Experimental Moral Philosophy

Experimental moral philosophy began to emerge as a methodology in the last decade of the twentieth century, a branch of the larger experimental philosophy (X-Phi, XΦ) approach. From the beginning, it has been embroiled in controversy on a number of fronts. Some doubt that it is philosophy at all. Others acknowledge that it is philosophy but think that it has produced modest results at best and confusion at worst.

1. Introduction
   1.1 What Counts as an Experiment?
   1.2 Who Conducts the Experimental Investigation?
   1.3 How Directly Must the Experimentation be Related to the Philosophy?
   1.4 What Counts as Experimental Philosophy, as Opposed to Psychology?

2. Moral Judgments and Intuitions
   2.1 Two “Negative” Programs
   2.2 Three “Positive” Programs
   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
   3.3 Emotion and Affect

4. Some Metaethical Issues
   4.1 Moral Disagreement
   4.2 Moral Language

5. Criticisms of Experimental Moral Philosophy
   5.1 Problems with Experimental Design and Interpretation
1. Introduction

Before the research program can be evaluated, we should have some conception of its scope. But controversy surrounds questions about its boundaries as well. Uncontroversially, the distinction between experimental and non-experimental philosophy is not identical to the distinction between \textit{a posteriori} and \textit{a priori} philosophy. Experimental evidence is a proper subset of empirical evidence, which is itself a subset of \textit{a posteriori} evidence. Can any more be said?

One reason it is so difficult to specify the boundaries of the program is that the current surge in interest in experimental philosophy is relatively new. Another is that practitioners of experimental philosophy are pulled in opposing directions. On the one hand, there is a desire to see experimental philosophy as continuous with the history of philosophy, and thus to view the class broadly. Admirers are likely to think that there is little room for doubt about the approach's claim to philosophical respectability, arguing that it is really just an extension of traditional philosophical approaches going back to Aristotle. He and many others in the philosophical canon would not have seen an important distinction between science and philosophy. They were as interested in “natural philosophy”—or science—as in anything else; distinctions among fields of study are relatively recent phenomena.\footnote{On the other hand, there is an inclination to think of the methodology as novel—even revolutionary—and thus to characterize it more narrowly. When we factor in issues about the scope of moral philosophy in general, things become even more complex. Still, there are a number of dimensions along which we might wish to understand the field more or less broadly. Among the most important are:}

1.1 What Counts as an Experiment?

Paradigmatic experiments involve randomized assignment to varying conditions of objects from a representative sample, followed by statistical comparison of the outcomes for each condition. The variation in conditions is sometimes called a \textit{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 payphone, and then measure the degree to which they subsequently help someone they take to be in need.\footnote{Finding or not finding the dime is the condition; degree of helpfulness is the outcome variable.} Finding or not finding the dime is the condition; degree of helpfulness is the outcome variable.

While true experiments follow this procedure, “studies” allow non-random assignment to condition. Astronomers, for example, sometimes speak of “natural experiments”. Although we are in no position to affect the 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 cultures' members, but such studies are valuable and empirically respectable. Still, for various reasons, the evidential value of studies is less than that of experiments, and indeed, most published research in experimental philosophy involves true experiments.

Arguably, other modes of inquiry such as historical research should also be treated as natural and legitimate extensions of paradigmatically experimental approaches. Indeed, some would treat history itself as a series of quasi-experiments in which different ways of life produce a
variety of outcomes. In all these cases, what matters most is whether the conclusions of the inquiry are grounded in evidence rising to the standards set by social and natural science.

The distinction between studies and experiments primarily concerns input variables. Another distinction has to do with outcome variables. One of the paradigmatic outcome measures of experimental philosophy is the survey, though others such as functional magnetic resonance imaging (fMRI), physiological measures, and behavioral measures are also in use. There are decades of top-flight research in personality and social psychology using the survey paradigm, but it is also subject to trenchant criticism. Instead of revealing what people think, surveys might establish what people think they think, what they think the experimenter wants them to think, what they think other participants think, or just something they made up because they had to provide a response. This is a worry that arises for all surveys, but it is especially pertinent in the context of experimental philosophy because participant responses are often treated as equivalent to “what one would say” or even “how one would judge.”

The survey paradigm suffers from other methodological drawbacks as well. Presumably, morality is not just a matter of what people would say, but also a matter of what they would notice, do and feel, as well as how they would reason and respond. Surveys provide only tenuous and indirect evidence about these activities, and so it might behoove experimental philosophers to employ or devise outcome measures that would more reliably and directly capture these features of morality.

How closely tied to true experiments must a philosophical investigation be, then, if it is to count as experimental moral philosophy? According to the most capacious view, the guiding principle of experimental philosophy is that when philosophical arguments invoke or make assumptions about empirical matters, those assumptions should be assessed according to the best natural and social scientific evidence available, and that if such evidence is not currently available it should where possible be acquired, either by seeking the help of specialists with the relevant scientific training or by having well-trained philosophers conduct the research themselves. Thus, the term, “experimental moral philosophy” might be replaced by a phrase like empirically well-informed moral philosophy. While this seems unobjectionable, there may still be local objections to the scientific methodology and to the relevance or probative force of empirical research in particular cases. Some of these are discussed in the concluding section of this entry.

1.2 Who Conducts the Experimental Investigation?

A good deal of the work in experimental moral philosophy involves experiments performed by philosophers in pursuit of answers to philosophical questions. In other cases, philosophers draw on experimental results produced by social scientists. It is hard to see why it would be important that the philosopher herself have done the experimental work. By way of illustration, consider the case of Richard Brandt, the well-known moral philosopher, who spent time living among the Hopi Native Americans as an anthropological observer in an effort to understand their ethical views. During that time, he conducted the empirical investigation himself (Brandt 1954), but in later work (1998) he called upon a great deal of research by others, many of whom were social scientists by training, not philosophers. It seems arbitrary to treat the former work as of a fundamentally different sort from the latter. Both qualify as fine examples of empirically well-informed moral philosophy.

Surely what matters, then, is not the identity or affiliation of the experimenter, but whether the experiment is well-designed, as free of confounds as possible, sufficiently highly powered, carefully interpreted, informed by the existing literature, and so on. Philosophers
can be especially insightful in helping to design experiments that are not plagued by such problems, partly because one of the capacities cultivated by philosophical training is the ability to imagine and construct potential counterexamples and counterarguments. When the conclusions aimed at are recognizably philosophical, there seems to be little room for doubt that philosophers should be included in both experimental design and the interpretation of results. They are likely to improve the operationalization of the relevant concepts and to better understand the philosophical implications of the results.

1.3 How Directly Must the Experimentation be Related to the Philosophy?

Moral philosophers often rely on beliefs about the way the world is, how it works, and how it got that way—assumptions whose empirical credentials are impeccable but difficult to reconstruct. In many cases, the justification for the beliefs is unimportant so long as their justificatory status is well-established. A philosopher considering the moral permissibility of stem cell research, for example, ought to understand the nature of these cells, gestation and embryonic development, the aims of the research projects employing (or seeking to employ) stem cells, and their prospects for success. Knowledge about these matters depends in many ways on the results of scientific experimentation, though perhaps conducted long ago and without any expectation of philosophical payoff.

Still, in some cases it is important that an experiment aim at answering a particular question, since this can improve experimental design. Experiments that merely aim to find out “what will happen when…” tend to be under-theorized, to rely on the wrong distinctions, and even to be uninterpretable. Well-designed confirmatory experiments make specific predictions in advance of data-collection, rather than spinning post hoc “just-so stories” after the data are in. Such stories can be useful in the context of exploring the hypothesis space, but they are not themselves evidence. Since confirmatory experiments can provide evidence that is useful in ethical (or more broadly philosophical) reasoning, philosophers have an important role to play in their design and interpretation.

1.4 What Counts as Experimental Philosophy, as Opposed to Psychology?

As is often the case, there are examples at either extreme for which it's easy to give an answer. Virtually any question in ethics will have an empirical dimension. Most people would agree that it is impermissible to torture an animal just for fun and that the reason has something to do with the pain the animal would experience. No doubt part of our evidence that the animal would indeed suffer is grounded in experimental or at least empirical knowledge of animal psychology. But it seems overbroad to think of this case as arising in experimental moral philosophy or psychology. Thinking in terms of empirically well-informed moral philosophy seems to moot the issue of how to characterize a case like this, and arguably captures the most important motivations underlying the more common phraseology.

Other cases are more easily classified as belonging to disciplines other than philosophy. It seems more accurate to think of research showing that people are much more inclined to be organ donors if donation is the default position and those wishing not to donate are permitted to opt out (Gigerenzer 2008, 2–3) as psychology or sociology rather than philosophy. But good thinking about the question of what sort of policies, practices, and institutions we morally ought to adopt should take note of this evidence.\textsuperscript{[5]} Once again, moving the focus from experimental to empirically well-informed moral philosophy might be thought to moot the concern about how to characterize such cases without sacrificing much that is useful to understanding the various research projects. Indeed, the
problem of what counts as experimental moral philosophy is just a particular application of the vexed (but perhaps unimportant) question of what counts as philosophy in the first place.

Here is a roadmap for the rest of this entry. Section 2 canvasses experimental research on moral judgments and intuitions, describing various programmatic uses to which experimental results have been put, then turning to specific examples of experimental research on moral judgment and intuition, including intuitions about intentionality and responsibility, as well as the so-called “linguistic analogy.” Section 3 discusses experimental results on “thick” (i.e., simultaneously descriptive and normative) topics, including character and virtue, wellbeing, and emotion and affect. Section 4 discusses questions about moral disagreement and moral language, both thought to be relevant to the longstanding debate over moral objectivity. Section 5 considers some potential objections to experimental moral philosophy.

2. Moral Judgments and Intuitions

As mentioned above, ethics encompasses not only moral judgment but also moral perception, behavior, feeling, deliberation, reasoning, and so on. Experimentalists have investigated moral intuition, moral judgments, moral emotions, and moral behaviors, among other things. The most thoroughly investigated are moral intuitions, discussed below.

2.1 Two “Negative” Programs

One project for the experimental ethics of moral judgment, associated with Stephen Stich and Jonathan Weinberg, is to determine the extent to which various moral intuitions are shared by philosophers and ordinary people. As some experimental philosophers are fond of pointing out, many philosophers appeal to intuitions as evidence or use their content as premises in arguments. They often say such things as, “as everyone would agree, p,” “the common man thinks that p,” or simply, “intuitively, p.” But would everyone agree that p? Does the common man think that p? Is p genuinely intuitive? These are empirical questions, and if the early results documented by experimental philosophers survive attempts to replicate, the answer would sometimes seem to be negative.

Still, even if we suppose that these results hold up, the philosophical implications remain in dispute, with some thinking that they would seriously undermine our reliance on moral (or metaethical) intuitions, and others thinking that they wouldn't matter much at all. At least this much should be relatively uncontroversial: If the claim “intuitively p” is meant to be evidence for p, then philosophers who make such claims should tread carefully, especially when it matters whether the philosopher's intuition is as widely shared as the philosopher believes it to be.

A second “negative” project involves examining the degree to which moral intuitions are influenced by factors that are widely agreed to be non-evidential. If the experimentalists conducting this research are right, women sometimes find p intuitive, whereas men find ~p intuitive (Buckwalter & Stich forthcoming); Westerners mostly agree that q, but East Asians tend to think ~q (Machery, Mallon, Nichols, & Stich 2004); and people find r plausible if they're asked about s first, but not otherwise (Nadelhoffer & Feltz 2008; Sinnott-Armstrong 2008d; Sinnott-Armstrong, Mallon, McCoy, & Hull 2008). Once again, the philosophical implications of this growing body of evidence are hotly disputed, with some arguing for the unreliability of moral intuitions, and, to the extent that moral judgments are a function of moral intuitions, those as well. Walter Sinnott-Armstrong (2008d) and Eric Schwitzgebel and Fiery Cushman (2012) 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, if they can be replicated and adequately explained, might again warrant skepticism about moral intuitions, or at least about some classes of intuitions or intuiters.\[^9\]

2.2 Three “Positive” Programs

Other philosophers are more sanguine about the upshot of experimental investigations of moral judgment and intuition. Joshua Knobe, among others, attempts to use experimental investigations of the determinants of moral judgments to identify the contours of philosophically interesting concepts and the mechanisms or processes that underlie moral judgment. He has famously argued for the pervasive influence of moral considerations throughout folk psychological concepts (2009, 2010; see also Pettit & Knobe 2009), claiming, among other things, that the concept of an intentional action is sensitive to the foreseeable evaluative valence of the consequences of that action (2003, 2004b, 2006).\[^{10}\] Such claims remain extremely controversial at present, with some philosophers strongly inclined to treat such differences as reflecting performance errors, widespread though they may be.

Another line of research, associated with Joshua Greene and his colleagues (2001, 2004, 2008), argues for a dual-system model of moral judgment. On their view (very crudely), a slower, more deliberative, system tends to issue in a person's utilitarian-like judgments, whereas a quicker, more automatic system tends to produce judgments more in line with a Kantian approach. Which system is engaged by a given moral reasoning task is determined in part by personal style and in part by situational factors.\[^{11}\] Still, even if something like this story is correct, it is not clear that it reflects in any way on the moral dispute between Kantians and Utilitarians.

A related approach, favored by Chandra Sripada (2011), aims to identify the features to which intuitions about philosophically important concepts are sensitive. Sripada thinks that the proper role of experimental investigations of moral intuitions is not to identify the mechanisms underlying moral intuitions. Such knowledge, it is claimed, contributes little of relevance to philosophical theorizing. It is rather to investigate, on a case by case basis, the features to which people are responding when they have such intuitions. On this view, people (philosophers included) can readily identify whether they have a given intuition, but not why they have it. An example from the debate over determinism (the view that all our actions are the products of prior causes) and free will: “manipulation cases” have been thought to undermine compatibilist intuitions—intuition supporting the notion that determinism is compatible with “the sort of free will required for moral responsibility” (Pereboom 2001). In such cases, an unwitting victim is described as having been surreptitiously manipulated into having and reflectively endorsing a motivation to j. Critics of compatibilism say that such cases satisfy compatibilist criteria for moral responsibility, and yet, intuitively, the actors are not morally responsible (Pereboom 2001). Sripada (2011) makes a strong case, however, through both mediation analysis and structural equation modeling, that to the extent that people feel the manipulees not to be morally responsible, they do so because they judge him them in fact not to satisfy the compatibilist criteria.\[^{12}\] Thus, by determining which aspects of the case philosophical intuitions are responding to, it might be possible to resolve otherwise intractable questions. Whether or not that is so is unsettled for now.

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).[13] 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, and chooses 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 entities, the arrows represent causal or functional processes, and the area in grey represents the mind of the participant, which is not directly observable but is the target of investigation. In broad strokes, the idea is that a participant first reads the text of the vignette and forms a mental model of what happens in the story. On the basis of this model (and almost certainly while the vignette is still being read), the participant begins to interpret, i.e., to make both descriptive and normative judgments about the scenario, especially about the mental states and character traits of the people in it. The participant then reads the experimenter’s question, forms a mental model of what is being asked, and—based on her judgments about the scenario—forms an answer to that question. That answer may then be pragmatically revised (to avoid unwanted implications, to bring it more into accord with what the participant thinks the experimenter wants to hear, etc.) and is finally
recorded as an explicit response to a statement about the protagonist's attitudes (e.g., “he brought about the side effect intentionally,” graded on a Likert scale.\[14\]

What we know is that vignette texts in which a norm violation is described tend to produce higher indications of agreement on the Likert scale responses. What experimental philosophers try to do is to explain this asymmetry by postulating models of the unobservable entities.

Perhaps the best known is Knobe's conceptual competence model, according to which the asymmetry arises at the judgment stage. He claims that normative judgments about the action influence otherwise descriptive judgments about whether it was intentional (or desired, or expected, etc.), and that, moreover, this input is 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.\[15\]

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 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, the model claims 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 the model, when assessing whether someone intentionally brings about some state of affairs, people determine (typically unconsciously) whether there exists sufficient concordance between 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. (Whether that distinction in turn is defensible, and how good people are at recognizing others' deep selves, are further questions.)

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.

According to the deliberation model (Alfano, Beebe, & Robinson 2012; Robinson, Stey, & Alfano forthcoming; Scaife & Webber forthcoming),
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 model 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 and leads to the attribution asymmetry. On the deliberation model, then, the attribution asymmetry originates much earlier than other experimental philosophers suppose, and is due to rational processes.

Single-factor models such as these are not the only way of explaining the attribution asymmetry. Mark Phelan and Hagop Sarkissian (2009, p. 179) find the idea of localizing the source of the asymmetry in a single stage or variable implausible, claiming that "attempts to account for the Knobe effect by recourse to only one or two variables, though instructive, are incomplete and overreaching in their ambition." While they do not propose a more complicated model, it's clear that many could be generated by permuting the existing single-factor models.

2.4 Another Example: The Linguistic Analogy


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 produce 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 produce 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.
Pair 1 suggests the “poverty of the stimulus” argument, according to which there must be an innate language (morality) faculty because it would otherwise be next to impossible for children to learn what and as they do. However, as Prinz (2008) points out, the moral stimulus may be less penurious than the linguistic stimulus: children are typically punished for moral violations, whereas their grammatical violations are often ignored. Nichols, Kumar, & Lopez (unpublished manuscript) lend support to Prinz’s contention with a series of Bayesian moral-norm learning experiments.

Pair 2 suggests the “principles and parameters” approach, according to which, though the exact content of linguistic (moral) rules is not innate, there are innate rule-schemas, the parameters of which may take only a few values. The role of environmental factors is to set these parameters. For instance, the linguistic environment determines whether the child learns a language in which noun phrases precede verb phrases or vice versa. Similarly, say proponents of the analogy, there may be a moral rule-schema according to which members of group G may not be intentionally harmed unless $p$, and the moral environment sets the values of $G$ and $p$. As with the first point of analogy, philosophers such as Prinz (2008) find this comparison dubious. Whereas linguistic parameters typically take just one of two or three values, the moral parameters mentioned above can take indefinitely many values and seem to admit of diverse exceptions.

Pair 3 suggests that people have knowledge of language (morality) that is not consciously accessed (and may even be inaccessible to consciousness) but implicitly represented, such that they produce judgments of grammatical (moral) permissibility and impermissibility that far outstrip their own capacities to reflectively identify, explain, or justify. One potential explanation of this gap is that there is a “module” for language (morality) that has proprietary information and processing capacities. Only the outputs of these capacities are consciously accessible. Whether the module is universal or varies across certain groups of people is a matter currently in dispute.

Pair 4 suggests the linguistic (moral) essentiality of recursion, which allows the embedding of type-identical structures within one another to generate further structures of the same type. For instance, phrases can be embedded in other phrases to form more complex phrases:

- the calico cat $\rightarrow$ the calico cat (that the dog chased) $\rightarrow$ the calico cat (that the dog [that the breeding conglomerate wanted] chased)
- the calico cat (that the dog [that the company {that was bankrupt} owned] chased)

Moral judgments, likewise, can be embedded in other moral judgments to produce novel moral judgments (Harman 2008, p. 346). For instance:

- 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 $x$

Such moral embedding has been experimentally investigated by John Mikhail (2011, pp. 43–8), who argues on the basis of experiments using variants on the “trolley problem” (Foot 1978) that moral judgments are generated by imposing a deontic structure on one’s representation of the causal and evaluative features of the action under consideration.

As with any analogy, there are points of disanalogy between language and morality. At least sometimes, moral judgments are corrigible in the face of argument, whereas grammaticality judgments seem to be less corrigible. People are often tempted to act contrary to their moral judgments, but not to their grammaticality judgments. Recursive embedding seems to be able to generate all of language, whereas recursive embedding may only be applicable to deontic judgments about actions, and not, for instance, to judgments about norms, institutions, situations, and character traits.
Indeed, it's hard to imagine what recursion would mean for character
traits: does it make sense to think of honesty being embedded in courage
to generate a new trait? If it does, what would that trait be?

3. Character, Wellbeing, and Emotion

Until the 1950s, modern moral philosophy had largely focused on either
either utilitarianism 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 this
area.

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
people things they 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 MacIntyre 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, p. 199). Julia
Annas (2011, pp. 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.

The philosophical situationists John Doris (1998, 2002) and Gilbert
Harman (1999, 2000a, 2003) were the first to mount an empirical
challenge to the virtue ethical conception of character, arguing on the basis
of evidence from personality and social psychology that the structure of
most people's dispositions does not match the structure of virtues (or
vices). Philosophical situationists contend that the social psychology of the
last century shows that most people are surprisingly susceptible to
seemingly trivial and normatively irrelevant situational influences, such as
mood elevators (Isen, Clark, & Schwartz 1976; Isen, Shalker, Clark, &
Karp 1978; Isen 1987), mood depressors (Apsler 1975; Carlsmit & Gross
1968; Regan 1971; Weyant 1978), presence of bystanders (Latané &
Darley 1968, 1970; Latané & Rodin 1969; Latané & Nida 1981; Schwartz
& Gottlieb 1991), ambient sounds (Matthews & Cannon 1975; Boles &
Haywood 1978; Donnerstein & Wilson 1976), ambient smells (Baron
1997; Baron & Thomley 1994), and ambient light levels (Zhong, Bohns,
& Gino 2010). Individual difference variables typically explain less
than 10% of the variance in people's behavior (Mischel 1968)—though, as
Funder & Ozer (1983) point out, situational factors typically explain less
than 16%.[16] 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 not because they possess the trait in question, but because they think they do or because they know that other people think they do. He calls such simulations of moral character factitious virtues, and even suggests that the notion of a virtue should be revised to include reflexive and social expectations.[18]

It might seem that this criticism misses its 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 (Badhwar 2009, Kupperman 2009) have conceded that virtue is extremely rare, but argued that it may still be a useful regulative ideal. Others (Hurka 2006, Merritt 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.[19]

### 3.2 Wellbeing

The study of wellbeing and happiness has recently come into vogue in both psychology (Kahneman, Diener, & Schwartz 2003; Seligman 2011) and philosophy (Haybron 2008), including experimental philosophy (Braddock 2010; Phillips, Misenheimer, & Knobe 2011; Phillips, Nyholm, & Liao forthcoming). It is helpful in this context to draw a pair of distinctions, even if those distinctions end up getting blurred by further investigation. First, we need to distinguish between a life that goes well for the one who lives it and a morally good life. It could turn out that these are extensionally identical, or that one is a necessary condition for the other, but at first blush they appear to involve different concepts. 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. As with the first distinction, it might turn out that these are extensionally identical, or that one is a necessary condition for the other, but at first blush they appear to involve different concepts. 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. As with the first distinction, it might turn out that a hedonically good life just is an overall good life, but that would be a discovery, not something we can simply take for granted.

With these distinctions in hand, there are a number of interesting...
experimental results to consider. 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. Arguably, this matters for any philosophical theory that treats pleasure as at least partially constitutive of wellbeing, and it matters a lot if you're inclined, like Bentham (1789/1961) and Singer (Singer & de Lazari-Radek forthcoming), to think that pleasure is the only intrinsic good and pain the only intrinsic ill. How exactly? If the data are to be trusted, philosophers' substantive claims about wellbeing based on hedonic intuitions might be thought to miss their mark because such claims are virtually all formulated in light of ex post evaluation of pleasures and pains. When a philosopher says that a good life is, to the extent possible, filled with pleasure and devoid of pain, and therefore that this way of life constitutes wellbeing, she may be thinking of pleasures and pains in a biased way.

A second interesting set of results has to do with (reports of) hedonic tone but (reports of) subjective wellbeing. The most prominent researcher in this field is Ed Diener,[20] 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.” It might be thought that these questions are more revelatory of wellbeing than questions about hedonic tone. However, two problems have arisen with respect to this data. 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. How could people be so confused about themselves? The questions in Diener's survey are hard to answer. What may happen is that respondents use heuristics to generate their responses, sometimes answering a different but related question than the one asked. 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 low or zero, but if the participant is asked the dating question first, the correlation tends to be high and positive (Strack, Martin, & Schwarz 1988).[21]

Another striking result in the literature on subjective wellbeing has to do with what has come to be known as the set-point. Some early studies suggested that, though major life events such as winning the lottery or suffering a severe spinal cord injury have an effect on subjective wellbeing, the effect wears off over time. People return to a set-point of (reported) subjective wellbeing (Brickman & Campbell 1971). The explanation offered for this striking phenomenon is that people adapt to more or less whatever life throws at them. If the set-point hypothesis is correct (and if reported subjective wellbeing is a reliable indicator of actual wellbeing), then moral theories focused on the promotion of wellbeing would undoubtedly make vastly different recommendations than they are now thought to. There may be little point in trying to promote something that is likely to return to a set-point in the end. More
recent research, however, has challenged the set-point finding by establishing the longitudinal impact of at least some important life events, such as divorce, death of a spouse, unemployment, and disability (Lucas 2007; Lucas, Clark, Georgellis, & Diener 2003).

One final interesting set of results centers on the idea of virtue as a pre-requisite for wellbeing. If the results of Braddock (2010), Phillips, Misenheimer, & Knobe (2011), and Phillips, Nyholm, & Liao (forthcoming) withstand scrutiny and replication, it would seem that ordinary people are willing to judge a life as good for the one living it only if it is both full of positive moods and affects, and is virtuous (or at least not vicious). This result resonates with the empirically-supported views of Seligman (2011) that happiness contingently turns out to be best-achievable in a life that includes both a good deal of positive emotion and the exercise of both moral and intellectual virtues.

3.3 Emotion and Affect

Experimental inquiries into morality and emotion overlap in myriad, distantly-related ways, only a few of which can be discussed here. One especially interesting application is based on what have come to be known as dual-system models of cognition, reasoning, decision-making, and behavior. While the exact details of the two systems vary from author to author, the basic distinction is between what Daniel Kahneman calls System 1, which is fast, automatic, effortless, potentially unconscious, often affect-laden, and sometimes incorrigible, and System 2, which is slow, deliberative, effortful, typically conscious, and associated with the subjective experience of agency, choice, and concentration (2011, pp. 20–21). Whereas System 2 exhibits a degree of functional unity, System 1 is better conceived as a loose conglomeration of semi-autonomous dispositions, states, and processes, which can conflict not only with System 2 but also with each other.

If correct, this might well bear on philosophical issues. For example, is emotionally-driven reasoning in general better or worse than “cold,” affectless reasoning? Likewise, we want to know whether moral judgments are always motivating. In other words, we want to know whether, insofar as one judges that $x$ is morally right (wrong), it follows that one is—perhaps defeasibly—motivated to $x$ (avoid $x$-ing). An affirmative answer is often labeled “internalist,” whereas a negative answer is labeled “externalist.” Emotions are intrinsically motivational, so if experimental investigation could show that emotion was implicated in all moral judgments, that could be a point in favor of internalism.[23] On the other hand, recent empirical/philosophical work on the phenomenon of sociopathy is sometimes thought to cast doubt on this sort of motivational internalism.[24]

The dual-system approach has been employed by various investigators, including Joshua Greene (2008, 2012), Jonathan Haidt (2012; Haidt & Björklund 2008), Joshua Knobe (Inbar et al. 2009), Fiery Cushman (Cushman & Greene forthcoming), and Daniel Kelly (2011). These researchers tend to claim that, though people usually associate their “true selves” with System 2, System 1 seems to be responsible for most intuitions, judgments, and behaviors. Haidt in particular has argued for a “social intuitionist” model according to which deliberation and reasoning are almost always post hoc rationalizations of System 1 cognition, affect, and behavior.

One process that relies heavily on System 1 is disgust. This emotion, which seems to be unique to human animals, involves characteristic bodily, affective, motivational, evaluative, and cognitive patterns. For instance, someone who feels disgusted almost always makes a gaping facial expression, withdraws slightly from the object of disgust, has a slight reduction in body temperature and heart rate, and feels a sense of nausea and the need to cleanse herself. In addition, she is 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, is inclined to treat anything that the object comes in contact with (whether physically or symbolically) as also disgusting, and is more inclined to make harsh moral judgments—both about the object and in general—when confronted with the object experienced as disgusting. The disgust reaction is nearly impossible to repress, is easily recognized, and—when recognized—empathically induces disgust in the other person.\[^{25}\] 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, including in some better-studied cases cuisines, sexual behaviors, out-group members, and violations of social norms.

In a recent book, Kelly (2011) persuasively argues that this seemingly bizarre combination of features is best explained by two theses. 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. According to the entanglement thesis (chapter 2), these initially distinct System 1 responses became entangled in the course of human evolution and now systematically co-occur. If you make the gape face, whatever you're attending to will start to look contaminated; if something disgusts you at a cognitive level, you will flash a quick gape face. According to the co-opt thesis (chapter 4), the entangled emotional system for disgust was 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. Because the disgust reaction is both on a “hair trigger” (it acquires new cues extremely easily and empathically, p. 51) and “ballistic” (once set in motion, it is nearly impossible to halt or reverse, p. 72), it was ripe to be co-opted in this way.

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” (p. 139), according to which the combination of disgust's hair trigger and its ballistic trajectory mean that it is extremely 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” (p. 148).

Furthermore, many theories of value incorporate a link between emotions and value. According to fitting-attitude theories (Rönnow-Rasmussen 2011 is a recent example), something is bad if and only if there is reason to take a con-attitude (e.g., dislike, aversion, anger, hatred, disgust, contempt) towards it, and good if and only if there is reason to take a pro-attitude (e.g., liking, love, respect, pride, awe, gratitude) towards it. According to response-dependence theories (Prinz 2007), something is bad (good) just in case one would, after reflection and deliberation, hold a con-attitude (pro-attitude) towards it. According to desire-satisfaction theories of wellbeing (Heathwood 2006 is a recent example), one's life is going well to the extent that the events towards which one harbors pro-attitudes occur, and those towards which one harbors con-attitudes do not occur. If Kelly's disgust skepticism is on the right track, it looks like it would be a mistake to lump together all con-attitudes. Perhaps it still makes sense to connect other con-attitudes, such as indignation, with moral (non-moral) badness, but it seems unwarranted to connect disgust with moral (nonmoral) badness, at least directly.\[^{26}\] Thus, experimental moral philosophy of the emotions leads to a potential insight into the evaluative diversity of con-attitudes.
Another potential upshot of the experimental research derives from the fact that disgust belongs firmly in System 1: it is fast, automatic, effortless, potentially unconscious, affect-laden, and nearly incorrigible. Moreover, while it is exceedingly easy to acquire new disgust triggers whether you want to or not, there is no documented way to de-acquire them, even if one wants to.[27] Together, these points raise worries about moral responsibility. It's a widely accepted platitude that the less control one has over one's behavior, the more responsible one is for that behavior. At one extreme, if one totally lacks control, many would say that one is not responsible for what one does. Imagine an individual who acts badly because he is disgusted: he gapes when he sees two men kissing, even though he reflectively rejects homophobia; the men see this gape and, understandably, feel hurt. Would it be appropriate for them to take up a Strawsonian (1962) reactive attitude towards him, such as indignation? Would it be appropriate for him to feel a correlative attitude towards himself, such as guilt or shame? If his flash of disgust had instead been something that he recognized and endorsed, the answers to these questions might be simpler, but what are we to say about the case where someone is, as it were, stuck with a disgust trigger that he would rather be rid of? We do not attempt to answer this question here, but instead aim to show that, while experimental moral philosophy of the emotions may provide new insights, it also raises thorny questions.

4. Some Metaethical Issues

Metaethics steps back from ethics to ask about the nature and function of morality. Much contemporary metaethics relies on assumptions about the nature of moral thought in the actual world. Experimental moral philosophy offers to help us make progress on some of these questions by securing a firm empirical foundation for the relevant claims or debunking them where necessary.

Traditional metaethics can be seen to focus on at least four sets of issues, the relations among them, and a number of other issues in the neighborhood. The issues arise in the areas of 1) Moral Metaphysics, 2) Moral Semantics, 3) Moral Reasons, and 4) Moral Epistemology.[28] We've mentioned ways in which experimental moral philosophy can be brought to bear on questions in moral epistemology (critiques of moral intuitions) as well as questions about moral reasons (the debate over internalism).[29] Here we focus on some of the most important applications of experimental moral philosophy thus far to questions about moral metaphysics and moral language.[30]

4.1 Moral Disagreement

The history of philosophical and social scientific approaches to moral disagreement provides a striking illustration of the way science in general can and cannot properly be brought to bear on a question about metaethics. For many years, students of human nature have been fascinated with moral differences. Herodotus chronicled a series of bizarre practices many of which his readers would have found immoral, and philosophers like Sextus Empiricus attempted to draw philosophical inferences (in Sextus’s case, skeptical ones) from such reports. Enlightenment thinkers like Hume, Locke, and Montaigne discussed cases of apparent moral disagreement as well. But the most serious investigations of moral differences came with the birth of cultural anthropology as a distinct discipline around the beginning of the 20th Century, especially under the influence of Finnish philosopher, sociologist, and anthropologist Edward Westermarck; American sociologist William Graham Sumner; and American anthropologist Franz Boas, along with his students including, Ruth Benedict, Melville Herskovits, and Margaret Mead.

These 20th Century thinkers for the most part advocated what came to be
known as cultural, ethical, or moral relativism, very roughly the view that the moral truth depends on the moral beliefs of various groups or individuals. Despite the empirical credentials of these approaches, the boundaries and distinctions among the variously named theories were often vague or imprecise, and the connections between them and the empirical data were not always very strong or even clear, despite a genuine debate among the social scientists employing them. By the middle of the century, philosophers had begun to engage the issues, taking a much more cautious and nuanced approach to them. Two in particular, Richard Brandt (1954) and John Ladd (1957), did their own significant philosophico-anthropological fieldwork among Hopi and Navaho Native Americans communities, respectively.

Many would credit Brandt and Ladd as godfathers of the contemporary experimental moral philosophy approach (and Hume and Nietzsche as its great-godfathers), though it would not pick up steam for decades to come. Alasdair MacIntyre's A Short History of Ethics (1998), while not based on experimental research, arguably bears mention as a work in which the author attempted to treat the history of ethics from a perspective that combined philosophical discussions with an informed perspective on history proper, though it was treated by some critics as endorsing moral relativism (or as downright incoherent). MacIntyre's subsequent work, along with that of others such as Martha Nussbaum (2001), continued this trend to view the history of morality through the twin lenses of philosophical work and empirical scholarship on history proper. J. L. Mackie drew an anti-realist (he called it skeptical) conclusion about morality based on abductive inference from vaguely-referenced anthropological knowledge with a brief but influential discussion of “the argument from relativity” in Ethics: Inventing Right and Wrong (1977). A few years later, naturalist moral realist Nicholas Sturgeon began his seminal paper, “Moral Explanations,” (1988) by affirming his respect for the argument from moral disagreement and his recognition that it could be resolved only through “piecemeal” and “frustratingly indecisive” a posteriori inquiry, sentiments echoed over a decade later by anti-realist Don Loeb.

Loeb (1998) offered a version of the argument that proceeded by way of an inference from irresolvable moral disagreement (what Brandt called fundamental disagreement) to the claim that we lack moral knowledge. If our apparent knowledge of morality is our best (or even our only) reason for believing morality to be a realm of fact, than an argument for moral skepticism is also indirectly an argument for moral anti-realism, he claimed, since it is not epistemically responsible to believe in things for which we lack good evidence. The question, he thought, was how much moral disagreement (and agreement, for that matter) would remain if various explanatory alternatives were exhausted.

In an important paper, John Doris and Alexandra Plakias (2008) turned their attention to the question Sturgeon and Loeb had posed, referring to the alternative explanatory strategies mentioned by Loeb and others as defusing explanations. Doris and Plakias considered an experiment suggesting that a so-called “culture of honor” is much more prevalent in the southern United States than in the north, and conducted a study tending to support the claim that East Asians react differently to “magistrate and the mob” scenarios (in which a person responsible for keeping order can do so only by framing a friendless but innocent man) than do Westerners. In both cases, they argued, no good defusing explanation for the moral disagreement seems available and thus this evidence seems to support moral anti-realism. Though the paper is a splendid example of experimental moral philosophy, bringing social science and philosophy together in addressing a central question in metaethics, its authors recognize that these two cases are suggestive but by no means dispositive and that much more would need to be done to resolve the kinds of questions they are asking, if the questions are indeed
While a few efforts at continuing and refining this research program have begun (Fraser and Hauser, 2010), it is nowhere near maturity, and philosophical criticisms of the entire project have been leveled. For example, Andrew Sneddon (2009 argues that disagreements among ordinary people—those who are not “moral experts”—are completely irrelevant to metaethics because non-experts lack relevant information (about normative theory and experimental psychology, for example) and we do not know whether these people's disagreements would persist if they were better informed.

4.2 Moral Language

One promising area for further research concerns the longstanding debate over moral language. In particular, cognitivists (or descriptivists) hold that 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). Non-cognitivists (including their contemporary representatives, the expressivists) hold that moral language has a fundamentally different function, to express attitudes or to issue commands, for example. All sides agree that a good understanding of moral language will almost certainly contain an element of reform and will not capture every idiosyncratic use of that language. But it is also agreed that reform can go too far, so far as to change the subject. A proper understanding of moral language must be grounded in the moral practices and intentions of actual speakers, though not, perhaps, in their beliefs about their practices and intentions.

Understanding moral language might be relevant to resolving questions about moral objectivity and the nature of morality. If moral language is not fact-asserting (or is incoherently both fact-asserting and something incompatible with that) then there is, in a way, nothing to be a realist about. Don Loeb (2008) calls this sort of argument “pulling a metaphysical rabbit out of a semantic hat.” In this spirit, Loeb has recently argued that both sides may have claim to part of the truth, suggesting that the best analysis of the terms in the moral vocabulary (and the concepts for which they stand) might well be incoherent, combining “non-negotiable” but inconsistent elements. Loeb notes that on virtually any reasonable theory of language, meaning and reference are at least partly determined by the semantic intentions and commitments of ordinary speakers, suitably adjusted to account for performance errors and other deviations. Thus, the issue is one requiring sophisticated and philosophically informed social science, directed at determining the semantic intentions and commitments of ordinary moral thinkers, not armchair reasoning in which only two possibilities are considered.

Yet, remarkably, the debate over moral language has been carried on for almost a century in a kind of empirical vacuum. Instead of gathering scientific evidence, philosophers have relied almost exclusively on their own linguistic and substantive intuitions and on the assumption that they already possess a good understanding of ordinary moral thought and talk.

This makes moral language ripe for treatment by experimental moral philosophers and their colleagues in linguistics and psychology. To date there has been little research focused directly on these questions by philosophers, as opposed to semanticists not pursuing philosophical questions. No doubt scientific inquiry in this area is fraught with difficulty and will require collaboration among philosophers and social scientists. But that does not make it any less important that it be conducted, nor make it any more reasonable to treat seat-of-the-pants judgments by philosophers as adequate.

Fortunately, a new line of research by philosophically inclined psychologists has emerged, and it seems highly relevant to these
questions. If ordinary language users have an implicit metaethical commitment, it is likely that this commitment would be reflected in the meaning of the words in their moral vocabularies. Thus, one way to discover what the words in the moral vocabulary are used to do—make factual assertions or something else—is to uncover people's implicit metaethical commitments. A promising line of research, beginning with Darley and Goodwin's important paper (2008), and continued by James Beebe (forthcoming) and others (Wright et al. 2013; Campbell & Kumar 2012; Goodwin & Darley 2010; and Sarkissian et al. 2011) aims to shed light on these very questions. So far, the results are quite mixed. People seem to treat some moral questions as questions of fact, and others as not matters of fact. Whether there is any principled way to explain these variations, and whether that would shed light on the debate over moral language, remains to be seen.

5. Criticisms of Experimental Moral Philosophy

Experimental moral philosophy far outstrips what we've 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. 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.[38] 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.[39] 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 at least .05), for every published, unexpected result there may be any number of unpublished, unexpected non-results. Thus, unexpected, published results are more likely than might appear to be false positives.

The best way to tell 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). 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.\[40\] Experimental philosophers have followed suit with their own replication project.\[41\]

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 (Alfano & Robinson forthcoming).

Another worry is that simultaneously testing many intuition-probes leads to overconfidence. If one were to test two random samples of people with several intuition-probes, the odds are that one would find some differences in their intuitions—even some statistically significant differences. We saw that a false-positive rate of 5% is often deemed acceptable, and that means that in as many as one out of twenty cases, the test indicates that the pattern of responses did not come from chance when in fact it did. The probability of at least one false positive when making multiple comparisons rises exponentially: \( P(\text{false positive}) = 1 - 0.95^n \), where \( n \) is the number of hypotheses tested simultaneously. This is known as the experimentwise error rate.\[42\]

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.

One final problem to note with experimental design is that ordinal scales are often treated as if they were cardinal scales. The widely used Likert scales, for example, are ordinal: the difference between a response of 0 and a response of 2 cannot be assumed to be the same “size” as the difference between a response of 2 and a response of 4. Only if a scale is cardinal can such an assumption be made (indeed, that's what it means for a scale to be cardinal), yet this mistake is not uncommon.

5.2 Philosophical Problems

One might object, however, that the real problem with experimental philosophy is not that science is hard but that science is irrelevant. “Real” moral philosophy, whatever that is, is in principle unaffected by empirical or experimental results.

One might base the argument for the irrelevance of experimental moral philosophy on the is-ought dichotomy. Whatever science turns up about what is, the argument would go, it cannot turn up anything about what ought to be. Arguably, however experimental evidence can help—at a minimum—to establish what is possible, necessary, and impossible.\[43\] If it can do that, it can at least constrain normative views, assuming the ought-implies-can principle to be true. What is not possible is not morally required, for example. Furthermore, if there are thick moral properties such as character, virtue, or wellbeing then it is possible to investigate at least some normative issues empirically—by investigating these thick properties, just as experimental moral philosophers have been doing.

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—that 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 principle are, if true at all, necessarily true, and that empirical research can establish at best only contingent truths.[44] 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


Fraser, B. and Hauser, M (2010), “The Argument from Disagreement and the Role of Cross-Cultural Empirical Data,” Mind and Language,


[Is this officially forthcoming? If so, it needs a publisher listed. If not, it needs to be removed from the bibliography.] D. Kuchler (ed., forthcoming), *Philosophy of Social Science*.


Ladd, J. (1957). *The Structure of a Moral Code: A Philosophical Analysis of Ethical Discourse Applied to the Ethics of the Navaho Indians*. 


*Nis this forthcoming somewhere?* Nichols, S., Kumar, S., & Lopez, T. (unpublished manuscript). Rational learners and non-utilitarian rules.


Cornell University Press.
Sidgwick and Contemporary Ethics.


Other Internet Resources


Related Entries

moral psychology: empirical approaches

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, Christian Miller, Sven Nyholm, Brian Robinson, Chandra Sripada, and Carissa Veliz, several anonymous referees, and the editors of this encyclopedia for helpful comments and suggestions.

Notes to Experimental Moral Philosophy

1. As Quine is said to have put it, “The universe is no university.”
3. John Stuart Mill (1977) had something like this in mind when he wrote of “experiments of living” (1977, 260–261).
4. Statistical power is the probability that a test will reject the null hypothesis (the statement that the conditions of the experiment do not differ) when the null hypothesis is false. It is therefore 1 minus the probability of a Type II error (failure to detect a relationship despite the reality of the relationship). Statistical power increases as the sample size grows.
5. There is a good deal of exciting new work on questions often characterized as matters of public policy, but which have important moral dimensions. See, among others, Sunstein & Thaler (2008), Sunstein (2013), Gigerenzer & Muir Gray (2011), and Conly (2012). Also see the work of behavioral economists such as George Lowenstein (Kriss, Loewenstein, Wang, and Weber, 2011; Cain, Loewenstein, and Moore, 2011).
6. The distinction between moral intuitions and moral judgments is fraught. Here, we treat moral intuitions as moral seemings and moral judgments as moral beliefs.
8. For incisive criticisms of this claim, see Banerjee, Huebner, & Hauser (2010), Sytsma & Livengood (2011), and Lam (2010).
9. Others prominently expressing concern about the bearing of experimental results such as these on philosophers' reliance on moral intuitions include, Kwame Anthony Appiah (2008) and Peter Singer (2005).
10. This and related research is discussed in more detail in subsection 2.3.
11. Fiery Cushman and Liane Young (2009, 2011) have developed an alternative dual-process model for moral (and non-moral reasoning), as
has Daniel Kahneman (2011). For more on dual-system approaches see Section 3.3, below.

12. Mediation analysis attempts to determine whether one variable (the predictor) affects a second variable (the outcome) by influencing a third, mediating variable (Baron & Kenny 1986). Structural equation modeling allows the analyst to assess and compare various models relating predictors, outcomes, mediators, and moderators (Kline 2005).


14. Such scales are named for their inventor, Rensis Likert (1932, pronounced “LICK-urt”). The participant is presented a statement and then asked to agree or disagree with it on a numeric scale. Commonly, scales run from 1 to 7, 1 to 5, −3 to 3, or −2 to 2. Almost always, the endpoints are labeled ‘strongly disagree’ and ‘strongly agree’. Quite often, the midpoint is labeled ‘neither agree nor disagree’. Sometimes other points on the scale are labeled as well.

15. The idea that seemingly predictive and explanatory concepts might also have a normative component is not entirely original with Knobe; Bernard Williams (1985, p. 129) pointed out that virtues and vices have such a dual nature.

16. Owen Flanagan (1991) considered some of the same evidence before Doris and Harman, but he was reluctant to draw the pessimistic conclusions they did about virtue ethics.

17. When it comes to explaining variance in behavior, the basic idea is that the statistical analysis of experimental results yields a correlation between a personality variable (such as extroversion) and a behavioral variable (such as an act of helping). Correlations range from −1 to +1. A correlation of 0 means that the individual variable is of literally no use in predicting the behavioral outcome; a correlation of 1 means that the individual variable is a perfect positive predictor; a correlation of −1 means that the individual variable is a perfect negative predictor. Actual correlations tend to be between −.3 and +.3. The amount of variance explained by a given predictor variable is the square of the correlation between that variable and the behavior in question. So, for instance, if extroversion is correlated with helping behavior at .25, then extroversion explains 6.25% of the variance in helping behavior. Although this is only one, rather simplistic, measure of explanatory power, personality variables do not look better on other measures, such as Cohen's $d$, $h^2$, or partial-$h^2$.

18. Merritt (2000) was the first to suggest that the situationist critique could be handled by offloading some of the responsibility for virtue onto the social environment in something like this way.

19. One might hope that philosophical reflection on ethics would promote moral behavior. Eric Schwitzgebel has recently begun to investigate whether professional ethicists behave better morally than their non-ethicist philosophical peers, and claims that, on most measures, the two groups are indistinguishable (Schwitzgebel 2009; Schwitzgebel & Rust 2010; Schwitzgebel et al. 2011).


21. See Schimmack & Oishi (2005) for a critical reply, which argues that chronically accessible information is a much better predictor of life satisfaction responses than temporarily accessible information, such as how many dates one went on last week.

23. See Nichols (2002; 2004, ch. 5.)

24. See Kennett (2006), Roskies (2003), and Shoemaker (2011). See Strandberg & Björklund (2013) for a different sort of experimentally-motivated argument against internalism. See also Buckwalter & Turri (unpublished manuscript) for an argument that internalism and externalism draw on distinct folk notions of belief. Some doubt the relevance of empirical considerations to the debate over internalism in ethics, arguing that since internalism claims a necessary or conceptual link between moral judgments and motivation, externalism is compatible with any merely contingent connection.

25. See May (forthcoming) for a criticism of these findings, and Kelly (2011, especially chapter 1) for a comprehensive literature review.

26. Fitting attitude theorists needn't deny this. Some experiences of disgust are not fitting attitudes upon which to ground wellbeing, they could say. And similar arguments could perhaps be deployed on behalf of the other metaethical views mentioned above. Even so, experimental moral philosophy could play a useful role, helping us to identify suspect experiences of disgust.

27. There does seem to be some potential for fine-tuning, however (Rozin 2008; Case et al. 2006).

28. 1) Moral Metaphysics: Is morality a realm of fact? 2) Moral Semantics: What is the function of moral language? How are moral terms to be defined? 3) Moral Reasons: Why be moral? What is the connection between moral principles and reasons for action, or between moral judgments and motivations to act in accord with them? What, indeed, are reasons for action and are they all grounded in our desires? 4) Moral Epistemology. How (if at all) is moral knowledge possible? Is there such a thing as good moral reasoning and if so, what are its precepts?

29. These may themselves be relevant to questions in category 1, moral metaphysics. The categories are somewhat vague, admit of a good deal of overlap, and are by no means exhaustive, but they will serve our purpose here.

30. We also leave aside important issues related to the evolution of morality (see, e.g., Trivers, 1971, and Cosmides and Tooby, 2013), itself the basis for an important challenge to moral realist (and other) views. The latter, evolutionary debunking arguments, are discussed in the entries, “Morality and Evolutionary Biology” and “Moral Epistemology”.

31. See Kluckholn (1959) and Benedict (1959).

32. Moral disagreement is still often cited in philosophical discussions of moral relativism. For example, John Cook (1999) and Richard Schweder (2012) take empirical evidence concerning apparent moral disagreement seriously.

33. Other philosophers notably recognizing the relevance of empirical inquiry to the argument from moral disagreement include Michele Moody Adams (1997), Francis Snare (1980 and 1984), and William Tolhurst (1987).

34. For a fuller discussion see the article in this encyclopedia by Stich and Doris, Moral Psychology: Empirical Approaches, Section 6.

35. Ayer (1936), who makes a similar argument, would have eschewed labeling the conclusion of such an argument as *metaphysical*, since he thought the issue of moral objectivity couldn't be articulated in a meaningful way. (Ayer used the word ‘metaphysical’ in a very different
36. This is Stich and Weinberg's (2001) term.

37. Frank Jackson (1998) displays what some might think is a too-cavalier attitude toward empirical investigation of moral language: “I am sometimes asked—in a tone that suggests that the question is a major objection—why, if conceptual analysis is concerned to elucidate what governs our classificatory practice, don’t I advocate doing serious opinion polls on people's responses to various cases? My answer is that I do—when it is necessary. Everyone who presents the Gettier cases to a class of students is doing their own bit of fieldwork, and we all know the answer they get in the vast majority of cases.” (36–37.)

38. The scandal over replication has (rightly or wrongly) assumed such proportions recently that John Doris has taken to calling it, “Repligate”.

39. There is also an ongoing controversy surrounding null-hypothesis significance testing (NHST). In a nutshell, the problem is that a p-value is a conditional probability, but not the conditional probability that one might expect. A p-value is the probability that the result in hand would have been observed given the null hypothesis, i.e., given that nothing interesting is happening (no positive correlations, no negative correlations, no interaction effects, and so on). This is sometimes inverted by sloppy researchers and interpreters, who gloss the p-value as the probability of the null hypothesis given the observation. Symbolically, the difference is between P(observation | null) and P(null | observation). The latter, more desirable, conditional probability can be estimated using Bayesian statistical analysis, but seldom is (and there are controversies surrounding Bayesian analysis, especially the arbitrariness of prior probabilities). For an introduction to these problems, see Abelson (1997), Cohen (1994), and Wagenmakers et al. (2012).


42. In a recent critique of this kind of fallacious statistical thinking, Peter Austin, Muhammad Mamdani, David Juurlink, and Janet Hux (2006) describe statistical arguments purporting to show that Canadian patients' astrological signs were often correlated with their pathologies. For instance, using the same statistical techniques favored by many experimental philosophers one would be led to conclude that Gemini are 30% more likely to be alcoholics (p < 0.02), Scorpios have an 80% higher risk of developing leukemia (p < 0.05), and Virgo women suffer 40% more from excessive vomiting during pregnancy (p < 0.04). These are presumably statistical anomalies, not indicators of genuine health risks.

43. If something is actual, for example, then it is also possible.

44. We are here indebted to Chris Heathwood.

Copyright © 2014 by the authors
Don Loeb and Mark Alfano