In this section I discuss Prinz's answer to a potential objection to his sentimentalism, namely, that the evidence lacks *modal strength*. In fact, objections of this kind have already been raised against Prinz's and other naturalistic metaethical theories already. I shall first argue that his answer doesn't get to the heart of the objection. Second, I propose a way in which Prinz can and should answer it. To do this I shall present two instances where this objection has been made. A helpful distinction by Jon

Tresan will then show that there are actually two kinds of internalist theses at play here. Only one of these is really relevant for Prinz's naturalized metaethics, I shall argue. The objection then loses its force in light of Prinz's project of a naturalistic methodology. The following reasoning can also be seen as a small case study in recent (naturalized) metaethics.

The first question in Prinz's decision tree is whether moral judgments are essentially affect-laden or not. This is Prinz's take on the internalist-externalist debate.[2] This debate is a classical debate in metaethics that can be traced back to the British moralists (Darwall 1995). It concerns the question of whether *motivation* is *internal* or *external* to moral judgments. Do moral judgments necessarily involve motivation to act accordingly? Or does the motivation come from a desire external to them (e.g., the desire to be a good person)?[3]

Prinz advocates a position that he calls sentimentalism:

> Sentimentalism $=_{\text{Df}}$ Moral Judgments essentially involve affective states, such as emotions, in one of two ways: such states as constituent parts of moral judgments (traditional sentimentalism); or moral judgments are judgments about the appropriateness of such states (neo-sentimentalism). (Prinz this collection, p. 6)

The evidence for a link between moral judgments and emotions is overwhelming (Prinz this collection, p. 10). But is it enough to warrant a stronger relation than mere accompaniment? Even if we grant Prinz the interpretation that affective states are not only mere *consequences* of moral judgments, could we not still question whether they are essential components of moral judgments? The objection is this: the empirical evidence lacks *modal strength* to support sentimentalism. Even if all our ordinary moral judgments are based on emotions, it could still be *possible* to judge dispassionately (Prinz this collection, p. 13). Therefore the evidence doesn't support sentimentalism.

Prinz answers that the empirical evidence gives us enough reason to infer that we *cannot* make moral judgments without emotions: "Every study suggests that emotions arise when we make moral judgments. All evidence also suggests that when emotions are eliminated, judgments subside as well" (Prinz this collection, p. 13).

According to Prinz, the theory that emotions are essential components of moral judgments explains the total pattern of data better than its rivals (this collection, p. 14). Furthermore, he argues that the sentimentalist can accept psychologically exotic cases, in which the connection between moral judgments and emotions doesn't occur, which conform rival theories.

This answer, I argue, misses the real core of the objection. Prinz confronts it upfront and just states what it questions. He puts the objection in the following way:

> The evidence shows that emotions are often consulted when making moral judgments, but this leaves open the possibility that we might also make moral judgments dispassionately under circumstances that have not yet been empirically explored. (Prinz this collection, p. 13)

But this does not represent the objection adequately. The objection doesn't rest on possible, not-yet-found empirical evidence against sentimentalism. Rather, it rests on opposing ideas about what kind of modal strength claims about the relation between moral judgments and emotions should possess. At the heart of this objection there is no disagreement about the empirical evidence, but an opposition in the underlying methodology.

Adina Roskies, for example, accepts that "[…] those [brain] areas involved in moral judgments normally send their output to areas involved in affect, resulting in motives

2 Although he doesn't explicitly put it like this, I think it's safe to frame it in this way. The option that denies affect-ladeness is called "externalist moral realism", and he states in various places that emotions are motivating or action-guiding (Prinz this collection, pp. 8, 11, 21). And one answer to the third question is a position called "internal realism". What I say about internalism in the following therefore applies equally to Prinz's sentimentalism. See also Prinz (2006), where he explicitly states motivational internalism.

3 See Björklund et al. (2012) for a short overview.

that in some instances cause us to act" (2008, p. 192).

But she thinks that this is not enough for internalism to be true.[4] In her view there is a connection between the cognitive and the affective system, but "this link is causal and thus contingent and not constitutive" (Roskies 2008, p. 192). In this sense the connection, according to her, is not necessary.

Antti Kauppinen sees the difference between internalism and externalism in a similar way. He depicts internalism as saying that there is a link between moral judgments and motivation that holds a priori and with conceptual necessity. externalism, in contrast, is the view that this link is contingent and a posteriori (Kauppinen 2008, p. 3). For Kauppinen, every internalist position then becomes an externalist position if it weakens the modality of the claim. When a metaethical account doesn't claim that the connection between moral judgments and motivation holds a priori and by necessity, it is an externalist account. No amount of empirical data can refute this criticism.

In Kauppinen's case the disagreement with Prinz about the underlying methodology is clear. He reacts to the proposal by Roskies, Prinz, and Alfred Mele (among others) that we clarify the debate empirically (Kauppinen 2008, 4). Because of his definitions of internalism and externalism as conceptual necessary claims he argues that "[...] findings in either actual or fictional experimental psychology or neuroscience have little relevance to the debate" (Kauppinen 2008, p. 4).

Kauppinen is opposed to methodological naturalism in philosophical moral psychology (2008, p. 4). That is why he would not be satisfied with Prinz's answer to this objection. Against him, Prinz would have to defend his metaethical naturalism. Interestingly enough, Roskies, on the other hand, thinks that we *can* clarify metaethical debates empirically.

In what follows I shall show how I think Prinz should meet this objection. Furthermore, I will argue that everyone who wants to apply empirical data to metaethical debates, such as, e.g., Adina Roskies, should side with Prinz on his methodological naturalism and accept internalism as a true a posteriori theory about moral judgments.

I will now present an analysis of the internalism–externalism debate offered by Jon Tresan that I think will be very helpful here (2009). He distinguishes different formulations of internalism along various dimensions. He claims that a very important distinction has been overlooked: most philosophers in the debate neglected the difference between the modality of the internalist claim and the stated relation between moral opinions and motivation. According to Tresan, there are two different kinds of necessity that can occur in such claims: wide-scope necessity, which operates over the entire proposition—*de dicto*—and narrow-scope necessity, which operates over the predicate—*de re* (Tresan 2009, p. 54). The first operates on the dimension of *Modality* and the second on the dimension of *Relation* (Tresan 2009, p. 55).

For example, the statement that parents have children can be formulated with both kinds of necessities:

> Necessarily, parents have children (*de dicto*).
>
> Parents have, necessarily, children (*de re*).

In the first case the proposition that parents have children is stated as holding necessarily. Parents have children, otherwise they would not be called parents. If someone has a child, she is a parent. But the second statement says that people who are parents have their kids necessarily. But this is obviously false. John and Mary don't have their children necessarily. They could easily never have had any children at all. True, they would not, then, be parents – but the fact that they are parents may have, initially, been quite accidental. We can easily see that there is a difference between *de dicto*- and *de re*-necessities because these two statements can have different truth-values at the same time.

4 Her critique is directed at internalism, not sentimentalism. But I regard both positions as similar enough to treat Roskies's critique as an argument against Prinz's sentimentalism (see also above). At the core of both positions is the connection between moral judgments and affective (motivational) states.

With this distinction at hand we can distinguish two different internalist theses: a strong Modality/weak Relation or *de dicto*-internalism, and a weak Modality/strong Relation or *de re*-internalism. The former states that, with necessity, there is a connection between moral judgments and motivation. The latter says that there is a necessary connection between these two things.

Tresan uses this distinction to argue that something has gone fundamentally wrong in the internalism–externalism debate. The neglect of the two features has led to the *internalist fallacy*: the strength in Modality of an internalist claim was taken to be strength in Relation, which led to an overestimation of the epistemic value of the claim (Tresan 2009, p. 55). The classical debate stated the connection between moral judgments and motivation in terms of conceptual necessity (a *de dicto*-internalism) (see Roskies's and Kauppinen's accounts above). Arguments for this claim were supposed to evoke the intuition that no one can make a moral judgment without being motivated to act. If we have such intuitions, the arguments go, the connection is a conceptual necessity. Likewise, arguments against this internalist claim consisted in thought experiments that were supposed to evoke contrary intuitions.

From Tresan's distinction follows that claims with *de dicto* necessity are claims about our concepts and not about the subject matter (2009, p. 57). *De dicto*-internalism, then, is a claim about our concept "moral judgment" and *de re*-internalism a claim about the subject matter—the phenomenon of moral judgments.

Returning to Prinz (and to Roskies's proposal), we can now see that there are really two empirical questions we can ask: First, what is our concept of "moral judgment"? And second, what are moral judgments? Traditionally the first was not regarded as an empirical question. Philosophers probed their intuitions and just assumed that others shared them. Prinz, on the other hand, regards these kinds of questions as empirical in nature and presents his own survey studies that probes *folk intuitions*. He concludes that most people do consider emotions necessary for moral judgments (Prinz this collection, p. 10; for other studies on this with different results see also Nichols 2002, p. 22; Strandberg & Björklund 2013, p. 325; Björnsson et al. 2014, p. 16).

These studies can answer the first question regarding our concept of moral judgments. But, as Prinz rightly points out, people could be wrong (Prinz this collection, p. 10). These studies do not tell us anything about the subject matter. This is a further point Tresan makes. He argues that even if we have internalist intuitions this is not enough to support internalism. He argues that strength in modality is not interesting for a substantial theory of moral opinions. A claim with strong modality doesn't tell us more about the subject of the claim than the same claim without it. That, necessarily, bachelors are unmarried (*de dicto*) tells us nothing more than that they need to be unmarried to be called bachelors. It's a claim about our concept "bachelor". It tells us simply that the subjects are unmarried—the same as this exact claim without modality tells us. But if bachelors were necessarily unmarried (*de re*) this would be bad news for the subjects and would tell us something substantial about them—that they're essentially unmarried, that they, the individuals, are unable to be married. He concludes that "[i]f we are interested in the nature of the Subject Matter, we must look to Relation not Modality" (Tresan 2009, p. 57; emphasis in original).

Only an internalist claim with a strong relation is interesting. But Tresan thinks that there are no arguments for a *de re*-internalism, which would tell us something interesting and substantial about the subject matter. A *de re*-internalism that states a strong Relation is wrong. This is because our intuitions regarding moral judgments and motivation can only support a *de dicto* internalism (Tresan 2009, p. 64). And traditional arguments for internalism provoke only such intuitions.

I think it is clear that Tresan misses one important possible source of evidence for a strong relation: empirical evidence. Here lies the connection to Prinz's work. The empirical findings, which he collected, all point to a strong relation between moral judgment and affective states. I take Prinz to be looking for a strong

Relation when he says that emotions are an "essential component" of moral judgments (Prinz this collection, p. 12).

What I have tried to show here is the following. Prinz raises a potential objection against his own sentimentalism: the relation between moral judgments and emotions lack modal strength. He answers by saying that we have enough evidence to conclude their necessary connection. I argued that this is not a satisfying answer because it misses the core of the objection.

I think the evidence that he has collected points to a strong Relation between moral judgments and affective (motivational) states. Therefore Prinz has an answer to objections that call this strong relation into question. But this is not an answer to an objection that operates with a *de dicto* internalism. Underlying these objections is an opposition to methodological naturalism in general. Antti Kauppinen is one example of someone holding this position (2008, p. 4). Kauppinen does not think we should ask what moral judgments *actually* are. In his view, metaethics is concerned with what moral judgments *necessarily* are. "This takes us from the realm of the actual to the realm of the metaphysical or conceptually possible, and thus beyond the empirical and the observable" (Kauppinen 2008, p. 22).

The evidence that Prinz presents in the target paper doesn't suffice to refute this position. But I hope to have shown that this need not be a cause of concern for Prinz, because this kind of necessity takes us away from the subject matter. At the heart of Prinz's account lies an interest in moral judgments as a natural phenomenon that we should study by empirical means.

Adina Roskies, on the other hand, is sympathetic to empirical philosophy. One of her aims in the internalist–externalist debate was to show that "[...] moral philosophy need not be, and perhaps ought not be, exclusively a priori" (Roskies 2003, p. 2003).

But this is in contrast to her understanding of the required modality of the internalist claim, as I tried to show using Tresan's analysis. If we want to clarify those kinds of debates empirically, it's not enough to just take traditional philosophical claims and look for evidence in their favor or evidence that can refute them. We have to formulate them as a posteriori synthetic claims that are part of a bigger explanatory project (Björnsson 2002, p. 329).

I hope that this can shed more light on the implications of naturalistic metaethics for philosophical claims. They shouldn't be regarded as conceptual a priori claims, but as hypotheses that need empirical confirmation. Naturalistic metaethics is not concerned with a priori conceptual necessities. It requires revising our concepts when they don't fit into the best theories. In that sense empirical philosophers should be revisionists (see Francén 2010, pp. 137 and 142 for a more detailed account of revisionism).

Before I go on, I want to offer one last thought about this. What might be the motivation for framing these positions as claims about conceptual necessity? Roskies writes:

> I take it that internalist philosophers have intended to offer something stronger than contingent claims about human wiring (...) Only a view involving necessity or intrinsicality can distinguish moral beliefs and judgments from other types by their special content. (2008, p. 193)

But why do we need a priori conceptual necessities to distinguish between different kind of beliefs and judgments? We could start with very simple observations. Apparently people play a game of blaming and blessing: they use words like "good" and "bad" that are somehow different than other terms. The task of defining what morality is could be a descriptive anthropological enterprise. And I think this is in the spirit of naturalistic metaethics.

I have argued that it is enough for Prinz's sentimentalism (and for internalism) to claim a strong Relation between moral judgments and emotions. But what kind of Relation is strong enough for it? A mere statistical connection is surely not enough. If the important part of the sentimentalist thesis is not the Modality of the whole claim, we have to analyze the terms "ne-

cessary" and "essential" in a non-modal way. One possibility, that harmonizes with naturalized metaethics, is to regard this connection as *functional*.[5]

In the next, and final, section I shall look at Prinz's critique of non-cognitivism. I shall present a speculative alternative to his view that I hope, again, is in agreement with his proposal for a naturalized metaethics.

3 Defending non-cognitivism as an empirical theory

Here, I want to argue against Prinz's attack on non-cognitivism. He thinks that there are good empirical reasons to reject non-cognitivism. His first argument is that cognitivism can predict the surface form of moral language better than non-cognitivism. First, I argue against this by pointing to non-cognitivist accounts of moral language that I think can predict this surface form. Second, I provide a speculative non-cognitivist theory of why moral language has the surface form we can observe. Again, I think my proposal is in agreement with Prinz's naturalized metaethics. I do think, however, that it challenges him to explore the space of possible accounts. My proposal shows, I hope, that the empirical evidence cannot, at this point, decide this question.

The second question in Prinz's decision tree is whether or not moral judgments are truth-apt. Can they be true or false? Theories that answer yes to this question are cognitivistic, while theories that answer negatively are non-cognitivistic. *Non-cognitivism* is a collective term that can refer to many different theories (Shafer-Landau 2003, p. 17). It consists of two theses (Roojen 2013, section 1.1): the first says that moral utterances do not express propositions; they're not truth apt. This is a semantic thesis about moral language. The second thesis says that moral beliefs are not representational. They do not refer to anything in the world. This is a thesis about the mental state of the moral agent. Here Prinz wants to defend cognitivism by providing empirically-informed reasons to reject non-cognitivism. He defines expressivism in the following way (we can think of Expressivism as one form of the first, semantic thesis of non-cognitivism):

> Expressivism $=_{\mathrm{Df}}$ Moral assertions express mere feelings or non-assertoric attitudes, and do not purport to convey facts. (Prinz this collection, p. 7)

Prinz denies both of the two theses that make up non-cognitivism. He argues that the most obvious empirical prediction of non-cognitivism fails, as he thinks that if non-cognitivism was true we would expect our moral language to have a non-cognitive form (Prinz this collection, p. 16). But this is not the case. It seems that our moral language mostly has declarative form.

If this is correct, and if I don't have reasons to disbelieve it, does it mean that non-cognitivism makes wrong predictions? I don't think this is the case. Much of the work in non-cognitivism is dedicated to explaining this apparent tension. But I don't think that this involves "elaborate logics", as Prinz puts it (this collection, p. 16). Rather, most non-cognitivists provide theories about the nature of moral discourse that show that we should expect the surface grammar to be declarative. I don't think that non-cognitivism has or needs to have these "obvious empirical predictions".

The starting point is to look at the way language is used. It is not the literal meaning of ethical terms that are of interest but their *function* (Björnsson 2002, p. 328). Expressivism entails a pragmatist theory of moral language:

> [T]he pragmatist attempts to describe the function that a word, phrase or concept plays in human life, and once he has satisfied his curiosity there, he does not think that there are any further questions to ask about utterances of that sort. (Smyth 2014, p. 608)

Arguably, such a pragmatist view is easier to naturalize because we have the social sciences,

5 For this proposal see Björnsson & Francén Olinder (2013) and Bedke (2009) and Schulte (2012). They detail the idea that we can think of this relation as *teleo-functionalistic*.

which offer large toolboxes for investigating human practices.

Although Prinz's definition of expressivism may be at the heart of non-cognitivism, in most cases this is not the whole story. According to expressivism, moral terms are not only used to express ones attitudes but also to provoke certain attitudes in the hearer. This idea goes back to the early emotivists. The "dynamic use" of language (Stevenson 1937, p. 21) involves the manipulation of others: "[E]thical terms are instruments used in the complicated interplay and readjustment of human interests" (Stevenson 1937, p. 20; emphasis in original).

Stevenson, and many others following him, analyze expressions like "x is good" as meaning "Hooray for x! Do hooray as well!" (Stevenson 1937, p. 25).[6] It expresses the speakers attitude and the wish or the prescription that the hearer should adopt this attitude as well.

At this point Prinz could reiterate his point and simply ask: "Why then do we say 'this is good' and not 'I like this, do so as well'?" Here I want to offer a speculative answer: because we don't like to be manipulated. If the function of moral language is, at least in part, to influence the attitudes and the behavior of others, I think we should expect it to take this form. This is because a declarative sentence has more *authority* than a mere expressive one. If I want someone to do something it is arguably more effective to disguise it in non-subjective form, to give it the appeal of a truth-aptness.[7] I want to disguise it so that it will serve this persuasive purpose.

I don't want to say that these ideas are correct. But they're plausible theories that predict the surface form of moral language, and which are no worse than cognitivistic theories. Expressivism focuses on what people do with language. It focuses on the speech act, not the literal meaning. Whether people express, declare, prescribe, describe, recommend, or evaluate is nothing we can easily read from the surface form. But this is what Prinz seems to presuppose when he says the most obvious empirical prediction fails. We have to look at their behavior and the pragmatic context in which the discourse happens.

I argue that this fits even better with Prinz's project of a naturalized metaethics. When Prinz discusses the last step in the decision tree, he writes: "Naturalizing relativism will require the marriage of cultural anthropology and sociolinguistics" (this collection, p. 24). I think this marriage could be more helpful at an earlier stage in the decision tree—to help answer the question of whether or not moral terms aim at truth.

4 Conclusion

In this commentary on Prinz's highly interesting and substantial target paper I welcomed his methodological naturalism, but argued that his project is not as conciliatory between traditional analytical philosophy and naturalized philosophy as he seems to think. The reason is that on closer scrutiny we find opposing views on the methodology and purpose of philosophy. In the second part of my contribution I looked at an objection against Prinz's sentimentalism. I argued, first, that he misses the real core of this kind of objections. Then I used Jon Tresan's distinction between *de dicto*- and *de re*-internalism as a conceptual tool to propose and develop another answer that Prinz could use against this objection. In particular, I claimed that, given Prinz's metaethical naturalism, we should not look for conceptual necessity but for fruitful hypotheses which we can test in *a posteriori*. In the third and last part I argued against Prinz's critique of non-cognitivism. Prinz thinks that the most obvious empirical prediction of non-cognitivism fails. Here, I tried to demonstrate how non-cognitivism, given a pragmatical view of moral language, actually predicts the surface grammar of moral discourse as well as cognitivistic alternatives. I proposed a speculative explanation for this interesting fact. This kind of explanation, I believe, fits even better with Prinz's project of a naturalized metaethics.

6 Stevenson (1937, p. 25) writes: "I do like this; do so as well!" But the first part looks suspiciously descriptive. Because this doesn't fit with Stevenson's account, I reformulated it in this way.

7 Mackie discusses this instrumental use when he discusses why people give their moral judgments the appeal of objectivity (1990, p. 42). But as we saw, Prinz thinks this premise is wrong.

References

Bedke, M. S. (2009). Moral judgment purposivism: Saving internalism from amoralism. *Philosophical Studies*, *144* (2), 189-209. 10.1007/s11098-008-9205-5

Björklund, F., Björnsson, G., Eriksson, J., Francén Olinder, R. & Strandberg, C. (2012). Recent work on motivational internalism. *Analysis*, *72* (1), 124-137. 10.1093/analys/anr118

Björnsson, G. & Francén Olinder, R. (2013). "Internalists beware" - We might all be amoralists! *Australasian Journal of Philosophy*, *91* (1), 1-14. 10.1080/00048402.2012.665373

Björnsson, G. (2002). How emotivism survives immoralists, irrationality, and depression. *Southern Journal of Philosophy*, *40* (3), 327-344. 10.1111/j.2041-6962.2002.tb01905.x

Björnsson, G., Eriksson, J., Strandberg, C., Francén Olinder, R. & Björklund, F. (2014). Motivational internalism and folk intuitions. *Philosophical Psychology*, 1-20. 10.1080/09515089.2014.894431

Darwall, S. L. (1995). *The british moralists and the internal 'ought', 1640-1740.* Cambridge, UK: Cambridge University Press.

DeLapp, K. M. (2011). Metaethics. *The Internet Encyclopedia of Philosophy.* http://www.iep.utm.edu/metaethi/

Francén, R. (2010). Moral motivation pluralism. *Journal of Ethics*, *14* (2), 117-148. 10.1007/s10892-010-9074-y

Kauppinen, A. (2008). Moral internalism and the brain. *Social Theory and Practice*, *34* (1), 1-24. 10.5840/soctheorpract20083411

Mackie, J. L. (1990). *Ethics : Inventing right and wrong.* London, UK: Penguin.

Nichols, S. (2002). How psychopaths threaten moral rationalism: Is it irrational to be amoral. *The Monist*, *85* (2), 285-303.

Prinz, J. (2006). The emotional basis of moral judgments. *Philosophical Explorations*, *9* (1), 29-43. 10.1080/13869790500492466

——— (2008). Empirical philosophy and experimental philosophy. In J. Knobe & S. Nichols (Eds.) *Experimental philosophy* (pp. 189-208). Oxford, UK: Oxford University Press.

——— (2015). Naturalizing metaethics. In T. Metzinger & J. M. Windt (Eds.) *Open MIND.* Frankfurt a. M., GER: MIND Group.

Roskies, A. (2003). Are ethical judgments intrinsically motivational? Lessons from "Acquired Sociopathy". *Philosophical Psychology*, *16* (1), 51-66. 10.1080/0951508032000067743

——— (2008). Internalism and the evidence from pathology. In W. Sinnott-Armstrong (Ed.) *Moral psychology: The neuroscience of morality: emotion, brain disorders, and development* (pp. 191-206). Cambridge, MA: MIT Press.

Schulte, P. (2012). Satan Und Der Masochist: Eine Nonkognitivistische Antwort auf den Amoralismus-Einwand. In A. Dunshirn, E. Nemeth & G. Unterthurner (Eds.) *Crossing Borders. Grenzen (über)Denken. Beiträge Zum 9. Internationalen Kongress der österreichischen Gesellschaft für Philosophie in Wien* (pp. 599-608). Wien, AUT: Österreichische Gesellschaft Für Philosophie.

Shafer-Landau, R. (2003). *Moral realism: A defence.* Oxford, UK: Oxford University Press.

Smyth, N. (2014). Resolute expressivism. *Ethical Theory and Moral Practice*, *17* (4), 607-618. 10.1007/s10677-014-9495-y

Stevenson, C. L. (1937). The emotive meaning of ethical terms. *Mind, New Series*, *46* (181), 14-31.

Strandberg, C. & Björklund, F. (2013). Is moral internalism supported by folk intuitions? *Philosophical Psychology*, *26* (3), 319-335. 10.1080/09515089.2012.667622

Tresan, J. (2009). Metaethical internalism: Another neglected distinction. *Journal of Ethics*, *13* (1), 51-72.

van Roojen, M. (2013). Moral cognitivism vs. non-cognitivism. *The Stanford Encyclopedia of Philosophy*, *Winter 2013* E. N. Zalta (Ed.) http://plato.stanford.edu/archives/win2013/entries/moral-cognitivism/

Should Metaethical Naturalists Abandon *de dicto* Internalism and Cognitivism?

A Reply to Yann Wilhelm

Jesse Prinz

Yann Wilhelm pursues three issues in response to my target article. First, he tries to expose my naturalism as more radical than I let on. I concede the point, though I also offer ways in which my radicalism might be mitigated. Second, he exposes a limitation in my argument for internalism, and suggests that naturalists should defend form on internalism that is neutral about conceptual claims (*de re* internalism, rather than *de dicto*). I welcome the suggestion, but also consider how naturalists might defend *de dicto* internalism. Third, Wilhelm challenges my argument against non-cognitivism, by offering a novel explanation of the fact that moral judgments have an assertoric form. I response, I note avenues for cognitivist resistance to Wilhelm's explanation.

Keywords
Cognitivism | Conceptual truth | Internalism | Metaethics | Naturalism | Non-cognitivism

Author

Jesse Prinz
jesse@subcortex.com
City University of New York
New York, NY, U.S.A.

Commentator

Yann Wilhelm
ywilhelm@students.uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

In "Naturalizing Metaethics," I try to establish that core questions in metaethics lend themselves to empirical investigation. I argue that we can potentially adjudicate long-standing debates by testing predictions made by competing metaethical theories. I also make some conjectures about how such empirical investigations will turn out. Based on a small selection of preliminary findings, I advance a case of a version of sentimentalism—the view that emotions are essential to moral judgments. I also suggest that sentimentalism commits me to internalism—the view that moral judgments are essentially motivating—and I advance an empirical case for cognitvism—the claim that moral judgments are, like other assertions, capable of being true or false.

In his insightful commentary, Yann Wilhelm offers clarifications and challenges to my

arguments. First, he asks whether my naturalism is compatible with traditional approaches in philosophy. I imply that the two can co-exist in a complementary way, but Wilhelm suggests that my naturalism is more radical that it appears. I am forced to agree, and to clarify the co-existence claim. Wilhelm also challenges my case for internalism, distinguishing two different forms and suggesting I am only in a position to argue for one of them. I am open to that possibility, but I also sketch a strategy for defending both forms. Wilhelm concludes with a challenge to my defense of cognitivism. He provides non-cognitivists with an explanation for findings that I say they cannot explain. I offer a cognitivist response, but grant that this proposal demands empirical attention.

Wilhelm's commentary provides a valuable contribution to empirically oriented metaethics. He offers strategies for avoiding certain kinds of debates with opponents of naturalism, and he identities empirical issues that can be used settle debates between card-carrying naturalists. Wilhelm deepens my understanding of these issues and strengthens my optimism about the prospects of naturalistic metaethics.

2 Is naturalism a radical position?

Before moving on to the first order debates that Wilhelm so helpfully pursues, I want to concede an important point that he makes in the opening of his commentary. Wilhelm rightly observes that I overstate the extent to which a thoroughgoing naturalism can preserve traditional approaches to philosophy. Though ostensibly a plea for consilience, I am, in fact, skeptical about the notion *a prioricity*. Rather, I claim that armchair methods are observational (intuitions are defeasible inner observations informed by prior experience, and open to empirical correction). As Wilhelm makes clear, traditionalists who view conceptual analysis as an *a priori* endeavor will not share my enthusiasm for naturalism.

In another respect, however, my position is conservative. I don't think traditional philosophers must stop working as they currently do. Armchair methods remain the primary source of philosophical theories and distinctions. They also are the primary source for philosophical thought experiments that can be used to test between theories. Thus, my invitation to interpret armchair methods as observation is intended as a vindication of traditional philosophy, though not a vindication of how some traditional philosophers understand their own endeavors.

Proof of this qualified vindication comes form the fact that empirically oriented philosophers regularly draw on traditional work in devising their studies. For example, experimental philosophers have used trolley problems, twin earth cases, and the thought experiments used to back contextualism in epistemology. In my target paper, I relied on theories that have been identified and articulated within traditional philosophy. Testing between theories requires observation, I believe, but it would be a great loss if every philosopher ran a laboratory. Instead, I envision a future for philosophy in which many researchers do no experimental work, others are primarily experimentalists, and still others do a combination of the two. If we begin to make empirical methods a standard part of philosophical training, then philosophers will be able to read psychological research more responsibly and conduct experiments when they see fit. But it doesn't mean that they will also suddenly stop thinking and blindly collect data. As in the sciences, theoretical work in required in philosophy. We can resist the idea of *a priori* truth without throwing away the armchair.

3 Must naturalist be content with *de re* internalism?

These methodological points bear on Wilhelm's first challenge to my metaethical conclusions. In the target paper I argue for a form of internalism (roughly, the view that moral judgments are essentially motivating). Wilhelm points out that my evidence for this claim will not satisfy many externalists. I primarily rely on evidence that moral judgments always co-occur with emotional states, but fee externalists will be impressed; they will say that such findings cannot address questions about whether it is necessar-

ily the case that moral judgments are motivating, even if they always happen to be are motivating.

Wilhelm helpfully replies to this objection on my behalf, using Jon Tresan's distinction between *de dicto* and *de re* internalism. The former is a thesis about the concept of moral judgment (viz., it is a conceptual truth that when that concept applies, motivation applies as well). The latter is a claim about moral judgments themselves (viz., moral judgments do in fact carry motivation force). Wilhelm concurs that my evidence can contribute to a defense of the *de re* claim. He suggests that I abandon the case for *de dicto* internalism, since naturalists should not concern themselves with conceptual claims.

I welcome Wilhelm's suggestion, and I am inclined to endorse it. Let me mention, however, a strategy available to the naturalist whose heart is set on defending the *de dicto* claim. Returning to Wilhelm's discussion of methodology, let's imagine that naturalists wage a successful campaign against the *a priori*. Properly pursued, such a campaign might also undermine metaphysical necessity. Metaphysical necessities, unlike nomological necessities, are alleged to be true in virtue of conceptual entailments rather than laws of nature or natural facts. The critique of *a prioricity* threatens metaphysical necessity because it advances the view that truths about concepts are open to empirical revision. Let's suppose that concepts are mental representations garnered through experience with the function of classifying things in the world. So construed, concepts are susceptible to improvement through empirical inquiry. Initial concepts are rough and ready pointers that we use to carve up the observational world, and revised concepts are carvings that remain after observation. Now let us define a "robust conceptual truth" as the conceptual entailments that survive after a concept has been subjected to empirical fine-tuning. Such truths would more or less coincide with how the world is, together with certain pragmatic assumptions that go into theory construction. Thus, they would coincide with truths that emerge from our study of the things themselves (which are also constrained by pragmatic assumptions). On this picture, *de dicto* collapses into *de re*. A defense of *de re* internalism would indicate that our concept of moral judgment will converge on internalism as well. Rather than bypassing *de re* internalism, we can try to defend it by naturalizing conceptual truth.

Wilhelm might reply that this defense of *de dicto* internalism would not persuade non-naturalists. The defense is based on the assumption that the naturalist critique *of a prioricity* goes through, but that is just what non-naturalists are inclined to deny. Thus, it might appear the that the debate over the *de dicto* position is hostage to unresolvable disputes about the nature of philosophy.

Here I'd balk at the claim that such disputes are unresolvable. Those who believe is *a prioricity* may dislike naturalism, but they certainly believe that their views require evidential support. Naturalists offer an account of what concepts are (mental representations) and an explanation of conceptual intuitions (introspection of mental representations). Non-naturalists are obliged to provide an alternative account of both, and the two accounts can then be compared by agreed upon standards. I venture that the naturalist account will find a resounding victory in such a head-to-head match. It is more parsimonious view, since both sides must grant the existence of mental representations, and I suspect it can fully account for our conceptual intuitions.

These are, of course, big debates, which I cannot settle here. My point is simply that we can imagine a two-stage process that begins with broad issues about naturalism, and then moves on to first-order views. On my prognosis, we won't end up abandoning the notion of conceptual truth, but rather revising it. If so, *de dicto* naturalism might turn out true. Wilhelm may be right, however, that until we come to greater consensus on the nature of philosophy, naturalists might be on firmer ground if they try to bypass conceptual questions. He is also right that, from a naturalist perspective *de re* internalism may be the more interesting thesis. Conceptual claims lose their distinctive interest if concepts are revisable and, ultimately, coincident with empirical theories.

4 Can non-cognitivists explain the assertoric form of moral judgments?

Let me turn, finally, to Wilhelm's constructive effort to defend non-cognitivism. Non-cognitivists claim that moral judgments are not like ordinary assertions; they cannot be assessed as true or false, but rather merely express the speakers attitudes and commendations. If so, I ask, why do we express moral judgments as assertions? This is a familiar challenge. In my discussion, I merely I point out that can be backed up by empirical data. Wilhelm has a two-part reply. First, he observes that, for non-cognitivists, the primary function of moral discourse is to persuade. C. L. Stevenson, for example, says that "x is bad" does not just mean "boo to x!"; it also means and "say boo to x as well!". Second, Wilhelm makes the original and plausible suggestion that this persuasive function is most effective when it covert. People, he notes, don't like to be manipulated. If I explicitly exhort you to say "boo!" you may resist, because no one likes being told what to do. But if I present my attitude in the form on an assertion, you might causally take it on board, as you would if I were presenting an ordinary statement of fact.

I think Wilhelm's proposal deserves serious exploration. Cognitivists can respond in two ways. First, they can try to show that moral discourse often occurs in contexts that don't aim at persuasion. This might seem implausible. After all, why should we bother engaging in moral discourse if we don't intend to persuade anyone? On closer analysis, however, it does seem that much of our moral discourse involves preaching to the choir. In political debates, for example, left wing pundits and right wing pundits engage in a lot of moral discourse, but they never seem to persuade each other. This raises the intriguing possibility that moral judgments are not primarily in the business of persuasion. An alternative possibility is that we make moral judgments to assert our identity, or express solidarity with like-minded individuals. Empirical tests might be designed to compare the persuasion model and the self-expression model.

Cognitivists might also try to resist Wilhelm's conjecture that people do not like to be manipulated by consulting research on explicit persuasion. In defense of Wilhelm's conjecture, there is a literature suggesting that people sometimes resist explicit persuasion (e.g., Petty & Cacioppo 1979). On the other hand, resistance does not occur in all contexts. Indeed, in a consumer product context, Reinhard et al. (2006) found that, when a person is regarded as likeable (or attractive!), they become even more persuasive when they make their intent to persuade explicit. Similarly, in studies of college drinking behavior, Neighbors et al. (2008) found that injunctive norms (which explicitly reference attitudes) are effective when and only when they are expressed by members of the students' social groups. Further work could test the effects of explicit injunctions in the moral domain.

I should underscore that I think more testing is required to settle these debates. Wilhem's explanation for surface discourse remains viable, and we can make progress on these issues by devising new ways to test it. These are manifestly empirical issues. While I wager with the cognitivists, I grant that the case is far from closed.

5 Conclusion

I am indebted to Yann Wilhelm for his generous and probing commentary. It brings welcome clarification and new challenges to the project I set out "Naturalizing Metaethics." I also welcome the spirit of Wilhelm's discussion, which moves beyond ideological debates about metaphilosophy, and offers promising strategies for answering core metaethical questions.

Wilhelm successfully establishes that my preferred form of naturalism is less compatible with traditional philosophy than I let on, but I also pointed out that work by traditionally minded philosophy remains an invaluable font of philosophical theories. Wilhelm then offers a helpful suggestion that naturalists might more easily defend internalism if they bypass conceptual versions of that view. In response, I suggested that the radical implications of naturalism may actually offer a way to defend the concep-

tual version of internalism, by advancing a naturalized account of conceptual truth. Finally, Wilhelm offered a new psychological cum functional account of moral discourse, which inoculates non-cognitivists against grammatical objections. While I hold out hope for cognitivism, Wilhelm has identified a genuine empirical challenge to the cognitivist. This challenge beautifully demonstrates the value of empirical testing in metaethics, and it also reminds us that there is much work to be done.

References

Neighbors, C., O'Connor, R. M., Lewis, M. A., Chawla, N., Lee, C. M. & Fossos, N. (2008). The relative impact of injunctive norms on college student drinking: The role of reference group. *Psychology of Addictive Behaviors: Journal of the Society of Psychologists in Addictive Behaviors*, *22* (4), 576-581. 10.1037/a0013043

Petty, R. E. & Cacioppo, J. T. (1979). Effects of forewarning of persuasive intent and involvement on cognitive responses and persuasion. *Personality and Social Psychology Bulletin*, *5* (2), 173-176. 10.1177/ 014616727900500209

Reinhard, M.-A., Messner, M. & Sporer, S. (2006). Explicit persuasive intent and its impact on success at persuasion – The determining roles of attractiveness and likeableness. *Journal of Consumer Psychology*, *16*, 249-259. 10.1207/s15327663jcp1603_7