Experimental Moral Philosophy

First published Wed Mar 19, 2014; substantive revision Thu Aug 9, 2018

Experimental moral philosophy emerged as a methodology in the last decade of the twentieth century, as a branch of the larger experimental philosophy (X-Phi) approach. Experimental moral philosophy is the empirical study of moral intuitions, judgments, and behaviors. Like other forms of experimental philosophy, it involves gathering data using experimental methods and using these data to substantiate, undermine, or revise philosophical theories. In this case, the theories in question concern the nature of moral reasoning and judgment; the extent and sources of moral obligations; the nature of a good person and a good life; even the scope and nature of moral theory itself. This entry begins with a brief look at the historical uses of empirical data in moral theory and goes on to ask what, if anything, is distinctive about experimental moral philosophy—how should we distinguish it from related work in empirical moral psychology? After discussing some strategies for answering this question, the entry examines two of the main projects within experimental moral philosophy, and then discusses some of the most prominent areas of research within the field. As we will see, in some cases experimental moral philosophy has opened up new avenues of investigation, while in other cases it has influenced longstanding debates within moral theory.

1. Introduction and History
   1.1 What are Experiments?
   1.2 What Types of Questions and Data?
2. Moral Intuitions and Conceptual Analysis
   2.1 The Negative Program in Experimental Moral Philosophy
   2.2 The Positive Program in Experimental Moral Philosophy
   2.3 An Example: Intentionality and Responsibility
   2.4 Another Example: The Linguistic Analogy
3. Character, Wellbeing, and Emotion
   3.1 Character and Virtue
   3.2 Wellbeing, Happiness, and the Good Life
   3.3 Emotion and Affect
4. Metaethics and Experimental Moral Philosophy
   4.1 Folk Metaethics and Moral Realism
   4.2 Moral Disagreement
   4.3 Moral Language
5. Criticisms of Experimental Moral Philosophy
   5.1 Problems with Experimental Design and Interpretation
   5.2 Philosophical Problems

Bibliography

Academic Tools

Other Internet Resources

Related Entries

1. Introduction and History

The idea that our actual moral judgments are an important source of information about the origins and justification of
Astronomers, for example, sometimes speak of natural experiments. Although we are in no position to affect the outcomes of astronomical events, we can observe and analyze them to gain insights into the workings of the universe. While true experiments follow this procedure, other types of studies allow non-random assignment to conditions. For example, finding or not finding the dime is the condition; degree of helpfulness is the outcome variable. If the experimenter simply throws the dime in the air and asks for help in catching it, the outcome will depend on the person's ability to catch a dime, which is not necessarily random. The variation in conditions is sometimes called a manipulation. For example, an experimenter might try to vary the degree of helpfulness by giving different amounts of cash as a reward for catching the dime.

Paradigmatic experiments involve randomized assignment to varying conditions of objects or people from a representative sample, followed by statistical comparison of the outcomes for each condition. The variation in conditions is sometimes called a manipulation. For example, an experimenter might try to vary the degree of helpfulness by giving different amounts of cash as a reward for catching the dime. The variation in conditions is sometimes called a manipulation. For example, an experimenter might try to vary the degree of helpfulness by giving different amounts of cash as a reward for catching the dime.

These examples show that ethnography has a valuable role to play in philosophical theory, but the remainder of this entry will focus primarily on experiments. Paradigmatic experiments involve randomized assignment to varying conditions of objects or people from a representative sample, followed by statistical comparison of the outcomes for each condition. The variation in conditions is sometimes called a manipulation. For example, an experimenter might try to select a representative sample of people, randomly assign them either to find or not to find a dime in the coin return of a pay phone, and then measure the degree to which they subsequently help someone they take to be in need (Ison and Levin 1972). Finding or not finding the dime is the condition; degree of helpfulness is the outcome variable.

While true experiments follow this procedure, other types of studies allow non-random assignment to conditions. For example, sometimes speak of natural experiments. Although we are in no position to affect the outcomes of astronomical events, we can observe and analyze them to gain insights into the workings of the universe. While true experiments follow this procedure, other types of studies allow non-random assignment to conditions. For example, sometimes speak of natural experiments. Although we are in no position to affect the outcomes of astronomical events, we can observe and analyze them to gain insights into the workings of the universe.
motions of heavenly bodies, we can observe them under a variety of conditions, over long periods of time, and with a variety of instruments. Likewise, anthropological investigation of different cultures’ moral beliefs and practices is unlikely to involve manipulating variables in the lives of the members of the culture, but such studies are valuable and empirically respectable. Still, such studies can at best demonstrate correlation, not causation, so their evidential value of studies is less than that of experiments; most published research in experimental philosophy involves true experiments.

Even within the category of experiments, we find a lot of diversity with respect to inputs, methods of measurement, and outputs. The experiment just described uses a behavioral measure and manipulation—finding the dime is the input, helping is the outcome measured. Other experiments measure not behavior but judgment or intuition, and this can be done using a survey or other form of self-report or informant-report where subjects respond explicitly to some question, situation, or dilemma. Studies measuring judgments might use either manipulations of the condition in which the subject makes the judgment, or they might look for correlations between judgments and some other factor, such as brain activity, emotional response, reaction time, visual attention, and so on (Strohminger et al. 2014).

The experimental methods available have also changed over time. Surveys have been the dominant method of experimental philosophy for the past few decades, but technology may change this: advances in virtual and augmented reality mean that philosophers can now immerse people in moral dilemmas such as Thompson’s (1971) violinist thought experiment and different versions of the trolley problem. Philosophers interested in the neural correlates of moral judgment can use transcranial magnetic stimulation (TMS) to investigate the effects of enhancing or lessening activity in certain areas of the brain. Even survey methods have seen advances thanks to technology; the ubiquity of smartphones allows researchers to ping people in real time, asking for reports on mood (see section 3.2 below for a discussion of surveys relating to happiness and mood).

Whether experimental moral philosophy has to use true experiments or can include studies and even ethnographies and other forms of qualitative data is partly a terminological question about how to define the field and whether we distinguish it from empirical moral psychology, a closely related research program. As we will see below, though, a diversity of both methods and subjects is important in helping experimental moral philosophy respond to its critics.

### 1.2 What Types of Questions and Data?

Like experimental philosophy more generally, experimental moral philosophy is interested in our intuitions and judgments about philosophical thought experiments and moral dilemmas. But research in this area also concerns the cognitive architecture underlying our moral judgments; the developmental processes that give rise to moral judgments; the neuroscience of moral judgment; and other related fields.

**Direct** experiments investigate whether a claim held (or denied) by philosophers is corroborated or falsified. This might mean investigating an intuition and whether it is as widely shared as philosophers claim, or it might mean investigating the claim that a certain behavior or trait is widespread or that two factors covary. For example, we find philosophers claiming that it is wrong to imprison an innocent person to prevent rioting; that a good life must be authentic; and that the claim that a certain behavior or trait is widespread or that two factors covary. For example, we find philosophers claiming that it is wrong to imprison an innocent person to prevent rioting; that a good life must be authentic; and that moral judgments are intrinsically motivating. Experimental research can be (and has been, as we will see below) brought to bear on each of these claims.

**Indirect** experiments look at the nature of some capacity or judgment: for example, whether certain types of moral dilemmas engage particular areas of the brain; how early children develop a capacity to empathize; and whether the moral/conventional distinction is universal across cultures. These claims have philosophical relevance insofar as they help us understand the nature of moral judgment.

In addition to distinctions involving the type of data and how directly it bears on the question, we can also distinguish among experimental applied ethics, experimental normative ethics, and experimental metaethics. The first involves the variables that influence judgments about particular practical problems, such as how a self-driving car should respond to sacrificial dilemmas (Bonnefon, Sharriff, and Rahwan 2016). The second involves investigations of how we ought to behave, act, and judge, as well as our intuitions about moral responsibility, character, and what constitutes a good life. The third involves debates over moral realism and antirealism. In this entry we focus on the latter two.

In many of these cases, the line between experimental moral philosophy and its neighbors is difficult to draw: what distinguishes experimental moral philosophy from empirical moral psychology? What is the difference between experimental moral philosophy and psychology or neuroscience that investigates morality? What is the difference between experimental moral philosophy and metaethics? We might try to answer these questions by pointing to the training of the experimenters—as philosophers or as scientists—but much of the work in these areas is done by collaborative teams involving both philosophers and social scientists or neuroscientists. In addition, some philosophers work in law and psychology departments, and graduate programs increasingly offer cross-disciplinary training. Another approach would be to look at the literature with which the work engages: many psychologists (e.g., Haidt, Greene) investigating moral judgment situate their arguments within the debate between Kantians and Humeans, so engagement with the philosophical tradition might be used as a criterion to distinguish experimental moral philosophy from experimental moral psychology. All this said, an inability to sharply distinguish experimental moral philosophy from adjoining areas of investigation may not be particularly important. Experimental philosophers often point out that the disciplinary divide between philosophy and psychology is a relatively recent phenomenon; early twentieth-century writers such as William James situated themselves in both disciplines, making vital contributions to each. If there is a problem here, it is not unique to experimental moral philosophy. For example, is work on semantics that uses both
2. Moral Intuitions and Conceptual Analysis

One role for experiments in moral philosophy, as indicated above, is to investigate our moral intuitions about cases, and to use the results gleaned from such investigations to guide or constrain our moral metaphysics, semantics, or epistemology. Philosophers often rely on such judgments—in the form of claims to the effect that “we would judge x” or, “intuitively, x”—as data or evidence for a theory (though see Cappelen 2012 and Deutsch 2016 for critical discussion of this point). Claims about what we would judge or about the intuitive response to a case are empirically testable, and one project in experimental moral philosophy (perhaps the dominant original project) has been to test such claims via an investigation of our intuitions and related moral judgments. In doing so, experimental moral philosophers can accomplish one of two things: first, they can test the claims of traditional philosophers about what is or is not intuitive; second, they can investigate the sources of our intuitions and judgments. These tasks can be undertaken as part of a positive program, which uses our intuitions and judgments as inputs and constructs theories that accommodate and explain them. Alternatively, these tasks can figure in a negative program, which uses experimental research to undermine traditional appeals to intuition as evidence in moral philosophy and conceptual analysis more broadly. The negative program can proceed directly or indirectly: either via testing and refuting claims about intuitions themselves, or by discrediting the sources of those intuitions by discovering that they are influenced by factors widely regarded as not evidential or unreliable. We discuss both the negative program and the positive program below.

2.1 The Negative Program in Experimental Moral Philosophy

Early work in experimental philosophy suggested cross-cultural differences in semantic, epistemic, and moral intuitions. For example, Machery, Mallon, and Stich (2004) argued that East Asian subjects were more likely to hold deontivist intuitions than their Western counterparts, who tended to embrace causal theories of reference. Haidt (2006) argued that the extent to which people judged harmless violations such as eating a deceased pet, or engaging in consensual sibling incest, to be wrong depended on socioeconomic status. Since, presumably, moral wrongness doesn’t itself depend on socioeconomic status, gender, or culture of the person making the moral judgment, these results have been marshaled to argue either against the evidential value of intuitions or against the existence of moral facts altogether.

Negative experimental moral philosophy generates results that are then used to discount the evidential value of appeals to intuition. For example, Walter Sinnott-Armstrong (2008d), Eric Schwitzgebel, Fiery Cushman (2012), and Peter Singer (2005) have recently followed this train of thought, arguing that moral intuitions are subject to normatively irrelevant situational influences (e.g., order effects), while Feltz & Cokely (2009) and Knobe (2011) have documented correlations between moral intuitions and (presumably) normatively irrelevant individual differences (e.g., extroversion). Such results might warrant skepticism about moral intuitions, or at least about some classes of intuitions or intuiters.[1] The studies just mentioned involve results that suggest that the content of intuitions varies along some normatively irrelevant dimension. Another source of evidence for the negative program draws on results regarding the cognitive mechanisms underlying or generating intuitions themselves. For example, studies that suggest that deontological intuitions are driven by emotional aversions to harming others have been used to argue that we ought to discount our deontological intuitions in favor of consequentialist principles—an idea discussed in more detail below (see Singer 2005, Weigman 2017).

The distinction between negative and positive experimental moral philosophy is difficult to draw, partly because the negative program often discounts particular classes or types of intuitions in favor of others that are supposed to be more reliable. For example, Singer offers an anti-deontological argument as part of the negative program insofar as his argument uses the emotional origins of deontological intuitions to discount them. But because he then argues for the superiority of consequentialist intuitions, his position also fits within the positive program. So the difference between the two programs is not in the data or the kinds of questions investigated, but in how the data are put to use—whether they are seen as debunking traditional philosophical appeals to intuition or as an addition to traditional philosophical appeals to intuition, by helping philosophers distinguish between reliable and unreliable intuitions and their sources. In the next section, we look at positive uses of intuitions as evidence.

2.2 The Positive Program in Experimental Moral Philosophy

Other philosophers are more sanguine about the upshot of experimental investigations of moral judgment and intuition. Knobe, for example, uses experimental investigations of the determinants of moral judgments to identify the contours of philosophically resonant concepts and the mechanisms or processes that underlie moral judgment. He has argued for the pervasive influence of moral considerations throughout folk psychological concepts (2009, 2010; see also Pettit &
represents the mind of the participant, which is not directly observable but is the target of investigation. In broad

In this model, the boxes represent constructs, the arrows represent causal or functional influences, and the area in grey

environment) than when it was perceived to be good (helping the environment). This effect has been replicated dozens

Participants are asked to attribute intentionality or an attitude (or, in the case of later work by Robinson et al. 2013, a

2.3 An Example: Intentionality and Responsibility

Since Knobe’s seminal (2003) paper, experimental philosophers have investigated the complex patterns in people’s
dispositions to make judgments about moral notions (praiseworthiness, blameworthiness, responsibility), cognitive
attitudes (belief, knowledge, remembering), motivational attitudes (desire, favor, advocacy), and character traits
(compassion, callousness) in the context of violations of and conformity to various norms (moral, prudential, aesthetic,
legal, conventional, descriptive). In Knobe’s original experiment, participants first read a description of a choice
scenario: the protagonist is presented with a potential policy (aimed at increasing profits) that would result in a side
effect (either harming or helping the environment). Next, the protagonist explicitly disavows caring about the side effect,
choices to go ahead with the policy. The policy results as advertised: both the primary and the side effect occur.
Participants are asked to attribute intentionality or an attitude (or, in the case of later work by Robinson et al. 2013, a
character trait) to the protagonist. What Knobe found was that participants were significantly more inclined to indicate
that the protagonist had intentionally brought about the side effect when it was perceived to be bad (harming the
environment) than when it was perceived to be good (helping the environment). This effect has been replicated dozens
of times, and its scope has been greatly expanded from intentionality attributions after violations of a moral norm to
attributions of diverse properties after violations of a wide variety of norms.

The first-order aim of interpreters of this body of evidence is to create a model that predicts when the attribution
asymmetry will crop up. The second-order aims are to explain as systematically as possible why the effect occurs, and
to determine the extent to which the attribution asymmetry can be considered rational. Figure 1, presented here for the
first time, models how participants’ responses to this sort of vignette are produced:

In this model, the boxes represent constructs, the arrows represent causal or functional influences, and the area in grey
represents the mind of the participant, which is not directly observable but is the target of investigation. In broad
Another Example: The Linguistic Analogy

than other experimental philosophers suppose, and is due to rational processes. and leads to the attribution asymmetry. On the deliberation model, then, the attribution asymmetry originates earlier.

Participants tend to think of the protagonist as deliberating about what to do when and only when a norm would be violated. Since deliberation leads to the formation of other mental states such as beliefs, desires, and intentions, this basal difference between participants' models of what happens in the story flows through the rest of their interpretation.

Participants are told that a policy he would ordinarily want to pursue violates a norm, he acquires a reason to deliberate further about what to do. However, when it comes time to tell the experimenter what they think, participants do not want to be taken as suggesting that the harm-causing protagonist is blameless, so they report that he acted intentionally. This is a reasonable goal, so according to the pragmatic revision stage, the asymmetry in attributions is a rational expression of the ordinary conception of intentionality (desire, belief, etc.), which turns out to have a normative component.

The Conceptual Competence Model

Perhaps the best known is Knobe's conceptual competence model, according to which the asymmetry arises at the judgment stage. On this view, normative judgments about the action influence otherwise descriptive judgments about whether it was intentional (or desired, or expected, etc.). Moreover, this influence is taken to be part of the very conception of intentionality (desire, belief, etc.). Thus, on the conceptual competence model, the asymmetry in attributions is a rational expression of the ordinary conception of intentionality (desire, belief, etc.), which turns out to have a normative component.

The Motivational Bias Model

The motivational bias model (Alicke 2008; Nadelhoffer 2004, 2006) agrees that the asymmetry originates in the judgment stage, and that normative judgments influence descriptive judgments. However, unlike the conceptual competence model, it takes this to be a bias rather than an expression of conceptual competence. Thus, on this model, the asymmetry in attributions is a distortion of the correct conception of intentionality (desire, belief, etc.).

The Deep Self Model

The deep self concordance model (Sripada 2010, 2012; Sripada & Konrath 2011) also locates the source of the asymmetry in the judgment stage, but does not recognize an influence (licit or illicit) of normative judgments on descriptive judgments. Instead, proponents of this model claim that when assessing intentional action, people not only attend to their representation of a person's "surface" self—her expectations, means-end beliefs, moment-to-moment intentions, and conditional desires—but also to their representation of the person's "deep" self, which harbors her sentiments, values, and core principles. According to this model, when assessing whether someone intentionally brings about their representation of the outcome the agent brings about and what they take to be her deep self. For instance, when the chairman says he does not care at all about either harming or helping the environment, people attribute to him a deeply anti-environment stance. When he harms the environment, this is concordant with his anti-environment deep self; in contrast, when the chairman helps the environment, this is discordant with his anti-environment deep self. According to the deep self concordance model, then, the asymmetry in attributions is a reasonable expression of the folk psychological distinction between the deep and shallow self.

The Conversational Pragmatics Model

Unlike the models discussed so far, the conversational pragmatics model (Adams & Steadman 2004, 2007) locates the source of the asymmetry in the pragmatic revision stage. According to this model, participants judge the protagonist not to have acted intentionally in both norm-conforming and norm-violating cases. However, when it comes time to tell the experimenter what they think, participants do not want to be taken as suggesting that the harm-causing protagonist is blameless, so they report that he acted intentionally. This is a reasonable goal, so according to the pragmatic revision model, the attribution asymmetry is rational, though misleading.

The Deliberation Model

According to the deliberation model (Alfano, Beebe, & Robinson 2012; Robinson, Stey, & Alfano 2013; Scaife & Webber 2013), the best explanation of the complex patterns of evidence is that the very first mental stage, the formation of a mental model of the scenario, differs between norm-violation and norm-conformity vignettes. When the protagonist is told that a policy he would ordinarily want to pursue violates a norm, he acquires a reason to deliberate further about what to do; in contrast, when the protagonist is told that the policy conforms to some norm, he acquires no such reason. Participants tend to think of the protagonist as deliberating about what to do when and only when a norm would be violated. Since deliberation leads to the formation of other mental states such as beliefs, desires, and intentions, this asymmetry in attributions is a rational expression of the ordinary conception of intentionality (desire, belief, etc.), which turns out to have a normative component.
Another positive program investigates the structure and form of moral intuitions, in addition to their content, with the aim of using these features to inform a theory of the cognitive structures underlying moral judgment and their etiology. Rawls (1971), drawing on Chomsky’s (1965) theory of generative linguistics, suggested that moral cognition might be usefully modeled on our language faculty, a parallel endorsed by Chomsky himself:

I don’t doubt that we have a natural moral sense…. That is, just as people somehow can construct an extraordinarily rich system of knowledge of language on the basis of rather limited and degenerate experience, similarly, people develop implicit systems of moral evaluation, which are more or less uniform from person to person. There are differences, and the differences are interesting, but over quite a substantial range we tend to make comparable judgments, and we do it, it would appear, in quite intricate and delicate ways involving new cases and agreement often about new cases… and we do this on the basis of a very limited environmental context available to us. The child or the adult doesn’t have much information that enables the mature person to construct a moral system that will in fact apply to a rich range of cases, and yet that happens…. whenever we see a very rich, intricate system developing in a more or less uniform way on the basis of rather restricted stimulus conditions, we have to assume that there is a very powerful, very rich, highly structured innate component that is operating in such a way as to create that highly specific system on the basis of the limited data available to it. (Chomsky, quoted in Mikhail 2005).

Here Chomsky points to three similarities between the development of linguistic knowledge and the development of moral knowledge:

L1: A child raised in a particular linguistic community almost inevitably ends up speaking an idiolect of the local language despite lack of sufficient explicit instruction, lack of extensive negative feedback for mistakes, and grammatical mistakes by caretakers.

M1: A child raised in a particular moral community almost inevitably ends up judging in accordance with an idiolect of the local moral code despite lack of sufficient explicit instruction, lack of sufficient negative feedback for moral mistakes, and moral mistakes by caretakers.

L2: While there is great diversity among natural languages, there are systematic constraints on possible natural languages.

M2: While there is great diversity among natural moralities, there are systematic constraints on possible natural moralities.

L3: Language-speakers obey many esoteric rules that they themselves typically cannot articulate or explain, and which some would not even recognize.

M3: Moral agents judge according to esoteric rules (such as the doctrine of double effect) that they themselves typically cannot articulate or explain, and which some would not even recognize.

L4: Drawing on a limited vocabulary, a speaker can both produce and comprehend a potential infinity of linguistic expressions.

M4: Drawing on a limited moral vocabulary, an agent can produce and evaluate a very large (though perhaps not infinite) class of action-plans, which are ripe for moral judgment.

We will now explain and evaluate each of these pairs of claims in turn.

L1/M1 refer to Chomsky’s poverty of the stimulus argument: despite receiving little explicit linguistic and grammatical instruction, children develop language rapidly and at a young age, and quickly come to display competence applying complex grammatical rules. Similarly, Mikhail and other proponents of the analogy argue that children acquire moral knowledge at a young age based on relatively little explicit instruction. However, critics of the analogy point out several points of difference. First, children actually receive quite a bit of explicit moral instruction and correction, and in contrast with the linguistic case, this often takes the form of explicit statements of moral rules: “don’t hit,” “share,” and so on. Secondly, there is debate over the age at which children actually display moral competence. Paul Bloom (2013; see also Blake, McAuliffe, and Warneken 2014) has argued that babies display moral tendencies as early as 3–6 months, but others (most famously Kohlberg (1969) and Piaget (1970); see also Turiel (1983) and Nucci (2002)) have argued that children are not fully competent with moral judgment until as late as 8–12 years, by which time they have received quite a bit of moral instruction, both implicit and explicit. A related point concerns the ability to receive instructions: in the case of language, a child requires some linguistic knowledge or understanding in order even to receive instruction: a child who has no understanding of language will not understand instructions given to her. But in the moral case, a child need not understand moral rules in order to be instructed in them, so there is less need to posit some innate knowledge or understanding. Nichols et al. (2016) have conducted a series of experiments using statistical learning and Bayesian modeling to show how children might learn complex rules with scant input. Finally, while children may initially acquire the moral values present in their environment, they sometimes change their values or develop their own values later in life in ways that present spontaneously and with little conscious effort. This is in marked contrast with language; as any second-language learner will recognize, acquiring a new language later in life is effortful and, even if one succeeds in achieving fluency, the first language is rarely lost. Lastly, as Prinz (2008) points out, children are punished for moral violations, which may explain why they are quicker to learn moral than grammatical rules.
Until the 1950s, modern moral philosophy had largely focused on either consequentialism or deontology. The

3. Character, Wellbeing, and Emotion

For further empirical investigation, and here again experimental evidence will play a crucial role.

How to individuate moralities, and how many (assuming there are multiple ones) there are, is a question

utterances relative to a language, it may be that we can only evaluate the morality of an action relative to a specific

rules to cases. Finally, the analogy has metaethical implications. Just as we can only evaluate grammaticality of

a kind of rationalism about moral judgment, since it posits that our judgments result from the unconscious application of

of competent speakers. A pressing task will be the systematic collection of such judgments. The analogy also suggests

outlines an important role for experiments in moral theory. If the analogy holds, then an adequate moral theory must be

The linguistic analogy relies on experimental evidence about the nature and pattern of our moral intuitions, and it

so moral judgments can be embedded in other moral judgments to produce novel moral judgments (Harman 2008, 346). For example:

It’s wrong to \(x \rightarrow\) It’s wrong to coerce someone to \(x \rightarrow\) It’s wrong to persuade someone to coerce someone to \(x\)

The point is twofold: first, we can create complex action descriptions; second, we can evaluate novel and complex

actions and respond with relatively automatic intuitions of grammatical or moral permissibility. Mikhail (2011: 43–48)

uses experimental evidence of judgments about trolley problems to argue that our moral judgments are generated by

imposing a deontic structure on our representation of the causal and evaluative features of the action under

consideration. Mikhail points to a variation on the poverty of the stimulus argument, which he calls the poverty of the

perceptual stimulus (Mikhail 2009: 37): when faced with a particular moral situation, we draw complex inferences about

both act and actor based on relatively little data. Mikhail (2010) uses this argument against models of moral judgment

as affect-driven intuitions:

Although each of these rapid, intuitive, and highly automatic moral judgments is occasioned by an identifiable

stimulus, how the brain goes about interpreting these complex action descriptions and assigning a deontic status to

each of them is not something revealed in any obvious way by the surface structure of the stimulus itself. Instead,
an intervening step must be postulated: an intuitive appraisal of some sort that is imposed on the stimulus prior to

every deontic response to it. Hence, a simple perceptual model, such as the one implicit in Haidt’s (2001) influential

model of moral judgment, appears inadequate for explaining these intuitions, a point that can be illustrated by

calling attention to the unanalyzed link between eliciting situation and intuitive response in Haidt’s model

Unlike the traditional poverty of the stimulus argument, this version does not rely on developmental evidence but on our

ability to quickly and effortlessly appraise complicated scenarios, such as variations on trolley problems, and then issue

normative judgments. The postulated intervening step is like the unconscious appeal to rules of grammar, and the evidence

for such a step must come from experiments showing that we do, in fact, have such intuitions, and that they

conform to predictable patterns.

The linguistic analogy relies on experimental evidence about the nature and pattern of our moral intuitions, and it

outlines an important role for experiments in moral theory. If the analogy holds, then an adequate moral theory must be

based on the judgments of competent moral judges, just as grammatical rules are constructed based on the judgments

of competent speakers. A pressing task will be the systematic collection of such judgments. The analogy also suggests

a kind of rationalism about moral judgment, since it posits that our judgments result from the unconscious application of

rules to cases. Finally, the analogy has metaethical implications. Just as we can only evaluate grammaticality of

utterances relative to a language, it may be that we can only evaluate the morality of an action relative to a specific

moral system. How to individuate moralities, and how many (assuming there are multiple ones) there are, is a question

for further empirical investigation, and here again experimental evidence will play a crucial role.

3. Character, Wellbeing, and Emotion

Until the 1950s, modern moral philosophy had largely focused on either consequentialism or deontology. The
revitalization of virtue ethics led to a renewed interest in virtues and vices (e.g., honesty, generosity, fairness, dishonesty, stinginess, unfairness), in eudaimonia (often translated as ‘happiness’ or ‘flourishing’), and in the emotions. In recent decades, experimental work in psychology, sociology, and neuroscience has been brought to bear on the empirical grounding of philosophical views in these areas.

3.1 Character and Virtue

A virtue is a complex disposition comprising sub-dispositions to notice, construe, think, desire, and act in characteristic ways. To be generous, for instance, is (among other things) to be disposed to notice occasions for giving, to construe ambiguous social cues charitably, to desire to give things people want, need, or would appreciate, to deliberate well about what they want, need, or would appreciate, and to act on the basis of such deliberation. Manifestations of such a disposition are observable and hence ripe for empirical investigation. Virtue ethicists of the last several decades have sometimes been optimistic about the distribution of virtue in the population. Alasdair Maclntyre claims, for example, that “without allusion to the place that justice and injustice, courage and cowardice play in human life very little will be genuinely explicable” (1984, 199). Julia Annas (2011, 8–10) claims that “by the time we reflect about virtues, we already have some.” Linda Zagzebski (2010) provides an “exemplarist” semantics for virtue terms that only gets off the ground if there are in fact many virtuous people.

Starting with Owen Flanagan’s Varieties of Moral Personality (1993), philosophers began to worry that empirical results from social psychology were inconsistent with the structure of human agency presupposed by virtue theory. In this framework, people are conceived as having more or less fixed traits of character that systematically order their perception, cognition, emotion, reasoning, decision-making, and behavior. For example, a generous person is inclined to notice and seek out opportunities to give supererogatorily to others. The generous person is also inclined to think about what would (and wouldn’t) be appreciated by potential recipients, to feel the urge to give and the glow of satisfaction after giving, to deliberate effectively about when, where, and how to give to whom, to come to firm decisions based on such deliberation, and to follow through on those decisions once they’ve been made. Other traits are meant to fit the same pattern, structuring perception, cognition, motivation, and action of their bearers. Famous results in social psychology, such as Darley and Batson’s (1973) Good Samaritan experiment, seem to tell against this view of human moral conduct. When someone helps another in need, they may do so simply because they are not in a rush, rather than because they are expressing a fixed trait like generosity or compassion.

In the virtue theoretic framework, people are not necessarily assumed to already be virtuous. However, they are assumed to be at least potentially responsive to the considerations that a virtuous person would ordinarily notice and take into account. Flanagan (1993), followed by Doris (1998, 2002), Harman (1999, 2000), and Alfano (2013), made trouble for this framework by pointing to social psychological evidence suggesting that much of people’s thinking, feeling, and acting is instead predicted by (and hence responsive to) situational factors that don’t seem to count as reasons at all—not even bad reasons or temptations to virtue. These include influences such as ambient sensibilia (sounds, smells, light levels, etc.), seemingly trivial and normatively irrelevant inducers of positive and negative moods, order of presentation of stimuli, and a variety of framing and priming effects, many of which are reviewed in Alfano (2013: 40–50). It’s worth emphasizing the depth of the problem these studies pose. It’s not that they suggest that most people aren’t virtuous (although they do suggest that as well). It’s that they suggest that they undermine the entire framework in which people are conceived as cognitively sensitive and motivationally responsive to reasons. Someone whose failure to act virtuously because they gave in to temptation can be understood in the virtue theoretic framework. Someone whose failure to act virtuously because they’d just been subliminally primed with physical coldness, which in turn is metaphorically associated with social coldness, finds no place in the virtue theoretic framework. These sorts of effects push us to revamp our whole notion of agency and personhood (Doris 2009).

Early estimates suggested that individual difference variables typically explain less than 10% of the variance in people’s behavior (Mischel 1968)—though, as Funder & Ozer (1983) pointed out, situational factors may explain less than 16%.[8] More recent aggregated evidence indicates that situational factors explain approximately twice as much of the variance in human behavior as the five main trait factors (Rauthmann et al. 2014). Convergent evidence from both lexicographical and survey studies indicates that there are at least five dimensions of situations that reliably predict thought, feeling, and behavior: (1) negative valence, (2) adversity, (3) duty, (4) complexity, and (5) positive valence (Rauthmann and Sherman 2018).

According to Doris (2002), the best explanation of this lack of cross-situational consistency is that the great majority of people have local rather than global, traits: they are not honest, courageous, or greedy, but they may be honest-while-in-a-good-mood, courageous-while-sailing-in-rough-weather-with-friends, and greedy-unless-watched-by-fellow-parishioners. In contrast, Christian Miller (2013, 2014) thinks the evidence is best explained by a theory of mixed global traits, such as the disposition to (among other things) help because it improves one’s mood. Such traits are global, in the sense that they explain and predict behavior across situations (someone with such a disposition will, other things being equal, typically help so long as it will maintain her mood), but normatively mixed, in the sense that they are neither virtues nor vices. Mark Alfano (2013) goes in a third direction, arguing that virtue and vice attributions tend to function as self-fulfilling prophecies. People tend to act in accordance with the traits that are attributed to them, whether the traits are minor virtues such as tidiness (Miller, Brickman, & Bolen 1975) and ecology-mindedness (Cornelissen et al. 2006, 2007), major virtues such as charity (Jensen & Moore 1977), cooperativeness (Grusec, Kuczynski, Simutis & Rushton 1978), and generosity (Grusec & Redler 1980), or vices such as cutthroat competitiveness (Grusec, Kuczynski, Simutis & Rushton 1978). On Alfano’s view, when people act in accordance with a virtue, they often do so
It might seem that the criticisms that motivate these novel approaches to virtue miss their mark. After all, virtue ethicists needn’t (and often don’t) commit themselves to the claim that almost everyone is virtuous. Instead, many argue that virtue is the normative goal of moral development, and that people mostly fail in various ways to reach that goal. The argument from the fact that most people’s dispositions are not virtues to a rejection of orthodox virtue ethics, then, might be thought a non sequitur, at least for such views. But empirically-minded critics of virtue ethics do not stop there. They all have positive views about what sorts of dispositions people have instead of virtues. These dispositions are alleged to be so structurally dissimilar from virtues (as traditionally understood) that it may be psychologically unrealistic to treat (traditional) virtue as a regulative ideal. What matters, then, is the width of the gap between the descriptive and the normative, between the (structure of the) dispositions most people have and the (structure of the) dispositions that count as virtues.

Three leading defenses against this criticism have been offered. Some virtue ethicists (Kupperman 2009) have conceded that virtue is extremely rare, but argued that it may still be a useful regulative ideal. Others (Hurka 2006, Merrill 2000) have attempted to weaken the concept of virtue in such a way as to enable more people, or at least more behaviors, to count as virtuous. Still others (Kamtekar 2004, Russell 2009, Snow 2010, Sreenivasan 2002) have challenged the situationist evidence or its interpretation. While it remains unclear whether these defenses succeed, grappling with the situationist challenge has led both defenders and challengers of virtue ethics to develop more nuanced and empirically informed views.

3.2 Wellbeing, Happiness, and the Good Life

Philosophers have always been interested in what makes a human life go well, but recent decades have seen a dramatic increase in both psychological and philosophical research into happiness, well-being, and what makes for a good life. It is important to distinguish here between ‘good life’ in the sense of a life that goes well for the one who lives it, and a morally good life, since philosophers have long debated whether a morally bad person could enjoy a good life. The empirical study of wellbeing focuses primarily on lives that are good in the former sense: good for the person whose life it is. Second, we need to distinguish between a hedonically good life and an overall good life. A life that is hedonically good is one that the subject experiences as pleasant; an overall good life might not contain much pleasure but might be good for other reasons, such as what it accomplishes. We might decide, after investigation, that an overall good life must be a hedonically good life, but the two concepts are distinct.

With these distinctions in mind, we can see the relevance of experimental evidence to investigations in this area. First and perhaps most obviously, experiments can investigate our intuitions about what constitutes a good life, thereby giving us insight into the ordinary concepts of happiness, well-being, and flourishing. To this end, Phillips, Nyholm, & Liao (2014) investigated intuitions about the relationship between morality and happiness. Their results suggest that the ordinary conception of happiness involves both descriptive and normative components: specifically, we judge that people are happy if they are experiencing positive feelings that they ought to experience. So, to use their examples, a Nazi doctor who gets positive feelings from conducting his experiments is not happy. By contrast, a nurse who gets positive feelings from helping sick children is happy, though a nurse who is made miserable by the same actions is not happy.

Another set of experimental findings bearing on this question involves Nozick’s (1974: 44–45) experience machine thought experiment. In response to the hedonist’s claim that pleasure is the only intrinsic good, Nozick asks us to consider the following:

Suppose there was an experience machine that would give you any experience you desired. Super-duper neuropsychologists could simulate your brain so that you would think and feel you were writing a great novel, or making a friend, or reading an interesting book. All the time you would be floating in a tank, with electrodes attached to your brain. Should you plug into this machine for life, preprogramming your life experiences? [...] Of course, while in the tank you won’t know that you’re there; you’ll think that it’s all actually happening [...] Would you plug in?

Nozick argues that our response to the experience machine reveals that well-being is not purely subjective: “We learn that something matters to us in addition to experience by imagining an experience machine and then realizing that we would not use it.” DeBriggard (2010) reports finding that subjects’ intuitions about the experience were different if they were told they were already in such a machine, in which case they would not choose to unplug; he explains this in terms of the status quo bias. Weijers (2014) goes further, asking subjects about Nozick’s original scenario while also asking them to justify their response; he found that many of the justifications implied a kind of imaginative resistance to the scenario or cited irrelevant factors. He also found that respondents were more likely to say that a ‘plugged in life’ would be better when they were choosing for someone else, rather than for themselves.

A second type of study involves investigating the causes and correlates of happiness, well-being, and life-satisfaction. These experiments look at the conditions under which people report positive affect or life-satisfaction, as well as the conditions under which they judge that their lives are going well. This is distinct from the first type of research, since the fact that someone reports an experience as being pleasurable does not necessarily tell us whether they would judge
that experience to be pleasurable for someone else; there may be asymmetries in first- and third-person evaluations. Furthermore, this type of research can tell us how various candidates contribute to well-being from a first-person perspective, but that doesn’t settle the question of the concept of a good life. I might judge that my life is going well, yet fail to notice that I am doing so because I am in a very good mood, and that in fact I am not accomplishing any of my goals; if confronted with another person in a similar situation, I might not make the same judgment. Which of these judgments best represents our concept of well-being is a tricky question, and a normative one since experimental evidence alone may not settle it. As it turns out, experiments have uncovered a number of factors that influence our own reports and assessments of pleasure and well-being. We will discuss two areas of research in particular: reports of pleasures and pains, and judgments of life satisfaction.

First, in the realm of hedonic evaluation, there are marked divergences between the aggregate sums of in-the-moment pleasures and pains and ex post memories of pleasures and pains. For example, the remembered level of pain of a colonoscopy is well-predicted by the average of the worst momentary level of pain and the final level of pain; furthermore, the duration of the procedure has no measurable effect on ex post pain ratings (Redelmeier & Kahneman 1996). What this means is that people’s after-the-fact summaries of their hedonic experiences are not simple integrals with respect to time of momentary hedonic tone. If the colonoscopy were functionally completed after minute 5, but arbitrarily prolonged for another 5 minutes so that the final level of pain was less at the end of minute 10 than at the end of minute 5, the patient would retrospectively evaluate the experience as less painful. This complicates accounts of wellbeing in terms of pleasure (for example, Bentham’s 1789/1961 hedonism) insofar as it raises the question whether the pleasure being measured is pleasure as experienced in the moment, or retrospectively: if, in the very act of aggregating pleasure, we change the way we evaluate it, this is a complication for hedonism and for some versions of utilitarianism. Since well-being is supposed to be a normative notion capable of guiding both individual actions and social policy, findings like these also call into question what, exactly, we ought to be seeking to maximize: pleasurable moments, or the overall retrospective evaluation of pleasure.

Such findings have led some philosophers to seek out alternatives to hedonism in hopes of establishing a more empirically stable underpinning for well-being: in particular, the idea that well-being consists in life satisfaction. The most prominent psychologist in this field is Ed Diener[11] whose Satisfaction with Life Scale asks participants to agree or disagree with statements such as, “I am satisfied with my life” and “If I could live my life over, I would change almost nothing.” These questions seem to get at more stable and significant features of life than hedonic assessments in terms of pleasure and pain, and one might expect the responses to be more consistent and robust. However, two problems arise. The first is that participants’ responses to life satisfaction questionnaires may not be accurate reports of standing attitudes. Fox & Kahneman (1992), for instance, showed that, especially in personal domains people seem to value (friends and love life), what predicts participants’ responses is not recent intrapersonal factors but social comparison. Someone who has just lost a friend but still thinks of herself as having more friends than her peers will tend to report higher life satisfaction than someone who has just gained a friend but who still thinks of himself as having fewer friends than his peers.

Life satisfaction surveys also seem to be subject to order effects. For instance, if a participant is asked a global life satisfaction question and then asked about his romantic life, the correlation between these questions tends to be near zero, but if the participant is asked the dating question first, the correlation tends to be high and positive (Strack, Martin, & Schwarz 1988).[12] This suggests that life-satisfaction judgments might be unduly influenced by the topics and questions considered just before making a life-satisfaction judgment. Another example is Schwarz and Clore’s oft-cited 1983 study, in which the authors reported a correlation between weather and life-satisfaction when subjects were asked about weather first, but not when the weather question followed the life-satisfaction query. We note, however, that there is significant controversy about this oft-cited example. Some recent papers claim to find a correlation between the two (e.g. Connolly 2013) while others claim there is no evidence for an effect of weather on life-satisfaction judgments (e.g. Lucas & Lawless 2013).

Findings like the above have led some researchers (e.g., Haybron 2008) to argue that life-satisfaction judgments are too arbitrary to ever satisfy the requirements of a theory of well-being. In response, Tiberius and Plakias (2011) argue for an idealized life-satisfaction theory they call value based life satisfaction, suggesting that by asking subjects to consider their life-satisfaction while attending to the domains they most value much of the instability plaguing the studies described above is removed, a claim they support with research showing that judgments made after priming subjects to think about their values demonstrate higher levels of retest stability (Schimmack & Oishi 2005).

### 3.3 Emotion and Affect

As much of the previous discussion reveals, the relationship between emotion and moral judgment is one of the central concerns of both traditional and experimental moral philosophy. Our discussion of this topic will focus on two types of research: the role of emotion in moral reasoning generally, and the role of one specific emotion—disgust—in moral judgments.

One debate concerns whether hot, emotion-driven reasoning is necessarily better or worse than reasoning based only on cooler, more reflective thinking—a distinction sometimes referred to using Kahneman’s terminology of system 1/system 2 thinking. The terminology is not perfect, though, because Kahneman’s terms map onto a quick, automatic, unconscious system of judgments (system 1) and a slow, effortful, deliberative decision-making process (system 2), and as we saw above, this is not a distinction between emotion and reason, since rule-based judgments can be automatic and unconscious while emotional judgments might be effortful and conscious. The debate between Mikhail...
and Haidt is a debate over the extent to which emotions rather than rules explain our moral judgments; Singer and Greene’s arguments against deontology rest on the claim that emotion-backed judgments are less justified than their utilitarian counterparts.

One reason for thinking that moral judgments essentially involve some emotional component is that they tend to motivate us. Internalism is the view that there is a necessary connection between moral judgment and motivation. This contrasts with externalism, which doesn’t deny that moral judgments are usually motivating, but does deny that they are necessarily so, as the link between judgment and motivation is only contingent. Since emotions are intrinsically motivational, showing that moral judgments consist, even in part, of emotions would vindicate internalism. One route to this conclusion has involved surveying people’s intuitions about moral judgment. Nichols (2002: 289) asked subjects whether an agent who “had no emotional reaction to hurting other people” but claims to know that hurting others is wrong really understands that hurting others is wrong. He found that most subjects did attribute knowledge to such agents, suggesting that the ordinary concept of moral judgment is externalist, a claim that is further supported by Strandberg and Björklund (2013). An additional source of evidence comes from psychopaths and patients with traumatic brain injuries, both of whom show evidence of impaired moral functioning—though how one regards this evidence depends on which perspective one begins with: while externalists (Nichols 2002, Roskies 2003) claim that the existence of psychopaths provides a counterexample to the claim that moral judgment is motivating (since psychopaths lack empathy and an aversion to harming others), internalists (Smith 1994, Maibom 2005, Kennett and Fine 2008) argue that psychopaths don’t actually make full-fledged moral judgments. Psychologist Robert Hare (1993: 129) quotes a researcher as saying that, they “know the words but not the music.” Psychopaths also display other cognitive and affective deficiencies, as evidenced by their poor decision-making skills in other areas. This may mean that they should not be taken as evidence against internalism.

A reason for thinking that moral judgments ought not involve emotion is that emotions sometimes seem to lead to judgments that are distorted or off-track, or that seem otherwise unjustified. One emotion in particular is often mentioned in this context: disgust. This emotion, which seems to be unique to human animals and emerges relatively late in development (from the ages of about 5–8), involves a characteristic gaping facial expression, a tendency to withdraw from the object of disgust, a slight reduction in body temperature and heart rate, and a sense of nausea and the need to cleanse oneself. In addition, the disgusted subject is typically motivated to avoid and even expunge the offending object, experiences it as contaminating and repugnant, becomes more attuned to other disgusting objects in the immediate environment, and is inclined to treat anything that the object comes in contact with (whether physically or symbolically) as also disgusting. This last characteristic is often referred to as “contamination potency” and it is one of the features that makes disgust so potent and, according to its critics, so problematic. The disgust reaction can be difficult to repress, is easily recognized, and empathically induces disgust in those who do recognize it. There are certain objects that almost all normal adults are disgusted by (feces, decaying corpses, rotting food, spiders, maggots, gross physical deformities). But there is also considerable intercultural and interpersonal variation beyond these core objects of disgust, where the emotion extends into food choices, sexual behaviors, out-group members, and violations of social norms. Many studies have claimed to show that disgust is implicated in harsher moral judgments (Schnall, Haidt, & Clore 2008), citing experiments in which subjects filling out questionnaires in smelly or dirty rooms evaluated moral transgressions more harshly. Others have gone further and argued that disgust might itself cause or comprise a part of moral judgment (Haidt 2001; Wheatley & Haidt 2005). If this is true, critics argue, we ought to be wary of those judgments, because disgust has a checkered past in multiple senses. First, it’s historically (and currently) associated with racism, sexism, homophobia and xenophobia; the language of disgust is often used in campaigns of discrimination and even genocide. Secondly, disgust’s evolutionary history gives us reason to doubt it. Kelly (2011) argues that the universal bodily manifestations of disgust evolved to help humans avoid ingesting toxins and other harmful substances, while the more cognitive or symbolic sense of offensiveness and contamination associated with disgust evolved to help humans avoid diseases and parasites. This system is later recruited for an entirely distinct purpose: to help mark the boundaries between in-group and out-group, and thus to motivate cooperation with in-group members, punishment of in-group defectors, and exclusion of out-group members.

If Kelly’s account of disgust is on the right track, it seems to have a number of important moral upshots. One consequence, he argues, is “disgust skepticism” (139), according to which the combination of disgust’s hair trigger and its ballistic trajectory mean that it is especially prone to incorrigible false positives that involve unwarranted feelings of contamination and even dehumanization. Hence, “the fact that something is disgusting is not even remotely a reliable indicator of moral foul play” but is instead “irrelevant to moral justification” (148).

It is important to note that the skeptical considerations Kelly raises are specific to disgust and its particular evolutionary history, so they aren’t intended to undermine the role of all emotions in moral judgment. Still, if Kelly is correct, and if disgust is implicated in lots of moral judgments, we may have a reason to be skeptical of many of our judgments. Plakias (2011, 2017) argues against the first half of this antecedent, claiming that Kelly and other ‘disgust skeptics’ are wrong to claim that the purposes of moral and physical disgust are totally distinct; she suggests that disgust is sometimes a fitting response to moral violations that protects against social contagion. May (2014) argues against the second half, claiming that disgust’s role in moral judgment has been significantly overblown; at most, we have evidence that disgust can, in certain cases, slightly amplify moral judgments that are already in place. However, more recent empirical work indicates that incidental disgust has little effect on the harshness of moral judgments, though dispositional disgust-sensitivity does (Landy and Goodwin 2015).
4. Metaethics and Experimental Moral Philosophy

Metaethics steps back from moral theorizing to ask about the nature and function of morality. While first-order ethics seeks to explain what we should do, metaethics seeks to explain the status of those theories themselves: what are their metaphysical commitments, and what evidence do we have for or against them? Which epistemology best characterizes our moral practices? What is the correct semantics for moral language? These questions might not seem obviously empirical, but insofar as it attempts to give an account of moral semantics, epistemology, and ontology, metaethics aims, in part, to capture or explain what we do when we engage in moral talk, judgment, and evaluation. To the extent that metaethics sees itself as characterizing our ordinary practice of morality, it is therefore answerable to empirical data about that practice. To the extent that a theory claims that we are mistaken or in widespread error about the meaning of moral language, or that we lack justification for our core moral beliefs, this is taken to be a strike against that theory. For example, relativism is often criticized on the grounds that it requires us to give up the (putatively) widespread belief that moral claims concern objective matters of fact and are true or false independently of our beliefs or attitudes. We have already seen several ways that experimental data bears on theories about moral reasons (the debate between internalists and externalists) and the epistemology of moral judgment (the debate over the role of intuitions). In this section we will examine experimental contributions to the debate over moral realism, arguments about moral disagreement, and moral language.

4.1 Folk Metaethics and Moral Realism

Much contemporary metaethics relies on assumptions about the nature of ordinary moral thought, discourse, and practice; metaethicists tend to see their project as essentially conservative. For example, Michael Smith writes that the first step in metaethical theorizing is to “identify features that are manifest in ordinary moral practice” and the second step is to “make sense of a practice having these features.” (1994) This assumption has had a major impact on the debate between realists and antirealists, with realists claiming to best capture the nature of ordinary moral discourse: “We begin as (tacit) cognitivists and realists about ethics,” says Brink (1989), and then “we are led to some form of antirealism (if we are) only because we come to regard the moral realist’s commitments as untenable… Moral realism should be our metaethical starting point, and we should give it up only if it does involve unacceptable metaphysical and epistemological commitments.”

But experimental work has cast doubt on this claim, beginning with Darley and Goodwin (2008), and continued by James Beebe (2014) and others (Wright et al. 2013; Campbell and Kumar 2012; Goodwin and Darley 2010; and Sarkissian et al. 2011). Goodwin and Darley asked subjects to rate their agreement with statements drawn from the moral, ethical, and aesthetic domain, and asked subjects whether they agreed with the statement (for example, “before the 3rd month of pregnancy, abortion for any reason (of the mother’s) is morally permissible,”), whether they thought it represented a fact or an opinion or attitude, and whether, if someone were to disagree with them about the statement, at least one of the disputants would have to be mistaken. (We will say more about the authors’ use of disagreement as a proxy for realism in the following section.) In general, subjects rated moral statements as less factual than obviously factual statements (e.g. “the earth is at the center of the universe,”) but more factual than statements about matters of taste or etiquette. What is striking about these findings is not just that people are not straightforwardly realist, but that they seem to treat moral questions variability: some are treated as matters of fact, others as matters of opinion. This pattern has been replicated in several studies, and persists even when subjects are allowed to determine for themselves which issues to assign to the moral domain, suggesting that subjects do not think moral claims are uniformly objective.

A question that remains is how to explain this variation: are people being inconsistent? Are they tracking something besides the moral question—perhaps the degree of consensus that exists about a question? Further research may help shed some light on these issues, but even if the questions are answered in the affirmative, the realist’s claim to capture folk morality is called into question, since these experiments suggest that folk intuitions are either not uniformly realist, or they are confused (and hence poorly suited for a role in metaethical theorizing). Another response to these findings is that they reveal a kind of folk metaethical pluralism. For example, Michael Gill (2008, 2009) and Walter Sinnott-Armstrong (2009) have suggested a noncognitivist treatment of some areas of morality and cognitivism about others. This response is under-motivated by empirical data, however, since such data shows at most that our judgments about the extent to which moral claims describe mind-independent facts is variable, and not that their semantics itself varies. Nor does such data settle the question whether our moral judgments are beliefs or another conative state: while experimental data can reveal a role for emotion in moral judgment, as discussed above, it does not show that in such cases belief is absent. We will discuss limitations on experimental moral philosophy in more detail in Section 5 below. For now, it seems that data about folk realism are best viewed as undermining one of the most commonly-cited sources of evidence for moral realism, namely, that it captures our ordinary moral discourse and practice better than its competitors. A further question, open to exploration, is whether an anti-realist view might do better at capturing and explaining our folk intuitions than realist competitors.

4.2 Moral Disagreement

As we saw at the beginning of this entry, some of the earliest thinking in empirically-informed moral philosophy concerns moral disagreement. Brandt and Ladd conducted in-depth investigations into the moral codes of other groups, and some contemporary moral philosophers have argued for continuing attention to the empirical details of moral disagreement. This is partly due to the influence of Mackie’s (1977: 36–37) formulation of what he dubbed ‘the
debate seems ripe for experimental investigation, the noncognitivist typically acknowledges that their view is a language has a fundamentally different function, to express attitudes or to issue commands, for example. While this actions or characters. Noncognitivists (including their contemporary representatives, the expressivists) hold that moral assertions, expressing propositions that attribute moral properties (such as rightness or goodness) to things (such as cognitivism and noncognitivism. According to the cognitivist, ordinary moral sentences are best understood as factual properties to acts and things, rather than reporting or expressing speakers' attitudes, can capture this intuition.

The claim here is not about realism per se, but about whether experimental evidence can give us insight into the semantics of moral language, since our intuitions about whether there is genuine faultlessness can tell us about the existence of moral disagreements. Rather than conjecture about the outcomes of possible moral disagreements, for example, Doris and Plakias (2008) discuss several instances of cross-cultural moral disagreement, and assess the prospects for applying what they call ‘defusing explanations’ to these cases. A defusing explanation is one that accounts for the disagreement in terms of a non-moral difference or an epistemic shortcoming on the part of one or more disputant, thereby showing that it is not, in fact, a moral disagreement. Their argument and the experiments they cite are discussed in detail in the entry on empirical moral psychology, so we will not discuss them in detail here; for now, we will note that such explanations are difficult to assess via survey methods alone. This is why it is especially helpful to look to the anthropological record, as Oliver Curry et al. (forthcoming) do in a recent landmark publication. Curry hypothesizes that morality comprises stable, cooperative solutions to recurrent problems that can be modeled as non-zero-sum games. There are a variety of such recurrent problems, and game theory has established a suite of analytical tools for diagnosing and solving them. Curry and his colleagues show that such solutions are always seen as either morally good or morally neutral in societies at various stages of development, from small-band hunter gatherers to industrialized democracies. This research suggests that philosophers may have greatly exaggerated the extent of moral disagreement that actually exists.

4.3 Moral Language

One reason disagreement is a useful measure of objectivity is that the impossibility of faultless disagreement is taken to be characteristic of questions concerning objective matters of fact. That is, while we might concede the possibility of faultless disagreements about matters concerning the deliciousness of foods or the merits of a sports team, when we disagree over moral issues, we think that there is a correct answer and that therefore at least one disputant is getting things wrong (for a discussion of the analogy between food and morality, see Loeb 2003). To the extent that people judge a disagreement to be faultless, we might think that they are evincing a kind of anti-realism about the issue under dispute. This is yet another way disagreement can bear on the realism/antirealism debate. Can experimental evidence involving disagreement tell us anything about the semantics of moral language?

One reason to think it can is that the very idea of faultless disagreement strikes some as incoherent. In cases where one individual thinks (or says) that pistachio ice cream is delicious and another thinks (or says) that it is disgusting, we understand the two parties as expressing their own personal preference. To explain the apparent faultlessness of the dispute, we reinterpret their thoughts or utterances as expressing something like, “pistachio ice cream is delicious [disgusting] to me.” Because the two parties do not genuinely contradict one another, we need not say that one of them is mistaken. Faultlessness is therefore supposed to be a point in favor of the moral anti-realist, since it seems to imply that there is no single content about which the parties disagree (and about which one might be mistaken).

By contrast, realists claim that their view is better able to capture our intuition that there is a genuine disagreement at stake when one person says that stealing is wrong and another says that stealing is not wrong. In these cases, realists argue, we have the intuition that there really is a conflict, and only a theory on which moral language involves attributing properties to acts and things, rather than reporting or expressing speakers’ attitudes, can capture this intuition.

We have already seen some data bearing on this issue in our discussion of folk realism above. The claim here is not about realism per se, but about whether experimental evidence can give us insight into the semantics of moral language, since our intuitions about whether there is genuine faultlessness can tell us about whether there is a single shared content between two utterances—a single proposition that one party asserts and another denies—or whether the two parties’ utterances contain implicit relativizations. This has, at least, been the assumption guiding much of the debate over disagreement and its implication for moral semantics. In recent work, however, John Khoo and Joshua Knobe (2016) have cast doubt on this assumption; their experiments indicate that subjects do not see disagreement as requiring exclusionary content—a single proposition that one party accepts and the other rejects. This research suggests that intuitions about disagreement may not be as straightforwardly tied to moral semantics as philosophers have thought: judgments that two individuals genuinely disagree about some claim do not necessarily imply a single shared content between the two speakers. This work is still in early stages, but it reveals the complications involved in reading semantics off disagreement.

A further challenge for attempts to experimentally investigate moral semantics is the debate within antirealism between cognitivism and noncognitivism. According to the cognitivist, ordinary moral sentences are best understood as factual assertions, expressing propositions that attribute moral properties (such as rightness or goodness) to things (such as actions or characters). Noncognitivists (including their contemporary representatives, the expressivists) hold that moral language has a fundamentally different function, to express attitudes or to issue commands, for example. While this debate seems ripe for experimental investigation, the noncognitivist typically acknowledges that their view is a
reforming one. Furthermore, the semantics of our terms may be opaque to ordinary users, to the extent that we cannot read off semantics from intuitions but must investigate them indirectly. Lastly, the cognitivist and noncognitivist can agree on a number of empirical features of moral judgment and discourse; the challenge for experimentalists is to find predictions that confirm one view while disconfirming the other. This is a relatively new area of investigation and, while not without challenges, ripe for exploration.

5. Criticisms of Experimental Moral Philosophy

Experimental moral philosophy far outstrips what we have been able to cover here, and many issues and areas of productive research have barely been touched upon. For example, experimental evidence is relevant to moral questions in bioethics, such as euthanasia, abortion, genetic screening, and placebogenic interventions. Likewise, experiments in behavioral economics and cognitive psychology are being employed in asking moral questions about public policy (Bicchieri and Chavez 2010; Bicchieri and Xiao 2009). We neglect these issues only because of lack of space. In the few words remaining, we explore some potential criticisms of experimental philosophy.

5.1 Problems with Experimental Design and Interpretation

As a young field, experimental philosophy suffers from various problems with experimental design and interpretation. These are not insurmountable problems, and they are problems faced by related fields, such as social psychology, cognitive psychology, and behavioral economics.

One issue that has recently come to the fore is the problem of replication. Statistical analysis is not deductive inference, and the mere fact that statistical analysis yields a positive result does not guarantee that anything has been discovered. Typically, an experimental result is treated as “real” if its \( p \)-value is at most .05, but such a value just indicates the probability that the observation in question would have been made if the null hypothesis were true. It is not the probability that the null hypothesis is false given the observation. So, even when statistical analysis indicates that the null hypothesis is to be rejected, that indication can be fallacious. Moreover, when multiple hypotheses are tested, the chance of fallaciously rejecting the null hypothesis at least once rises exponentially.

We should expect other failures of replication because of a bias built into the system of funding experimentation and publishing results. The proportion of published false-positives is much higher for unexpected and unpredicted results than for expected and predicted results. Since experimentalists are reluctant to report (and even discouraged by journal editors and referees from reporting) null results (i.e., results where the \( p \)-value is more than .05), for every published, unexpected result there may be any number of unpublished, unexpected non-results. Thus, unexpected, published results are more likely to be false positives than they might appear.

The best way to tell whether such a result carries any evidential value is for the experiment to be replicated—preferably by another research group. If a result cannot be replicated, it is probably a mirage. Such mirages have turned up to a disturbing extent recently, as Daniel Kahneman has famously pointed out (Yong 2012; see also Wagenmakers et al. 2012). Kahneman proposed a “daisy chain” of replication, where no result in psychology would be published until it had been successfully replicated by another prominent lab. This proposal has not yet been (and may never be) instituted, but it has raised the problem of replication to salience, and a related project has taken off. The reproducibility project in psychology aims to establish the extent to which prominent, published results can be replicated. Experimental philosophers have followed suit with their own replication project (see the Other Internet Resources). This work has recently yielded encouraging fruit: a large collaborative replication effort suggests that approximately 7 out of 10 experimental philosophy results can be reproduced, whereas similar efforts show that less than half of the results in biomedicine, economics, and psychology are replicable (Cova et al. 2018, Other Internet Resources).

A related issue with experimental design and interpretation has to do with “fishing expeditions.” An experiment is sometimes pejoratively referred to as a fishing expedition if no specific predictions are advanced prior to data collection, especially when multiple hypotheses are tested. Likewise, critics argue that researchers such as Weinberg, Nichols, & Stich (2008) have put forward hypotheses that are formulated so vaguely as to make them nearly unfalsifiable. This problem is exacerbated when no guiding positive heuristic is offered to explain observed patterns.

Another worry is that simultaneously testing many intuition-probes can lead unwary experimenters on snipe hunts. Suppose an experimental philosopher conducts an experiment with two conditions: in the experimental condition, participants are primed with deterministic ideas, while in the control condition they are not primed one way or another. She asks participants twenty different questions about their moral intuitions, for instance, whether there is free will, whether malefactors deserve to be punished, whether virtue deserves to be rewarded, and so on. She then makes pairwise comparisons of their responses to each of the questions in an attempt to figure out whether deterministic priming induces changes in moral intuitions. She thus makes twenty independent comparisons, each at the industry-standard 5% level. Suppose now for the sake of argument that there is no effect—that all null hypotheses are true. In that case, the probability that at least one of these tests will result in a Type I error (rejecting the null hypothesis even though it is true) is 64%. More generally, when an experimenter performs \( n \) independent comparisons at the 5% level, the probability of at least one Type I error is \( 1 - (1 - .05^n) \). This problem can be addressed by various procedures, most notably the Tukey method and the Bonferroni method.\[17\]

One way to establish that a result is real is to see whether the experiment has a respectable effect size. The most common measure of effect size is Cohen’s \( d \), which is the ratio of the difference in means between conditions to the
standard deviation of the relevant variable. So, for example, a \( d \) of 1.0 would indicate that a manipulation moved the mean of the experimental condition an entire standard deviation away from the mean of the control condition, a huge effect. Unfortunately, effect sizes have, until recently, rarely been reported in experimental philosophy, though that appears to be changing.

### 5.2 Philosophical Problems

One might object that the real problem with experimental moral philosophy is not with science, but with its relevance (or more precisely, its irrelevance) to moral philosophy. The objection is that moral philosophy is concerned not with how we are and what we do, but with how we ought to be and what we ought to do. As such, it is a normative enterprise, and is unaffected by empirical results. Science is hard, but so is morality; if we fail to understand and live up to the demands of the latter, that doesn’t show a problem with our moral theory but with our moral natures.

Experimental philosophers can agree with this objection, up to a point. No one is suggesting that we read morality directly off survey results. But, to return to the point with which we began, while morality is normative, it is also essentially practical. Moral theory is a theory about what we ought to do: a theory about how creatures like us ought to conduct ourselves and interact with one another. A morality completely divorced from our natures, that demanded acts impossible from us, would surely be unacceptable. This is not just a point about experimental philosophy; utilitarianism is sometimes critiqued as making unrealistic demands of impartiality. Alongside the famous proscription against deriving an *ought* from an *is*, we also find the dictum that *ought implies can*. The exact extent to which morality can be demanding without being unrealistic is itself a philosophical question, and so experimental moral philosophy works in conjunction with philosophical analysis; it does not aim to eliminate it. Exactly how the two relate to each other, and how empirical evidence will bear on and influence debates in moral theory in the future, is a contested issue, but surely traditional moral theory and experimental moral philosophy have much to learn from one another.

We need to proceed cautiously here. No one doubts that what we ought to do depends on how things are non-morally. For example, the moral claim that a given man deserves to be punished presupposes the non-moral fact that he committed the crime. It should come as no surprise, then, that experimental evidence might be relevant in this way to morality. Whether experimental evidence is relevant to discovering the fundamental moral principles—those meant to be put into action one way or another depending on how the world is non-morally—is still subject to debate.

Another version of this argument says that fundamental moral philosophical principles are, if true at all, necessarily true, and that empirical research can establish at best only contingent truths. But if fundamental moral theories are necessary, then they are necessary for creatures like us. And one thing that empirical investigation can do is to help establish what sorts of creatures we are. Imagination needs material to work with. When one sits in one’s armchair, imagining a hypothetical scenario, one makes a whole host of assumptions about what people are like, how psychological processes work, and so on. These assumptions can be empirically well or poorly informed. It’s hard to see why anyone could doubt that being well informed is to be preferred. Likewise it’s hard to see how anyone could doubt the relevance of experimental evidence to better grounding our empirical assumptions, in this case our assumptions relevant to moral philosophy. Exactly how experimental—or more broadly, empirical—evidence is relevant, and how relevant it is, are at present hotly contested matters.

### Bibliography


Hare, R., 1993. Without Conscience: The Disturbing World of the Psychopaths Among Us, New York: Guilford Press.


Other Internet Resources

- Moral Psychology Research Group
- Experimental Philosophy Blog
- Replications in Experimental Philosophy
- Blogging Heads TV: Mind Report
- Positive Psychology Center
- Online Experiments
- University of Missouri Experimental Philosophy Lab
- Yale Experimental Philosophy Lab
- Porto Experimental Philosophy Lab
- Moral Foundations
- U Penn Behavioral Ethics Lab

Related Entries

character, moral | character, moral: empirical approaches | cognitivism vs. non-cognitivism, moral | emotion | ethics: virtue | intuition | moral anti-realism | moral epistemology | morality: and evolutionary biology | moral psychology: empirical approaches | moral realism | moral relativism | well-being

Acknowledgments

The authors thank James Beebe, Gunnar Björnsson, Wesley Buckwalter, Roxanne DesForges, John Doris, Gilbert Harman, Dan Haybron, Chris Heathwood, Antti Kauppinen, Daniel Kelly, Joshua Knobe, Clayton Littlejohn, Edouard Machery, Josh May, John Mikhail, Sven Nyholm, Brian Robinson, Chandra Sripada, Carissa Veliz, several anonymous referees, and the editors of this encyclopedia for helpful comments and suggestions.

Copyright © 2018 by
Mark Alfano <mark.alfano@gmail.com>
Don Loeb
Alexandra Plakias <aplakias@hamilton.edu>
The emerging field known as experimental philosophy has expanded into moral philosophy: by presenting experimental subjects with vignettes describing scenarios with moral implications, data about people’s moral intuitions are gathered and analyzed. This paper examines the adequacy of applying the common methodology of experimental philosophy to the study of moral thought. By employing Raimond Gaita’s notion of Experimental moral philosophy emerged as a methodology in the last decade of the twentieth century, as a branch of the larger experimental philosophy (X-Phi) approach. Experimental moral philosophy is the empirical study of moral intuitions, judgments, and behaviors. Like other forms of experimental philosophy, it involves gathering data using experimental methods and using these data to substantiate, undermine, or revise philosophical theories.