Principles and Intuitions in Ethics:
Historical and Contemporary Perspectives*

David O. Brink

This essay situates some recent empirical research on the origin, nature, role, and reliability of moral intuitions against the background of nineteenth-century debates between ethical naturalism and rational intuitionism. The legitimate heir to Millian naturalism is the contemporary method of reflective equilibrium and its defeasible reliance on moral intuitions. Recent doubts about moral intuitions—worries that they reflect the operation of imperfect cognitive heuristics, are resistant to undermining evidence, are subject to framing effects, and are variable—are best addressed by ethical naturalism as part of a broad dialectical equilibrium.

Experimental ethics is a diverse body of research at the intersection of philosophical and empirical moral psychology. One strand concerns the origin, nature, role, and reliability of moral intuitions in moral deliberation and theorizing. This new wave of research on moral intuitions itself has disparate strands, to which it would be impossible to do justice here. However, we can begin to assess some of this empirical work on intuitions for philosophical ethics by situating it in historical context. Toward this end, I will reconstruct in some detail nineteenth-century debates between intuitionists and their naturalistic rivals about the nature and significance of moral intuitions. In particular, I will examine John Stuart Mill’s naturalistic polemic against intuitionism and Henry Sidgwick’s defense of philosophical intuitionism. The legitimate heir to nineteenth-century naturalism, I will argue, is the contemporary method

* I would like to thank the participants and audience at the Experimental Ethics conference at Georgetown University for useful discussion of an earlier version of this material. Special thanks to Matt Braich, John Mikhail, Jesse Prinz, Peter Railton, Henry S. Richardson, and three anonymous reviewers at Ethics for their comments and suggestions.

Ethics 124 (July 2014): 1–30
2014 by The University of Chicago. All rights reserved. 0014-1704/2014/12404-0002$10.00
of reflective equilibrium and its defeasible reliance on moral intuitions. This background will put us in a better position to understand and assess the philosophical significance of some strands in the new research on moral intuitions. It will show that the new intuitionism is very unlike the old intuitionism of the nineteenth century and more like its naturalistic rivals. Moreover, some of the more surprising and potentially unsettling aspects of recent research on intuitions can be accommodated within the sort of methodological framework inspired by these naturalistic alternatives.

I. INTUITIONISM, NATURALISM, AND UTILITARIANISM

The tradition of intuitionism in British moral philosophy dates back at least to the seventeenth century with Ralph Cudworth and Samuel Clarke. It received important articulation in the eighteenth century by Richard Price and Thomas Reid, in the nineteenth century by Dugald Stewart, William Whewell, and Henry Sidgwick, and in the twentieth century by W. D. Ross and H. A. Prichard, among others. The central intuitionist claim, for our purposes, is that moral knowledge rests on moral precepts or principles that we can see, on proper reflection, to be self-evident. Different intuitionists take different principles to be self-evident, but most recognize a plurality of such principles and, in particular, juridical or deontic principles about duty and obligation that are distinct from teleological principles concerning self-love and benevolence. Among the deontic principles recognized by some intuitionists are principles of fidelity, veracity, piety, and gratitude. Moreover, several of the intuitionists claim that these principles are innate or, at least, recognized as self-evident by people employing innate moral faculties. As such, intuitionism traditionally took a rationalist form, asserting that moral knowledge is built a priori on a foundation of general moral precepts recognized by an innate cognitive faculty as self-evident.\(^1\)

We get a better sense of the commitments of traditional intuitionism by looking at its naturalistic rivals. Strictly speaking, what separates intuitionism and naturalism are their epistemological or methodological commitments. Debates about the content of fundamental normative principles would seem orthogonal to the primary debate between intuitionists and naturalists. Nonetheless, with the main exception of Sidgwick (discussed below), the intuitionists were not utilitarians and the naturalists were. Mill was one of the clearest nineteenth-century naturalistic critics of intuitionism, and his utilitarianism informed his naturalism.

---

Mill explains utilitarianism in chapter 2 of *Utilitarianism* in terms of the Proportionality Doctrine: “The creed which accepts as the foundations of morals ‘utility’ or the ‘greatest happiness principle’ holds that actions are right in proportion as they tend to promote happiness; wrong as they tend to produce the reverse of happiness.”\(^2\) It is controversial how to interpret the Proportionality Doctrine. I believe it is correct and, in any case, convenient for present purposes to interpret it as the direct (act) utilitarian claim that an act is right insofar as it promotes utility or happiness. We might expect such a utilitarian to be motivated by pure disinterested benevolence and to deliberate by calculating expected utility on each and every occasion for action. But Mill tells us that utilitarianism is a standard of right conduct and warns us that we should not always use it as a guide.\(^3\) It is a practical question how to reason or be motivated, and direct utilitarianism implies that this practical question, like all practical questions, is correctly answered by what would in fact promote utility. We often lack important information about the causal structure of the world and so are not in a position to make reliable calculations of expected utility. Moreover, such calculations are time consuming and so often have opportunity costs. Furthermore, calculations of expected utility are subject to familiar biases. For instance, we are subject to a temporal bias insofar as we tend to invest temporally proximate benefits and harms with significance out of proportion to their actual magnitude, and we are subject to a personal bias insofar as we tend to inflate the magnitude of benefits and harms to ourselves and those near and dear. For such reasons, we may better approximate the utilitarian standard if we don’t always try to approximate it. Later utilitarians, such as Sidgwick, have echoed this idea that utilitarianism is a standard of duty, not necessarily a guide to action.\(^4\)

But if utilitarianism is itself the standard of right conduct, not a guide, then what should guide our deliberations, and what role should the principle of utility play in moral reasoning? Here, Mill appeals to the need for secondary principles. In chapter 2 of *Utilitarianism* and in the discussion of the Art of Teleology at the end of *A System of Logic*, Mill explains the need for internalizing secondary precepts, concerning such things as honesty, fidelity, and fair play, as a way of satisfying the utilitarian first principle. On this view, secondary principles are generally but imperfectly reliable guides to doing what will maximize happiness.

---


Secondary principles, so understood, might sound like mere rules of thumb. But Mill does not regard them as mere heuristics in a utility calculation. They don’t themselves make reference to utility, and he thinks they should be adhered to uncritically in ordinary circumstances. He goes so far as to describe the rule against lying as “sacred.” He seems to believe that secondary principles, such as the principle against lying, satisfy two conditions.

1. Following the principle generally but imperfectly leads to optimal results.
2. One cannot in general reliably discriminate whether or, if so, when adherence to the principle would produce suboptimal results.

When these two conditions are met, Mill believes, agents should follow these principles automatically and uncritically most of the time. In these cases, agents consult only secondary principles; they do not use them as heuristics in a utility calculation. The principles have genuine deliberative autonomy. But to say this is not to say that agents should never consult the utilitarian first principle or assess the acceptance value of secondary principles. They should periodically step back and review, as best they can, whether the principle continues to satisfy conditions (1) and (2). Also, they should set aside these secondary principles and make direct appeal to the principle of utility in unusual cases in which it is especially clear that the effects of adhering to the principle would be very bad, and not merely marginally suboptimal, and in cases in which secondary principles, each of which has a utilitarian justification, conflict.

Regulating one’s behavior in this way by secondary principles is what will best promote happiness. Mill summarizes this picture in *A System of Logic*.

I do not mean to assert that the promotion of happiness should be itself the end of all actions, or even all rules of action. It is the justification, and ought to be the controller, of all ends, but it is not itself the sole end. There are many virtuous actions, and even virtuous modes of action (though the cases are, I think, less frequent than is often supposed) by which happiness in the particular instance is sacrificed, more pain being produced than pleasure. But conduct of which this can be truly asserted, admits of justification only because it can be shown that on the whole more happiness will

exist in the world, if feelings are cultivated which will make people, in certain cases, regardless of happiness.\textsuperscript{7}

For Mill, secondary principles are an essential supplement to the utilitarian first principle, and they have an important role in ordinary moral reasoning.\textsuperscript{8}

Mill’s utilitarian justification of discrete secondary principles is intended as a contrast with the intuitionism of Whewell and others. In \textit{Utilitarianism} Mill contrasts intuitionism with his own “inductive” or naturalistic approach to matters of justification.\textsuperscript{9} He dismisses any version of intuitionism that would lay claim to a faculty of moral perception that could discern what is right or wrong in particular cases. To understand Mill’s disagreement with the intuitionists, it will be helpful to distinguish three different levels of generality in our moral claims. At the most particular level, our judgments concern particular actions or action tokens. For instance, we might judge this particular act of lying, in these specific circumstances, to be wrong. We often try to subsume and explain such particular truths under more general claims about morally relevant factors that operate at the level of types or classes of actions. For instance, we might judge lying to be \textit{pro tanto} wrong. Such claims could be thought of as mid-level moral principles. Finally, we might try to subsume and explain disparate morally relevant factors under some single or small number of very general principles. If these principles cannot be subsumed and explained in terms of other, more basic principles, then these principles are ultimate or first principles.\textsuperscript{10} Utilitarianism is one such putative first principle, because it purports to ground all mid-level principles in utility.

\textsuperscript{7} Mill, \textit{A System of Logic}, CW 8:951.

\textsuperscript{8} Whether Mill’s claims about the importance of secondary principles imply rule utilitarianism depends on whether he wants to define right action in terms of principles with optimal acceptance value or whether these principles are just a generally reliable way of doing what is in fact best and right. If the former, Mill is a rule utilitarian; if the latter, he is an act utilitarian.

\textsuperscript{9} Mill, \textit{Utilitarianism}, CW 10:206.

\textsuperscript{10} There is, of course, controversy about the existence, nature, and role of moral principles, some of which is reflected in debates between generalists and particularists. For some flavor of these debates, see John McDowell, “Virtue and Reason,” \textit{Monist} 62 (1979): 331–50; Jonathan Dancy, \textit{Ethics without Principles} (Oxford: Clarendon, 2004); Sean McKeever and Michael Ridge, \textit{Principled Ethics: Generalism as a Regulative Ideal} (Oxford: Clarendon, 2006); Pekka Väyrynen, “A Theory of Hedged Moral Principles,” \textit{Oxford Studies in Metaethics} 4 (2009): 91–132; and Luke Robinson, “Moral Principles as Moral Dispositions,” \textit{Philosophical Studies} 156 (2011): 289–309. My own view is that most moral principles should be thought of as statements of \textit{pro tanto} or contributory moral factors (e.g., good-making or right-making factors) that make an invariant contribution to the overall moral status of a situation or act. Most moral situations involve multiple moral factors, and an overall moral verdict depends upon which factors are most important. Moral principles,
We might characterize Mill’s discussion of intuitionism in terms of distinctions Sidgwick would later draw within intuitionism. Sidgwick distinguishes perceptual, dogmatic, and philosophical intuitionism. Perceptual intuitionism locates intuitions at the level of action tokens, dogmatic intuitionism locates intuitions at the intermediate level of action types or mid-level principle, and philosophical intuitionism locates intuitions at the level of first principles. Sidgwick himself rejects both perceptual and dogmatic intuitionism but ultimately defends philosophical intuitionism. Mill dismisses perceptual intuitionism and focuses on, but ultimately rejects, dogmatic intuitionism.

According to Mill, intuitionists and naturalists need not and typically don’t disagree about the content of mid-level principles or moral factors. They disagree about the nature of moral knowledge. It is characteristic of intuitionism to claim that moral principles are knowable a priori, or independently of experience, and can be seen to be self-evident. By contrast, as Mill makes clear in his essay “Whewell on Moral Philosophy,” he thinks that the intuitionist wrongly treats familiar moral precepts as ultimate moral factors whose justification is supposed to be self-evident to reason. By contrast, Mill’s account of secondary principles recognizes their importance in moral reasoning but insists that they are neither innate nor infallible; they are precepts that have been adopted and internalized because of their acceptance value, and their continued use should be suitably regulated by their ongoing comparative acceptance value. Far from undermining utilitarian first principles, Mill thinks, appeal to the importance of such moral principles actually provides support for utilitarianism. He makes this argument in considerable detail in the case of precepts of justice in chapter 5 of Utilitarianism, where he argues that justice is not, as intuitionists allege, a prin-

11. Sidgwick, Methods, 97–102.
ciple independent of utility but is rather a principle and associated set of emotions protecting security and other essentials of happiness, and, hence, justified by their good consequences.\textsuperscript{13}

For Mill, this is a naturalist critique of intuitionism. He contrasts top-down and bottom-up strategies of justification and associates naturalism with a bottom-up strategy in which general or theoretical claims are justified by appeal to their ability to subsume and explain particulars, in particular, by appeal to “observation and experience.”\textsuperscript{14} Another aspect of Mill’s naturalism is his fallibilism. Any given belief about the world might be mistaken, and corrections must come from observation and experience.\textsuperscript{15} We might identify three distinct claims that seem to be part of Mill’s ethical naturalism.

1. \textit{Fallibilism}: Moral beliefs are fallible and so revisable in principle.
2. \textit{Dialectical Method}: Moral beliefs are not self-evident; they are to be justified by their fit with other beliefs we hold, including our beliefs about particulars.
3. \textit{Empiricism}: Moral knowledge is a posteriori insofar as the justification of moral beliefs depends on experience and observation.

Though one might embrace fallibilism and dialectical methods without a commitment to a posteriori justification, Mill’s naturalism embraces all three. He thinks that common moral precepts deserve to be taken seriously but insists that they are fallible. They depend for their justification on their acceptance value, and this is in significant part an empirical matter.

\section*{II. PHILOSOPHICAL INTUITIONISM}

Mill’s critique of intuitionism is a critique of perceptual and dogmatic forms of intuitionism, but not a critique of intuitionism per se, as Sidgwick’s brand of philosophical intuitionism shows. Intuitionism, of course, is one of Sidgwick’s three methods of ethics, along with egoism and utilitarianism. In its broadest sense, intuitionism for Sidgwick is a form of epistemological foundationalism that claims that our moral knowledge rests on self-evident moral beliefs. In \textit{The Methods of Ethics} Sidgwick distinguishes different forms of intuitionism. Though he is critical of extant forms of intuitionism, he ends up defending a form of intuitionism about first principles. Sidgwick claims that whereas prior intuitionists were either perceptual or dogmatic intuitionists, the most defensible form of intui-

\textsuperscript{13} Mill, \textit{Utilitarianism}, \textit{CW} 10:240--41, 255--59.
\textsuperscript{14} Ibid., 206.
\textsuperscript{15} Mill, \textit{A System of Logic}, \textit{CW} 7:568.
tionism is philosophical intuitionism, which treats first principles as the object of genuine intuition.

Sidgwick sees a structure in a system of moral truths in which the more particular moral truths are asymmetrically metaphysically dependent on the more general ones. It is the possession of right-making factors and the truth of associated moral rules or secondary principles that explain why particular claims about right conduct are true, but not vice versa. Similarly, if these secondary principles are not themselves ultimate, then they, in turn, can be subsumed and explained under some more general or abstract principle. Eventually, we must reach moral factors that are genuinely ultimate, which explain why other moral claims are true, but which cannot be subsumed under any more ultimate principle. Hence, it is in the nature of first principles to state ultimate moral factors, on which other moral truths depend but which do not themselves depend on other truths.

Sidgwick thinks that this metaphysical structure has epistemological significance. Moral reflection typically begins with particulars. We begin with a conviction about what ought to be done in a particular case or perhaps with moral perplexity about what to do in a particular case. Reflecting on perplexities about particular cases or perhaps trying to justify some judgment about a particular case, we attempt to subsume and explain disparate particular claims under more general morally relevant factors. Similarly, if we experience perplexity about a putative moral factor or someone challenges our apparently settled convictions about a putative moral factor, we might try to get leverage on that issue by looking to the more general first principles, if any, that might subsume and explain the factor in question. In this way, Sidgwick sees a justificatory ascent in ethics. But, he believes, this justificatory ascent must come to a stop when we reach genuine first principles. We cannot justify them in terms of anything else—otherwise, they wouldn’t be ultimate moral principles. They must be self-evident. Sidgwick concludes that particular moral truths are epistemically asymmetrically dependent on more general truths too. This explains why he is an intuitionist, but a philosophical intuitionist, one who locates rational intuitions at the level of first principles.


17. Sidgwick does claim that self-evident moral beliefs must satisfy four conditions: (1) their content must be clear and precise, (2) their truth must be evident upon reflection, (3) they must be mutually consistent, and (4) they must be capable of sustaining a consensus (Methods, 338–42). Whereas (1) and (2) seem to be part of self-evidence, (3) and (4) seem to be part of discursive evidence, rather than self-evidence.
III. NATURALISM ABOUT FIRST PRINCIPLES

Sidgwick’s philosophical intuitionism invokes a common idea in the history of modern ethics, one that appeals to Bentham and Mill as well, namely, that ultimate or first principles cannot be derived from other claims or principles, because, if they could, they would not be ultimate principles after all. The British utilitarians sometimes refer to the principle of utility as an axiom, and it’s true that other things admit of proof by derivation from axioms but that axioms themselves do not admit of proof within the axiomatic system. Sidgwick concludes that the axioms themselves can only be self-evident.

But this makes it hard to see how we can answer doubts about whether a putative first principle is true or provide reasons for adopting one axiomatic system in preference to another. Fortunately, proof, in this sense, is not essential for justification. We can ask why we should accept some axioms as true and why we should prefer one axiomatic system to another. One naturalistic answer appeals to things we can prove within the system and how that compares with what we believe, independently, to be true. Does the axiomatic system allow us to prove all the things that we believe to be true? Does it entail claims that we don’t believe to be true? Sometimes axiomatic systems have surprising implications. Here, we need to decide if we should accept these implications after all, partly on the strength of all the welcome implications of the system, or whether it is a defect in the system that it has such consequences. Typically, we are trying to decide if we should accept a given axiom or axiomatic system, rather than another. To do so, we want to make comparative assessments of the implications of the alternative axioms or systems. Within the axiomatic system, particular truths are asymmetrically metaphysically dependent on more general ones, including, ultimately, the axioms. But epistemic dependence can be bidirectional.

The situation is similar with ethical principles. We can justify more particular moral claims by tracing them to the first principle(s) on which they depend, but we cannot justify first principles in terms of anything more ultimate—otherwise, they wouldn’t be ultimate moral principles. We can’t ask in virtue of what other morally relevant factor a first principle obtains. But we can ask whether a putative first principle is true, or why we should accept one putative first principle, rather than another. A plausible naturalistic answer is that we should accept those putative first principles as true that subsume and explain more particular judgments that, on reflection, we think are true. No principle may subsume and explain all and only the things we take to be true antecedently, in part because our antecedent beliefs are likely to be incomplete in some respects and inconsistent in others. But we can reasonably prefer to accept a principle or set of principles that subsumes and explains more that we
are prepared on suitable reflection to accept over principles that sub-
sume and explain less of what we are prepared to accept. Metaphysical
dependence may be top-down, but epistemic dependence can be bottom-
up and bidirectional.

This picture of the justification of first principles fits with Mill’s
naturalistic methodological commitments and with much of his argu-
mentative strategy in *Utilitarianism*. The proper interpretation of Mill’s
proof of utilitarianism in chapter 4 is controversial. But on one plausi-
ble reading, Mill’s argument for utilitarianism rests on two independent
assumptions—that concern for someone for her own sake focuses on
her happiness and that the moral point of view should be impartial in
its concern with happiness. He defends this assumption of impartiality
by noting its progressive influence historically in undermining institu-
tions and practices of class and privilege. This appeal to the moral les-
sons of history is a good example of Millian naturalism.

But the proof is just the tip of a methodological iceberg. Bottom-
up justification is really quite pervasive in *Utilitarianism*. In chapter 2 he
considers a number of objections to utilitarianism. Typically, he responds
by defending the implications of utilitarianism and arguing that the ob-
jection rests on a confusion of some kind. But in other cases he concedes
the point and tries to show that utilitarians have the conceptual resources
to accommodate it. In all cases, Mill tries to show that utilitarianism can
accommodate moral and evaluative commitments that reflection and ex-
perience show are worth preserving. In doing so, he is defending utili-
tarianism using naturalistic methods.

Or consider again his polemic with intuitionism. Whereas the intu-
itionist treats the precepts of commonsense morality as ultimate moral
factors whose justification is supposed to be self-evident, Mill treats these
precepts as mid-level generalizations that function as secondary prin-
ciples in relation to the utilitarian first principle. The importance of such
precepts, Mill thinks, supports, rather than undermines, utilitarianism.
As we saw, he makes this argument in the case of precepts of justice, ar-
guing that they protect security and other essentials of happiness and,
hence, are justified by their good consequences.

It will help to understand Mill’s naturalistic commitments to distin-
guish two different relations between putative first principles and com-
monsense moral precepts. Principles might aim to accommodate or re-
form commonsense morality. We have been discussing ways in which Mill
thinks that utilitarianism, properly understood, can accommodate com-

---

18. I make this case in more detail in David O. Brink, *Mill’s Progressive Principles* (Ox-
ford: Clarendon, 2013), chap. 5.
mon moral distinctions and claims. But, like most utilitarians, Mill mixes accommodation and reform. Indeed, utilitarianism could not have been a progressive influence historically, as it was, unless it had contained elements of reform. Reformist demands assume that our existing outlook is in some way incomplete, inconsistent, or incorrect. But reformist demands would have little appeal if they couldn’t be anchored in accommodation. Revisionary claims are acceptable on naturalistic grounds provided they are forced on us by principles that subsume and explain our experience and observations well. In such cases, local revision or reform is required by considerations of global or systematic accommodation.

So Mill employs a naturalistic defense of utilitarianism insofar as he engages in accommodation. Because accommodation is background for principled reform, Mill also employs a naturalistic defense of utilitarian calls for reform. In his polemic with intuitionism, he argues that we should accept the utilitarian first principle because of the way it subsumes, explains, and regulates common moral precepts. We can now see that Mill is making a claim very much like the claim that Sidgwick will later make when he says that commonsense morality is “unconsciously” and “inchoately and imperfectly” utilitarian.\(^{21}\)

It may be shown, I think, that the Utilitarian estimate of the consequences not only supports broadly the current moral rules, but also sustains their generally received limitations and qualifications: that, again, it explains the anomalies in the Morality of Common Sense, which from any other point of view must seem unsatisfactory to the reflective intellect; and moreover, where the current formula is not sufficiently precise for the guidance of conduct, while at the same time difficulties and perplexities arise in the attempt to give it additional precision, the Utilitarian method solves these difficulties and perplexities in general accordance with the vague instincts of Common Sense, and is naturally appealed to for such solution in ordinary moral discussions. It may be shown further, that it not only supports the generally received view of the relative importance of different duties, but is also naturally called in as arbiter, where rules commonly regarded as co-ordinate come into conflict.\(^{22}\)

Mill’s mix of global accommodation and local reform commits him to Sidgwick’s claim about the imperfect coincidence between utilitarianism and commonsense morality.

Mill’s naturalism commits him to thinking that this imperfect coincidence provides a justification of the utilitarian first principle. Though


\(^{22}\) Ibid., 425–26.
Sidgwick recognizes that commonsense morality is inchoately and imperfectly utilitarian, he seems more ambivalent about the probative significance of this fact. On the one hand, Sidgwick does sometimes treat the imperfect coincidence of utilitarianism and commonsense morality as evidence for the truth of utilitarianism: “If systematic reflection upon the Morality of Common Sense thus exhibits the Utilitarian principle as that to which Common Sense naturally appeals for the further development of its system which this same reflection shows to be necessary, the proof of Utilitarianism seems as complete as it can be made.”

However, this naturalistic justification of utilitarianism is inconsistent with Sidgwick’s own defense of philosophical intuitionism, which insists that moral knowledge rests on first principles that must be self-justifying.

IV. ACCOMMODATION, REFORM, AND EQUILIBRIUM

The contemporary heir to Mill’s naturalism about first principles is the method of reflective equilibrium that John Rawls discusses in *A Theory of Justice*. According to Rawls, we test candidate moral principles by seeing how well they cohere with our considered moral judgments about particular cases, both actual and hypothetical. An attractive principle subsumes and explains familiar moral principles and provides a good fit with our considered moral judgments, especially about particular cases. If a principle has counterintuitive implications, that counts against the principle and may be reason to reject or at least modify the principle. But an isolated conflict with our intuitive judgments need not be a decisive objection to a principle. The test here is one of systematic comparative plausibility. If a candidate principle subsumes and explains many of our considered moral judgments, especially if it does so better than rival principles, then that may be a reason to judge the recalcitrant intuition mistaken and revise it. Indeed, since there is no reason to believe that our existing considered moral judgments are complete or consistent, there’s no reason to expect that a principled account of our moral beliefs is available that won’t require some modification of our existing moral outlook. Ideally, we make trade-offs among our principles, considered moral convictions, and other commitments, making adjustments here

and there, as coherence seems to require, until our ethical views are in reflective or dialectical equilibrium.26

Considered moral judgments are presumably a subclass of moral judgments. Moral judgments are considered when they have been made under conditions of reflection conducive to forming true beliefs—they are not rushed, they are fairly consistent and stable over time, they are robust under familiar dialectical pressures of consistency and full information, and they are not subject to any obvious form of bias or distortion. We can call these considered moral judgments intuitions, if we like, as many have, and we can even call the resulting method of reflective or dialectical equilibrium intuitionistic, if we like. But we should not confuse this method with rational intuitionism. Reflective equilibrium does not assume that its regulative precepts are innate, and it denies that they are self-evident. Rather, they are among the starting points in a dialectical investigation, and they are justified by their dialectical fit with other commitments.27

Rawls’s discussion is also helpful, because it reminds us that utilitarianism is but one candidate principle. Mill and Sidgwick endorse the idea that commonsense morality is “unconsciously” and “inchoately and imperfectly” utilitarian. Call this Sidgwick’s thesis. If we treat Sidgwick’s thesis as part of a bottom-up justification of utilitarianism, then we are essentially saying that utilitarianism provides a dialectical fit with commonsense morality. Rawls agrees that if Sidgwick’s thesis were true, that would be a good justification of it. Of course, he in fact rejects Sidgwick’s thesis and defends Justice as Fairness as providing a superior fit with our considered judgments about social justice. So whereas the original naturalists were utilitarians, naturalism does not presuppose utilitarianism.

26. Rawls describes the search for equilibrium as an “attempt to describe our moral capacity” or “characterize our moral sensibility” (A Theory of Justice, 46). This raises the question whether dialectical equilibrium only aspires to uncover the principles that implicitly guide our moral sensibility or whether it also treats these principles as tracking objective moral factors. This corresponds roughly to constructivist and realist interpretations of reflective equilibrium. My sympathies lie with the realist interpretation, in part because I think that the process of seeking equilibrium asks us to square our theories with the deliverances of intuition, the intuiteds, rather than the intuitings. But much of the discussion to follow can remain agnostic on this question of metaethical interpretation.

27. I assume that dialectical equilibrium denies that considered moral beliefs are self-evident. Rawls explicitly says this at A Theory of Justice, 21. Elsewhere I have argued against various conceptions of self-evidence and in favor of a coherentist interpretation of dialectical equilibrium. See David O. Brink, Moral Realism and the Foundations of Ethics (New York: Cambridge University Press, 1989), chap. 5. However, dialectical equilibrium could be given a foundationalist interpretation according to which considered moral beliefs are pro tanto self-evident but their further or complete justification requires showing them to be in equilibrium with other beliefs we hold.
We can also see the interplay of accommodation and reform at work in reflective equilibrium. Insofar as a principle subsumes and explains considered moral judgments, it accommodates our moral intuitions. But insofar as a principle conflicts with considered moral judgments, it calls for reform of our moral intuitions. We should not expect perfect accommodation, because our moral judgments are incomplete, inconsistent, and subject to various kinds of bias and distortion. But complete reform is no more plausible than complete accommodation. Indeed, complete reform threatens to introduce a change in subject matter and should be explored only as a last resort, if we concluded that our considered moral judgments are hopelessly muddled. Typically, reform is like the hole in a doughnut, made possible by a surrounding substance of accommodation. We accept local reforms as part of a process of global accommodation. This means that reform is always, or at least typically, partial. But to say that reform is partial does not imply that it cannot be significant. How revisionary a principle can be is something we cannot decide in advance of looking at the principle, its degree of accommodation, and the nature of its reforms.

How revisionary a dialectical fit or equilibrium can be will depend, in part, on how many and diverse the inputs to the process are. The dialectical fit described so far has been fairly narrow, consisting primarily of considered moral judgments about particular cases and moral factors, candidate principles, and claims about circumstances that allow us to derive implications from the principles for factors or particular cases. Even a narrow equilibrium of this sort can be quite revisionary. But a naturalist should think that the fit or equilibrium we seek should be broad, including more and more diverse inputs. Some additional inputs might include auxiliary moral or other philosophical commitments. Sidgwick thinks that utilitarianism might, in some circumstances, be an “esoteric morality”—known only to a select few, with the majority best left to accept apparently deontological secondary principles. Rawls famously rejects this idea, claiming that moral theories should pass a “publicity test,” according to which acceptable moral principles have to be stable when publicly affirmed and provide the basis for public assessment of actions and institutions. He argues that utilitarianism is unstable in this way and violates the strains of commitment. Relatedly, some have insisted that moral principles pass a test of “psychological

realism,” which requires that moral principles make demands that can be reconciled with actual or potential motivational capacities of agents.\textsuperscript{31} Mill thought that utilitarianism was psychologically realistic, but more recent writers have argued that utilitarianism, and perhaps other impartial moral principles, offend against this realism constraint. Indeed, psychological realism is just one kind of connection between normative ethics and moral psychology. There is a long tradition in moral philosophy of wanting to make moral principles hostage to plausible claims about human nature. Mill exemplifies this tradition when he defends a form of utilitarianism that employs a distinctive conception of happiness reflecting assumptions about our natures as “progressive” beings. Sidgwick and others think that the plausibility of principles of distributive justice, distinct from considerations of utility, may depend on the right theory of personal identity. And, of course, many recognize that some of our beliefs about moral responsibility may be affected by metaphysical debates about the compatibility of freedom, moral responsibility, and determinism. These are just a few additional philosophical commitments and issues that might figure as independent inputs to our dialectical process.

Not all of these additional inputs will be philosophical in nature. For example, sociology, psychology, and cognitive science might well have things to tell us about the contours of a psychological realism constraint. Moreover, social science may also tell us things about the moral judgments that are already input into dialectical fit. For instance, as already noted, we are familiar with at least two kinds of biases—inter-temporal and interpersonal. Many of our judgments are prone to a temporal bias insofar as we tend to invest temporally proximate benefits and harms with significance out of proportion to their actual magnitude. Sometimes, our temporal bias involves a constant discount rate. However, some agents discount hyperbolically, resulting in dynamic preference instability.\textsuperscript{32} In such cases, we judge in a cool hour that the distant benefits offset a short-term sacrifice; but as the sacrifice draws near, we undergo a preference reversal and come to judge that the sacrifice is not worthwhile; and then with hindsight we revert to our earlier cool judgment and experience regret at not persevering for the delayed reward. On most views, both kinds of temporal discounting are

\textsuperscript{31} Psychological realism is a theme in much of Bernard Williams’s moral philosophy. See, e.g., several of the essays collected in Bernard Williams, \textit{Moral Luck} (London: Cambridge University Press, 1981). Also see Owen Flanagan, \textit{Varieties of Moral Personality} (Cambridge, MA: Harvard University Press, 1991); and Scheffler, \textit{Human Morality}.

mistakes. Similarly, our judgments can be subject to a personal bias insofar as we tend to overestimate benefits and costs to ourselves and those near and dear to us and underestimate benefits and costs to comparative strangers. To the extent that some of our moral intuitions reflect the operation of these or other biases, we should be less confident in those intuitions and more prepared to revise them should they conflict with an otherwise attractive principle.

These are ways in which a broad equilibrium can be even more revisionary than a narrow one. Given the dependence of many of our moral beliefs on complex empirical and philosophical issues and the dependence of many of our moral beliefs on biases and blind spots, there is every reason to expect a broad dialectical fit to be revisionary. The test of broad equilibrium must also ultimately be one of systematic comparative plausibility. At the end of the day, we want to know whether to accept one moral theory or claim, rather than another. To determine this, we need to know not only how good its fit is, but how its fit compares with that of other, rival theories or claims.

Of course, it’s an extraordinarily complex and difficult matter to determine when the standard of systematic comparative plausibility is or would be met. Dialectical fit is a regulative ideal that we can only hope to approximate. For the most part, we apply the standard in a manner that is fragmentary and piecemeal, focusing on particular apparent strengths and weaknesses of particular theories and seeing how well the particular strengths and weaknesses survive closer scrutiny. These piecemeal inquiries bear directly on the justificatory prospects of the theories under examination, but they do not themselves, individually, answer questions of systematic comparative plausibility.

V. EXPERIMENTAL ETHICS AND MORAL INTUITIONS

The method of dialectical equilibrium is naturalistic insofar as it holds moral deliberation and theorizing answerable to our defeasible moral convictions about particular cases and includes these sorts of moral intuitions as input into a broad equilibrium that includes other defeasible philosophical and scientific commitments. For this reason, we could say that there is one sense in which the search for dialectical equilibrium is intuitionist, though this should be sharply distinguished from the commitments of rational intuitionism. To assess this naturalistic appeal to intuition, it might be instructive to look at strands in recent empirical

---

moral psychology that concern the origin, nature, and regulation of moral intuitions. In particular, I would like to explore four kinds of claims about moral intuitions: (1) that they reflect the operation of imperfect cognitive heuristics, (2) that they are resistant to undermining evidence, (3) that they are subject to framing effects, and (4) that they are variable, rather than universal. Some of this empirical work on intuitions, in general, and moral intuitions, in particular, may give rise to questions about the reliability or probative value of reliance on moral convictions or intuitions.

Just as the naturalistic methods of broad dialectical equilibrium need to be distinguished from rational intuitionism, so too does the new empirical work on intuitions. The intuitions that are the object of recent empirical investigation are quite different from the sort of intuitions prized by traditional philosophical forms of intuitionism. Traditional philosophical forms of intuitionism did not inquire much about the origin or function of moral intuitions and took the moral faculty to be innate and to be the source of a priori moral knowledge. Also, in treating moral intuitions as apprehensions of self-evident moral particulars or principles, the tradition presupposed that a distinction could be drawn between genuine intuitions and false intuitions. By contrast, new work on moral intuitions operates with a much more inclusive conception of intuitions as moral beliefs that concern moral particulars and types of actions and that arise more or less spontaneously without a great deal of explicit inference and processing. These intuitionists are very much interested in the origin and function of these intuitions and make these issues the object of empirical naturalistic study. The new intuitionists and the old need not disagree about whether moral intuitions are the product of an innate faculty. But the new intuitionists raise a variety of questions about the reliability and probative value of moral intuitions (canvassed below). Though empirical work on the intuitions need not undermine moral knowledge, it will require a naturalistic commitment to the defeasible character of the evidential value of moral intuitions based on a more adequate empirical moral psychology. In this way, the new empirical intuitionism will be more closely allied with naturalism than with rational intuitionism.

VI. INTUITIONS AND HEURISTICS

For some time, psychologists and cognitive scientists have noted that we frequently operate with fairly coarse-grained rules of thumb or heuristics to help process complex information and make decisions in complex situations in a manageable and timely fashion. On some models of cognition, heuristics occupy one side in a cognitive division labor between processes that are fast, automatic, and intuitive and those that are slow,
deliberative, and self-conscious. On such dual process models of cognition, while deliberative processing (the second kind) can intervene and override intuitive processing (the first kind), it is essential to our survival and efficient task management that we often employ reliable heuristics. While such heuristics operate reliably in a familiar range of cases, they tend to be either overinclusive or underinclusive in cases at the margins. This is typically because the heuristics employ a salient feature of a situation that is a rough proxy for a more complex set of relevant properties that are harder to detect and discriminate for. On the one hand, the heuristics serve us well in a large run of cases. As a result, natural selection or cultural selection may well favor the internalization of such heuristics. On the other hand, such heuristics lead us to overgeneralize and undergeneralize in other cases, and so lead to error, sometimes potentially very significant error. In his interesting article, “Moral Heuristics,” Cass Sunstein extends these ideas about heuristics to the moral domain. He notes the utilitarian reliance on such heuristics or secondary principles as a more reliable means of promoting utility than case-by-case calculations of expected utility but suggests that the need for heuristics is more general and not limited to utilitarianism. The value of such heuristics in a large run of cases can explain why such precepts might be the product of natural or cultural selection and transmission. However, insofar as these moral heuristics are just rough proxies for what really seems to matter, they are bound to lead us into moral error in some cases. So, for instance, if you are hiding Jews in your attic, otherwise sensible heuristics about honesty won’t produce the right result when the storm trooper appears at your door and asks if you are protecting Jews. Or, for example, familiar heuristics about protecting loved ones from harm or suffering may interfere with helping them secure necessary medical attention that involves significant short-term pain and hardship or emotionally difficult psychological counseling. Presumably, there will be similar limitations to other useful heuristics about fidelity and nonmaleficence.

One of the conclusions that Sunstein draws from his discussion of imperfect moral heuristics is that some of our moral intuitions are not


to be trusted. In particular, he thinks, it is a mistake to rely on intuitions in novel or exotic cases, because these are precisely the cases for which our heuristics were not designed—the idea that hard cases make bad law. Insofar as the method of dialectical equilibrium tests principles by comparing their implications for novel or exotic cases with our intuitions about such cases, that method becomes suspect.

Of course, the importance of heuristics and their imperfections are not new ideas. As we have seen, Mill and Sidgwick thought that many common precepts, including secondary principles, were justified, to the extent that they were, by the fact that following the precept in question tended to promote the utilitarian standard. While they appealed to a utilitarian standard, other first principles can also subsume and explain such precepts. Mill explicitly claims that while secondary principles might satisfy the standard in most cases, they will not do so in all cases. This isn’t a defect in such principles if we cannot reliably discriminate for the cases in which following the principle is suboptimal and it would be optimal to depart from the principles. Sunstein seems to have in mind imperfect heuristics where it is possible to identify cases in which the heuristic is overinclusive or underinclusive. Where this is true, the heuristics would not capture optimal secondary principles. This seems closer to the picture Sidgwick holds of the coarse-grained precepts of commonsense morality enjoining veracity and fidelity.37

On the one hand, we want to be able to explain the general reliability and probative value of common moral heuristics. The basic idea is familiar enough. Much of morality is concerned with the appropriate distribution of the benefits and burdens of social interaction and requires agents to regulate the pursuit of their own self-interest by a fair division of resources and responsibilities. Commonsense morality recognizes norms of fidelity, fair play, and nonaggression the general observance of which is mutually beneficial. Though each might prefer to enjoy the benefits of others complying with such norms without incurring the costs of one’s own compliance, noncompliance is generally detectable and others won’t be cooperative and forbearing toward agents known to be noncompliant. If so, compliance generally secures greater long-term benefits than noncompliance. Moreover, because one benefits from the compliance of others, each has an interest in the cooperation and restraint of others and so has an incentive to reinforce compliant behavior and discourage noncompliant behavior. In this way, communities will tend to foster the internalization of cooperative virtues. Because of the individual and collective benefits of regulating one’s conduct according to such norms, it is no accident that such norms are often culturally transmitted and internalized. If such a moral capacity

37. Sidgwick, Methods, 342, 360–61.
is adaptive, it could have been selected for by biological evolution. If so, we have a sketch of a plausible explanation of why, if morality is roughly what we think it is, it is to be expected that, at least under favorable conditions, we are reasonably good detectors of some basic moral requirements.\textsuperscript{38}

On the other hand, we also want to be able to explain how heuristics can be imperfect and why this doesn’t infect moral methodology. The recognition of imperfect heuristics, which are both overinclusive and underinclusive, would be a problem for an intuition-driven methodology only if these intuitions were unrevisable.\textsuperscript{39} But, of course, they can be revised when they conflict with principles that explain not only where the heuristic is reliable but also where it is not. In doing so, we rely on other intuitions, including intuitions about the moral valence of cases where the heuristic seems, on reflection, to give the wrong result. When we have a principle that explains why the heuristic gets the right results in a large run of cases but the wrong result in other cases, then we have a principle that is better supported by intuitions than the heuristic, and the heuristic should be revised or be recognized as a useful but imperfect heuristic. Such principles can be thought of as characterizing our moral competence and allow us to treat imperfect heuristics as performance errors. For example, Sidgwick’s thesis, if true, implies that our underlying competence is utilitarian and asks us to treat moral intuitions that are not the product of genuine secondary principles as performance errors. But, of course, one might reject Sidgwick’s thesis and conclude that the principles that correctly characterize our underlying competence are nonutilitarian.\textsuperscript{40} The fact that we can recognize the value of imperfect heuristics and so why they might prove adaptive or ripe for cultural transmission makes it easier to treat their justification as derivative and envision the possibility of modifying or abandoning them.

VII. RECALCITRANT INTUITIONS

Jonathan Haidt is among those doing empirical work on moral intuitions who thinks that this work has philosophically unsettling implications. Haidt describes his work on intuitions as a new form of intui-

\textsuperscript{38} This explanation of the reliability of moral heuristics might be compared with the nice account offered by Peter Railton in “The Affective Dog and Its Rational Tale: Intuition and Attunement” (in this issue).

\textsuperscript{39} Though Sunstein does think that imperfect heuristics raise worries about reflective equilibrium and its appeal to intuition, he ultimately concludes that reflective equilibrium can go forward if it recognizes the limits of heuristics (“Moral Heuristics,” 542).

tionism—social intuitionism. Social intuitionism, he thinks, implies that many of our moral intuitions are less sensitive to undermining evidence than we might suppose. In “The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment,” Haidt describes research in which he presents subjects with a scenario in which an adult brother and an adult sister engage in consensual and enjoyable sex without adverse psychological consequences, while taking adequate birth control precautions. Haidt notes that the incest taboo is robust even in scenarios, such as this, in which no harm is caused, though he also notes that subjects are often unable to explain the basis for their moral conviction that incest is wrong in such circumstances. He thinks that this finding tends to undermine traditional rationalist assumptions about moral reasoning.

Haidt, in fact, draws a number of conclusions from his examples. One aspect of Haidt’s research concerns moral dumbfounding, in which respondents have confident moral beliefs but are unable to articulate reasons or principles supporting those beliefs. But dumbfounding should not be problematic. Indeed, the value of heuristics in dual processing models of cognition explains why we should expect agents to possess confident moral beliefs for which they cannot easily supply an articulate justification. As long as we can explain their general reliability, there is no reason not to assign them probative value. However, as we have just seen, recognizing some heuristics as generally reliable does not prevent us from recognizing circumstances in which they might be mistaken and in need of revision.

Haidt also suggests that moral judgment is more a matter of perception than ratiocination, that moral judgments are a product of emotions, such as disgust, and that moral reasoning is often post hoc rationalization of antecedent and fixed intuitions. These are large claims that I cannot address here properly. But their target seems to be a very simplistic and unrealistic form of rationalism that assumes that moral deliberation is always conscious and articulate and that moral emotions aren’t themselves sensitive to moral norms and reasoning. We shouldn’t assume that moral judgments are not subject to norms of correctness just because they are emotionally laden; moral emotions typically involve moral beliefs as constituent elements. So, for instance, to be motivated by shame to endorse a taboo is to accept a negative moral assessment of the action, which can then be evaluated as true or false. Nor should we assume that moral judgments are not subject to norms of correctness just because they rely on inarticulate hunches or convictions, since, as we have seen, many of our intuitive judgments rely on such heuristics, and

with good reason. So dumbfounding does not imply that the respondents are not relying on internalized moral generalizations of various kinds. Finally, the fact that judgments can be made without the benefit of conscious reasoning doesn’t mean that those judgments aren’t sensitive to the ability or inability to offer discursive justifications for those convictions. If I accept taboos that affect others negatively, even in situations in which the justifications we normally recognize do not apply, this will put pressure on me to provide an alternative rationale for the taboo, to limit its scope, or to revise it altogether.42

Haidt’s claim on which I would like to focus is that moral intuitions, such as the incest taboo, seem to be recalcitrant, that is, insensitive to potentially defeating evidence. Haidt clearly assumes that harmfulness is the primary moral factor and concludes from the fact that the incest taboo is robust for many people even when incest is harmless that many people’s intuitions about incest are evidence-insensitive. If moral intuitions are recalcitrant in this way, then they would seem to lack probative value and so should not constrain moral deliberation or theory construction.

Even if we grant Haidt’s assumption that no harm in fact ensues in his incest vignette—not only are no offspring produced but also the incest does not damage the relationship between the incestuous siblings or their relationships to other family members—they certainly risked harms of these kinds. If it is appropriate for taboos to track harms risked, as well as caused, then these robust taboos may not be recalcitrant after all.43

For this reason, we might look for better examples of recalcitrant intuitions. Consider phenomena not discussed by Haidt. The criminal law, criminal jurisprudence, and much moral philosophy assume that reactive attitudes, such as blame, and punishment aim to track culpable wrongdoing. That perspective assumes that blame and punishment depend on two independent variables, wrongdoing and responsibility for the wrongdoing. On plausible views, responsibility is a scalar concept that tracks in significant part degree of normative competence, which should be affected by, among other things, the maturity of the wrong-

42. The parallels between Haidt’s claims about the post hoc character of moral justifications and the critique by legal realists (e.g., Jerome Frank) of principled adjudication and defense of the post hoc character of judicial reasoning are quite striking. I take the legal realist critique to be misguided in several ways, which have direct analogues in this part of Haidt’s argument. That case will have to be made on another occasion.

43. Railton (in his contribution) motivates a similar doubt about whether the incest taboo is recalcitrant by comparing the harm risked there with a parallel case in which a father dangles a baby out a window, the baby loved the experience, and the father’s grip was secure. It seems we rightly condemn such behavior, though no harm ensued, on the ground that harm was unnecessarily risked.
doer. But in some important studies involving assessments of the severity of punishment appropriate for heinous crimes the assessments of respondents were surprisingly insensitive to information about the age of the wrongdoer. In such cases, respondents seemed to track the scale of the wrongdoing, largely ignoring differences in the degree of maturity and normative competence of the wrongdoer. So in these cases, the intuitions of respondents are not tracking evidence that should have mitigated their assessments of blameworthiness.

What should we say about genuinely recalcitrant intuitions? Not all intuitions are equal. Even a narrow equilibrium distinguishes considered moral judgments from other intuitive moral beliefs. On one conception of the dialectical process, only considered moral judgments are part of the inputs to moral deliberation and theorizing. On the more inclusive conception I favor, all intuitions figure as inputs, but considered moral judgments are assigned greater weight. For present purposes, we needn’t decide between these two conceptions. Both allow us to discriminate among intuitions as more and less valuable. Considered moral convictions are those intuitions that are not rushed, are stable over time, are robust under dialectical pressures for consistency and full information, and are not subject to obvious forms of bias. We might decide to include within the class of considered moral convictions only those moral convictions that are sensitive to potentially defeating evidence. So if we could identify intuitions that were genuinely recalcitrant in this sense, that would be reason not to regard them as considered moral judgments and so either to exclude them as inputs or, more plausibly, to allow them in but attach less weight to them.

Of course, we can’t let the determination of whether intuitions are sensitive to the evidence depend on a selection of the relevant evidence by the principle we are trying to test. That rationale for dismissing counterevidence would be viciously circular. To dismiss intuitions as recalcitrant in this sense, we would have to have more ecumenical evidence that the intuitions were unresponsive to relevant evidence. In the case of the recalcitrance of intuitions concerning just deserts, both criminal law and moral theory treat responsibility for wrongdoing as a...
condition of deserved punishment, and immaturity is, on many conceptions of responsibility, a basis for diminished responsibility.

It will be easier to justify this sort of dismissal if we have available a good explanation of why we might be subject to such convictions even if they were not true and responsive to evidence. Whether or not the incest taboo is a genuinely recalcitrant intuition, we have available in that case just such an explanation, inasmuch as there are plausible natural and cultural evolutionary explanations for why we would have developed an incest taboo in normal circumstances that may not always provide reasons to avoid incest in circumstances where concerns about interbreeding are absent. In the case of overly punitive intuitions about just deserts in cases involving immaturity we have an overgeneralization about the sufficiency of wrongdoing for punishment from cases in which there is normal responsibility to cases in which there is not.

There is another reason that it is unclear whether the intuitions in Haidt’s incest study are genuinely recalcitrant. The evidence that Haidt provides (as I understand it) has a snapshot quality. In response to the incest vignette, many respondents express the conviction that incest is wrong even in circumstances when it is consensual and does not lead to interbreeding or psychological harm, often saying it is wrong while confessing an inability to say what makes it wrong. This suggests that these respondents initially have moral convictions that they cannot support with acceptable reasons. It does not show that these respondents are incapable of reasoning further about their convictions and either finding an acceptable rationale for them or revising them for want of a suitable rationale. This would be further reason for questioning whether their initial reactions count as considered moral convictions insofar as they might prove unstable over time. It also points out the importance of conceiving of reflective equilibrium as a dynamic dialectical process.

VIII. INTUITIONS AND FRAMING EFFECTS

For some time, behavioral economists, psychologists, and cognitive scientists have studied ways in which the framing or presentation of alternatives elicits different preferences among options. Standard decision theory assumes the invariance of preferences over equivalent options. Framing effects seem to challenge the invariance assumption. Consider the following example from Daniel Kahneman and Amos Tversky.45 Respondents are asked to imagine that the government is preparing for an outbreak of an unusual Asian disease that would otherwise kill 600

people. They are asked to express a preference between two proposed programs to combat the disease.

- If Program A is adopted, 200 people will be saved.
- If Program B is adopted, there is a one-third probability that 600 will be saved and a two-thirds probability that no one will be saved.

In the responses, 72 percent of respondents prefer Program A to B, while 28 percent prefer Program B to A. Another study was run premised on the same background story. In this case, respondents were asked to express a preference between these two programs.

- If Program C is adopted, 400 people will die.
- If Program D is adopted, there is a one-third probability that no one will die and a two-thirds probability that 600 people will die.

Here, 78 percent of respondents prefer Program D to C, while 22 percent prefer Program C to D.

But, of course, A and C describe the same policy, and B and D describe the same policy. Here, the framing effects seem to depend on whether the effects of the policies are described in terms of lives saved or in terms of lives lost. Intuitively, these different descriptions of the policies and their consequences should not affect their value or our preferences among them.

What is true of preferences can also be true of moral judgments or intuitions. Our moral beliefs about which action or policy is better sometimes vary depending upon apparently inconsequential differences in the ways in which the two options are described or presented—the order in which options or their components are described, whether outcomes are described in terms of what is gained or in terms of what is lost, and whether negatively valenced outcomes are viewed as costs or losses. Indeed, the public health options described above could easily be represented as policies among which respondents are asked to express a belief about which policy is morally better.

It is tempting to say that framing effects do challenge invariance and that they also reveal that our preferences and intuitions are inconsistent. But not all (alleged) framing effects actually challenge invariance. Sometimes what framing effects show is that what we thought were equivalent options are not in fact equivalent. For example, someone might claim that the preference for Program A over B already violates invariance, because a sure thing for \( X \) utiles is equivalent to any lottery of the form \( (X \times Y) / Y \). But one might well respond that the preference for sure gains shows that we don’t view these as equivalent options; the certainty of a gain contributes to its utility. Similarly, so-called prospect theory models people’s preferences by appeal to the importance of assess-
ing possible gains and losses relative to a baseline or status quo: all else being equal, we are more concerned to avoid losses than we are eager to realize comparable gains; and, all else being equal, we are more averse to losses than to costs. One way to think of prospect theory is as an attempt to reconcile framing effects with the invariance assumption.

However, we cannot reconcile all framing effects with invariance without giving up any independent grip on when two options are equivalent. On reasonable assumptions about when options are equivalent, we cannot assume that framing effects won’t sometimes challenge invariance. For instance, in the health policy example, it’s hard not to regard A and C as equivalent and B and D as equivalent. Here, it seems, actual preferences do not obey the invariance assumption, and the resulting preferences and moral intuitions seem genuinely inconsistent. We may not find inconsistency wherever we find framing effects, but some framing effects do seem to reveal inconsistency. To the extent that our moral methods appeal to intuitions that are subject to framing effects, there is a worry that our methods can be unreliable.

As I said, not all framing effects really violate the invariance assumption. Sometimes the correct conclusion to draw is that what we assumed were equivalent descriptions of the same option elicited different preferences because the options were really subtly different options. But sometimes framing effects really do violate invariance, because people have conflicting preferences about one and the same option, depending on how it is presented. Here, we have genuine inconsistency of preference and belief. But this is not itself a worry for dialectical equilibrium, which assumes that perfect accommodation is impossible because of inconsistent beliefs and commits us to principled revision of some of our beliefs. Much of the initial literature on framing effects also assumes a snapshot perspective, focusing on the inconsistent preferences elicited in subjects unaware of the framing effects and resulting inconsistency. However, reflective equilibrium is a dynamic process. Presumably, we might try to resolve the inconsistency in part by seeing how, if at all, convictions change when subjects are made aware of the inconsistency and the phenomenon of framing effects. In some cases, this will produce a new comparatively stable preference, and that is presumably the one to rely on going forward. Alternatively, the inconsistency may remain, in which case we have a diachronically robust framing effect and have to treat this as a case of intrapersonal inconsistency. The conflicting intuitions will cancel each other out, and dialectical fit will have to be settled on other grounds or cases. If and when it is, the

---

principles that provide the best fit elsewhere can provide guidance in the

case of perplexity due to intrapersonal inconsistency.

IX. VARIABLE INTUITIONS

It is not just ethicists that test their principles or theories at the bar of

intuition. Philosophers test different systems of modal logic by seeing

whether their axioms allow us to prove all and only those claims about

possibility and necessity that we find evident on reflection. Edmund

Gettier was widely seen as undermining a conception of knowledge that

had lasted two millennia by drawing our attention to cases of justified

true belief that intuition seemed to reveal were not cases of knowledge.47

Similarly, in ethics, many people think that we can test moral principles

by intuitions. For instance, many think that the possibility that utili-

tarianism might justify the punishment of an innocent person if the

consequences of doing so were good enough shows that utilitarianism

should be rejected because we have the firm intuition that it is imper-

missible to punish the innocent. Of course, it is well known that not

all intuitions are universally shared. But the hope seems to be that we

can build and test principles against widely shared intuitions first and

then extend principles that fare well there to cases where there is less

agreement or more uncertainty. But one thing that research in experi-

mental ethics seems to show is that there is more widespread disagree-

ment about moral intuitions, not just at the margins but also in central

cases. Some of this evidence shows intercultural variation.48 But there

is also intracultural variation. In a very interesting recent meta-study,

“Gender and Philosophical Intuition,” Wesley Buckwalter and Stephen

Stich notice that there is significant variation in philosophical intuition

along gender lines.49 Significantly, these differences in intuition aren’t

confined to cases widely regarded as controversial but concern cases

in which it had been assumed that intuitions were widely shared, such


48. Early studies of intercultural moral disagreement include William Sumner, Folk-

ways (Boston: Ginn, 1906); and Edward Westermarck, Ethical Relativity (New York: Harcourt

& Brace, 1952). For a recent discussion, see Jesse Prinz, The Emotional Construction of Morals


49. Wesley Buckwalter and Stephen Stich, “Gender and Philosophical Intuition,” in


Buckwalter and Stich make the interesting suggestion that gender differences in philo-

sophical intuition and methodology might help explain the disproportionately low num-

bers of women in graduate programs and the professoriate within academic philosophy.

For some skepticism about whether there are significant gender differences in philosophi-

cal intuition, see Toni Adleberg, Morgan Thompson, and Eddy Nahmias, “Do Men and

Women Have Different Philosophical Intuitions? Further Data,” Philosophical Psychology

(forthcoming).
as Gettier cases and cases involving punishment of the innocent. Of course, the idea of gender-specific differences in moral intuitions can be traced back to earlier work by Carol Gilligan, who claimed to find gender-specific differences in both moral intuition and in forms of moral reasoning. If disagreement in intuition is so widespread, how can a methodology that tests principles by intuition hope to resolve disagreement and produce convergence?

Interpersonal moral disagreement would be no problem if we each pursued our own dialectical inquiries solipsistically, worrying only about the fit between principles and our own intuitions. But, of course, we are not solipsistic reasoners. We can try to achieve a dialectical equilibrium collectively, and then our disagreements will be unavoidable. To some extent, analytical normative ethics can be seen as the collective pursuit of reflective equilibrium. But even when we conduct our dialectical inquiries individually, we want our reasoning to hold up in public, with actual and potential interlocutors, and so we try to be answerable in our inquiries to what others might think and where they might disagree with us. So there is no getting around the problem of disagreement. It is a worry for dialectical methods that hope to defend calls for principled reform of some our moral beliefs on the basis of principles that draw support from other moral beliefs. Why think that moral principles will be able to resolve our disagreements if they depend for their support on just the sort of moral beliefs about which we disagree?

Moral disagreement is a real worry for any bottom-up strategy of justification in ethics. Moreover, it is not a worry that can be fully addressed here, in the course of a discussion of the significance of experimental ethics for traditional assumptions about moral methodology. But two things can and should be noted.

First, the phenomenon of moral disagreement is not new; it was not something discovered recently by experimental ethics. So it is not some new problem or challenge posed by experimental ethics. Second, a naturalistic version of reflective equilibrium has resources for dealing with moral disagreement that traditional appeals to intuition do not. If moral intuitions are supposed to be self-evident, it is unclear what to say or do when intuitions conflict. If genuine intuitions are self-evident or noninferentially justified, it is hard to see how we could give reasons in their support to someone who disagrees with us. But if intuitions are inferentially justified by how well they fit with other principles, intuitions, and auxiliary commitments, then we can hope to discover that

considerations of systematic comparative plausibility would give at least one of us reason to revise his initial moral beliefs. Since we expect a dialectical fit to be revisionary, there’s no reason that interlocutors need share all moral beliefs at the outset. Indeed, since there is no guarantee that any moral beliefs will survive reflective equilibrium unmodified, there is no need for them to share any common ground at the beginning of the inquiry. Though common ground may not be strictly necessary for convergence in principle at the end of the inquiry, it is certainly helpful and increases the prospects for convergence. There may be metadisagreement about how pervasive moral disagreement is, but no one denies that there is some common ground. Moreover, a broad dialectical fit can be even more revisionary than a narrow one provided that it is driven by enough philosophical theory and social science. Given the dependence of many of our moral beliefs on complex empirical and philosophical issues, there is every reason to expect a broad equilibrium to be revisionary. Moreover, because a broad dialectical fit is an intellectual ideal that we at most hope to approximate, the existence of pre-reflective and even reflective moral disagreement provides little reason to think that moral disagreement is in principle unresolvable. Indeed, the claim that moral disagreements are in principle unresolvable by dialectical methods is just one claim about what the results of a systematic dialectical inquiry among different interlocutors would be and enjoys no privileged a priori position in relation to its nonskeptical competitors. The ethical naturalist should think that the prospect for convergence is itself an empirical matter. But we can be justified in pursuing a broad dialectical equilibrium even if there is no a priori guarantee of convergence.

X. CONCLUDING REMARKS

We have looked at some interesting and potentially unsettling empirical claims about moral intuition—that they reflect the operation of imperfect cognitive heuristics, that they are resistant to undermining evidence, that they are subject to framing effects, and that they are variable, rather than universal. To assess the significance of this empirical work on intuitions for the role of intuitions in philosophical ethics it helps to view these issues in the context of traditional debates between rational intuitionists and ethical naturalists about the nature and epistemological role of moral intuitions. This empirical work on moral intuitions is very different from rational intuitionism and has more in common with Mill’s naturalist critique of rational intuitionism and his dialectical conception of the justification of first principles, which was further developed in the method of reflective equilibrium. This empir-
ical work raises natural questions about the reliability of intuition-driven methods in ethics. But there are resources to address such questions within ethical naturalism. Indeed, becoming aware of the origins and limitations of some of our moral beliefs provides us with an additional resource in deciding how best to balance the demands for accommodation and reform that are part of the search for broad dialectical equilibrium.
QUERIES TO THE AUTHOR

q1. AU: It is Press and Ethics style to use arabic numerals for all divisions in documentation for several reasons: for the sake of consistency of cited works across all articles in the journal, because it’s too time consuming for us to check the originals of all cited works or to keep track of official styles for all philosophical editions, because various cited works may appear in different editions and formats, and because a consistent style with one kind of numeral is visually less distracting to the reader.

q2. AU: Initially, “72%” was changed to “Seventy-two percent” because it is Ethics style not to start a sentence with a numeral. I agree with your point that the two numbers in this sentence should appear consistently, and to accommodate that I added an opening phrase. Feel free to revise, as long as the sentence doesn’t start with a numeral.