BEHAVIORAL AND BRAIN SCIENCES (2005) **28,** 531–573
Printed in the United States of America

Moral heuristics

Cass R. Sunstein
University of Chicago Law School and Department of Political Science, University of Chicago, Chicago, IL 60637.
csunstei@uchicago.edu
http://www.law.uchicago.edu/faculty/sunstein/

Abstract: With respect to questions of fact, people use heuristics – mental short-cuts, or rules of thumb, that generally work well, but that also lead to systematic errors. People use moral heuristics too – moral short-cuts, or rules of thumb, that lead to mistaken and even absurd moral judgments. These judgments are highly relevant not only to morality, but to law and politics as well. Examples are given from a number of domains, including risk regulation, punishment, reproduction and sexuality, and the act/omission distinction. In all of these contexts, rapid, intuitive judgments make a great deal of sense, but sometimes produce moral mistakes that are replicated in law and policy. One implication is that moral assessments ought not to be made by appealing to intuitions about exotic cases and problems; those intuitions are particularly unlikely to be reliable. Another implication is that some deeply held moral judgments are unsound if they are products of moral heuristics. The idea of error-prone heuristics is especially controversial in the moral domain, where agreement on the correct answer may be hard to elicit; but in many contexts, heuristics are at work and they do real damage. Moral framing effects, including those in the context of obligations to future generations, are also discussed.

1. Introduction

Pioneering the modern literature on heuristics in cognition, Amos Tversky and Daniel Kahneman contended that "people rely on a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations" (Tversky & Kahneman 1974, p. 1124). Intense controversy has developed over the virtues and vices of such heuristics, most of them "fast and frugal," that play a role in many areas (see Gigerenzer et al. 1999; Gilovich et al. 2002). But the relevant literature has only started to investigate the possibility that in the moral and political domains, people also rely on simple rules of thumb that often work well but that sometimes misfire (see Baron 1994a; 1998; Messick 1993). In fact, the central point seems obvious. Much of everyday morality consists of simple, highly intuitive rules that generally make sense, but that fail in certain cases. It is wrong to lie or steal, but if a lie or a theft would save a human life, lying or stealing is probably obligatory. Not all promises should be kept. It is wrong to try to get out of a longstanding professional commitment at the last minute, but if your child is in the hospital, you may be morally required to do exactly that.

One of my major goals in this article is to show that heuristics play a pervasive role in moral, political, and legal judgments, and that they sometimes produce significant mistakes. I also attempt to identify a set of heuristics that now influence both law and policy, and try to make plausible the claim that some widely held practices and beliefs are a product of those heuristics. Often moral heuristics represent generalizations from a range of problems for which they are indeed well-suited (see Baron 1994a), and hence, most of the time, such heuristics work well. The problem comes when the generalizations are wrenched out of context and treated as freestanding or universal principles, applicable to situations in which their justifications no longer operate. Because the generalizations are treated as freestanding or universal, their application seems obvious, and those who reject them appear morally obtuse, possibly even monstrous. I contend that the appearance is misleading and even productive of moral mistakes. There is nothing obtuse, or monstrous, about refusing to apply a generalization in contexts in which its rationale is absent.

Because Kahneman and Tversky were dealing with facts and elementary logic, they could demonstrate that the heuristics sometimes lead to errors. Unfortunately, that cannot easily be demonstrated here. In the moral and political domains, it is hard to come up with unambiguous cases in which the error is both highly intuitive and on reflection uncontroversial – that is, cases in which people can ultimately be embarrassed about their own intuitions. Nonetheless, I hope to show that whatever one's moral commitments, moral heuristics exist and indeed are omnipresent. We should not treat the underlying moral intuitions as fixed points for analysis, rather than as unreliable and potentially erroneous. In the search for reflective equilibrium, understood to be coherence among our judgments at all levels of generality (Rawls 1971), it is important to see

CASS R. SUNSTEIN is Karl N. Llewellyn Distinguished Service Professor, Law School and Department of Political Science, University of Chicago. His many books include *Laws of Fear* (2005), *Why Societies Need Dissent* (2003), and *Risk and Reason* (2002). Much of his research explores the intersection of psychology, economics, and law; he has been particularly interested in the role of cognitive errors and social influences on juries, legislators, judges, and those who are the objects of law's commands. He is now working on a book on the production of social knowledge, with special reference to information aggregation across persons.

© 2005 Cambridge University Press 0140-525X/05 $12.50

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

that some of our deeply held moral beliefs might be products of heuristics that sometimes produce mistakes.

If moral heuristics are in fact pervasive, then people with diverse foundational commitments should be able to agree, not that their own preferred theories are wrong, but that they are often applied in a way that reflects the use of heuristics. Utilitarians ought to be able to identify heuristics for the maximization of utility; deontologists should be able to point to heuristics for the proper discharge of moral responsibilities; and those uncommitted to any large-scale theory should be able to specify heuristics for their own more modest normative commitments. And if moral heuristics exist, blunders are highly likely not only in moral thinking, but in legal and political practice as well. Conventional legal and political arguments are often a product of heuristics masquerading as universal truths. Hence, I will identify a set of political and legal judgments that are best understood as a product of heuristics, and that are often taken, wrongly and damagingly, as a guide to political and legal practice, even when their rationale does not apply.

2. Ordinary heuristics and an insistent homunculus

2.1. Heuristics and facts

The classic work on heuristics and biases deals not with moral questions but with issues of fact. In answering hard factual questions, those who lack accurate information use simple rules of thumb. How many words, in four pages of a novel, will have "ing" as the last three letters? How many words, in the same four pages, will have "n" as the second-to-last letter? Most people will give a higher number in response to the first question than in response to the second (Tversky & Kahneman 1984) – even though a moment's reflection shows that this is a mistake. People err because they use an identifiable heuristic – the availability heuristic – to answer difficult questions about probability. When people use this heuristic, they answer a question of probability by asking whether examples come readily to mind. How likely is a flood, an airplane crash, a traffic jam, a terrorist attack, or a disaster at a nuclear power plant? Lacking statistical knowledge, people try to think of illustrations. For those without statistical knowledge, it is far from irrational to use the availability heuristic; the problem is that this heuristic can lead to serious errors of fact, in the form of excessive fear of small risks and neglect of large ones.

Or consider the representativeness heuristic, in accordance with which judgments of probability are influenced by assessments of resemblance (the extent to which A "looks like" B). The representativeness heuristic is famously exemplified by people's answers to questions about the likely career of a hypothetical woman named Linda, described as follows: "Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice and also participated in antinuclear demonstrations" (see Kahneman & Frederick 2002; Mellers et al. 2001). People were asked to rank, in order of probability, eight possible futures for Linda. Six of these were fillers (such as psychiatric social worker, elementary school teacher); the two crucial ones were "bank teller" and "bank teller and active in the feminist movement."

More people said that Linda was less likely to be a bank teller than to be a bank teller and active in the feminist movement. This is an obvious mistake, a conjunction error, in which characteristics A and B are thought to be more likely than characteristic A alone. The error stems from the representativeness heuristic: Linda's description seems to match "bank teller and active in the feminist movement" far better than "bank teller." In an illuminating reflection on the example, Stephen Jay Gould observed that "I know [the right answer], yet a little homunculus in my head continues to jump up and down, shouting at me – 'but she can't just be a bank teller; read the description'" (Gould 1991, p. 469). Because Gould's homunculus is especially inclined to squawk in the moral domain, I shall return to him on several occasions.

Of course, the early work on heuristics has been subject to intense criticism, sometimes with the claim that in the real world, most heuristics work quite well (Gigerenzer 2000; Gigerenzer et al. 1999). In this view, many findings of cognitive errors are an artifact of the laboratory setting and of clever experimental designs involving unfamiliar problems; in the real world, people may be much less likely to err, perhaps because the heuristics are sensible, perhaps because they are not applied indiscriminately (Krueger & Funder 2004; cf. Kahneman & Tversky 1996). In addition, some of the key ideas, including availability and representativeness, have been challenged as inadequately specified and as subject to ad hoc applications (Gigerenzer 1996; cf. Kahneman & Tversky 1996, p. 591). For present purposes, it is unnecessary to resolve these debates here. No one denies that with respect to facts, human beings use simple rules of thumb that can produce serious mistakes; even the "recognition heuristic," said to enable people to make remarkably accurate judgments about the size of cities or prospects in sports events (Goldstein & Gigerenzer 2002), will produce severe and predictable errors. If this is true for facts, it is highly likely to be true for political and moral judgments as well.

With respect to moral heuristics, existing work is suggestive rather than definitive; a great deal of progress remains to be made, above all through additional experimental work on moral judgments. Some of the moral heuristics that I shall identify might reasonably be challenged as subject to ad hoc rather than predictable application. One of my primary hopes is to help stimulate further research, testing when and whether people use moral heuristics that produce sense or nonsense in particular cases.

2.2. Attribute substitution and prototypical cases

What is a heuristic? Kahneman and Frederick have recently suggested that heuristics are mental shortcuts used when people are interested in assessing a "target attribute" and when they substitute a "heuristic attribute" of the object, which is easier to handle (Kahneman & Frederick 2002). Heuristics therefore operate through a process of *attribute substitution*. The use of heuristics gives rise to intuitions about what is true, and these intuitions sometimes are biased, in the sense that they produce errors in a predictable direction. Consider the question of whether more people die from suicides or homicides. Lacking statistical information, people might respond by asking whether it is easier to recall cases in either class (the availability heuristic). The approach is hardly senseless, but it might also lead to errors, a result of "availability bias" in the domain of risk percep-

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

tion (see Kuran & Sunstein 1999). For the size of cities, the recognition heuristic, which is a close cousin of the availability heuristic, has the same problem, leading to what might be called "recognition bias." Sometimes heuristics are linked to affect, and indeed affect has even been seen as a heuristic (Slovic et al. 2002); but attribute substitution is often used for factual questions that lack an affective component.

Similar mechanisms are at work in the moral, political, and legal domains. Unsure of what to think or do about a target attribute (what morality requires, what the law is), people might substitute a heuristic attribute instead – asking, for example, about the view of trusted authorities (a leader of the preferred political party, an especially wise judge, or a religious figure). Every law professor in the United States knows that in approaching difficult constitutional questions, many law students are drawn to a kind of "Justice Antonin Scalia heuristic." If students are unsure how to analyze a constitutional problem, they might ask instead what Justice Scalia (an influential conservative on the United States Supreme Court) thinks – and either follow him or do the opposite. In the areas of morality, politics, and law, attribute substitution is pervasively involved. Often the process works by appeal to *prototypical cases*. Confronted by a novel and difficult problem, observers often ask whether it shares features with a familiar problem. If it seems to do so, then the solution to the familiar problem is applied to the novel and difficult one. Of course it is possible that in the domain of values as well as facts, real-world heuristics generally perform well in the real world – so that moral errors are reduced, not increased, by their use, at least compared to the most likely alternatives (see my remarks on rule-utilitarianism later). The only claim here is that some of the time, our moral judgments can be shown to misfire.

The principal heuristics should be seen in light of dual-process theories of cognition (Kahneman & Frederick 2002). Those theories distinguish between two families of cognitive operations, sometimes labeled System I and System II. System I is intuitive; it is rapid, automatic, and effortless (and it features Gould's homunculus). System II, by contrast, is reflective; it is slower, self-aware, calculative, and deductive. System I proposes quick answers to problems of judgment, and System II operates as a monitor, confirming or overriding those judgments. Consider, for example, someone who is flying from New York to London in the month after an airplane crash. This person might make a rapid, barely conscious judgment, rooted in System I, that the flight is quite risky; but there might well be a System II override, bringing a more realistic assessment to bear. System I often has an affective component, but it need not; for example, a probability judgment might be made quite rapidly and without much affect at all.

There is growing evidence that people often make automatic, largely unreflective moral judgments, for which they are sometimes unable to give good reasons (see Greene & Haidt 2002; Haidt 2001; cf. Pizarro & Bloom 2003). Moral, political, or legal judgments often substitute a heuristic attribute for a target attribute; System I is operative here as well, and it may or may not be subject to System II override. Consider the incest taboo. People have moral revulsion against incest even in circumstances in which the grounds for that taboo seem to be absent; they are subject to "moral dumbfounding" (Haidt et al. 2004), that is, an inability to give an account for a firmly held intuition. It is plausible, at least, to think that System I is driving their judgments, without System II correction. The same is true in legal and political contexts as well.

3. Heuristics and morality

To show that heuristics operate in the moral domain, we have to specify some benchmark by which we can measure moral truth. On these questions I want to avoid any especially controversial claims. Whatever's one view of the foundations of moral and political judgments, I suggest, moral heuristics are likely to be at work in practice. In this section I begin with a brief account of the possible relationship between ambitious theories (understood as large-scale accounts of the right or the good) and moral heuristics. I suggest that for those who accept ambitious theories about morality or politics, it is tempting to argue that alternative positions are mere heuristics; but this approach is unpromising, simply because any ambitious theory is likely to be too contentious to serve as the benchmark for measuring moral truth. Progress is best made, not by opposing (supposedly correct) ambitious theories to (supposedly blundering) common sense morality, but in two more modest ways: first, by showing that moral heuristics are at work in *any* view about what morality requires; and second, by showing that such heuristics are at work in a minimally contentious view about what morality requires. I will identify a number of such heuristics in Section 4.

3.1. Theories, heuristics, and ambitious starts

Many utilitarians, including John Stuart Mill and Henry Sidgwick, argue that ordinary morality is based on simple rules of thumb that generally promote utility but that sometimes misfire (see Mill 1861/1971, pp. 28–29; Sidgwick 1907/1981, pp. 199–216; cf. Hare 1981; Smart 1973). For example, Mill emphasizes that human beings "have been learning by experience the tendencies of experience," so that the "corollaries from the principle of utility" are being progressively captured by ordinary morality (Mill 1861/1971, p. 29).[1] Is ordinary morality a series of heuristics for what really matters, which is utility?

With the aid of modern psychological findings, utilitarians might be tempted to make exactly this argument (see Baron 1998). They might contend that ordinary moral commitments are a set of mental shortcuts that generally work well, but that also produce severe and systematic errors from the utilitarian point of view. Suppose that most people reject utilitarian approaches to punishment and are instead committed to retributivism; this is their preferred theory. Are they responding to System I? Might they be making a cognitive error? (Is Kantianism a series of cognitive errors?) Note that with respect to what morality requires, utilitarians frequently agree with their deontological adversaries about concrete cases; they can join in accepting the basic rules of criminal and civil law. When deontologists and others depart from utilitarian principles, perhaps they are operating on the basis of heuristics that usually work well but sometimes misfire.

But it is exceedingly difficult to settle large-scale ethical debates in this way. In the case of many ordinary heuristics, based on availability and representativeness, a check of the facts, or of the elementary rules of logic, will show that peo-

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

ple err. In the moral domain, this is much harder to demonstrate. To say the least, those who reject utilitarianism are not easily embarrassed by a demonstration that their moral judgments can lead to reductions in utility. For example, utilitarianism is widely challenged by those who insist on the importance of distributional considerations. It is far from clear that a moderate utility loss to those at the bottom can be justified by a larger utility gain for many at the top (Rawls 1971; see also Nussbaum 1984; Sen 1980–1981).

Emphasizing the existence of moral heuristics, those who reject utilitarianism might well turn the tables on their utilitarian opponents. They might contend that the rules recommended by utilitarians are consistent, much of the time, with what morality requires – but also that utilitarianism, taken seriously, produces serious mistakes in some cases. In this view, utilitarianism is itself a heuristic, one that usually works well but leads to systematic errors. And indeed, many debates between utilitarians and their critics involve claims, by one or another side, that the opposing view usually produces good results, but also leads to severe mistakes and should be rejected for that reason (see Smart & Williams 1973).

These large debates are not easy to resolve, simply because utilitarians and deontologists are most unlikely to be convinced by the suggestion that their defining commitments are mere heuristics. Here, there is a large difference between moral heuristics and the heuristics uncovered in the relevant psychological work, in which the facts or simple logic provide a good test of whether people have erred. If people tend to think that more words, in a given space, end with the letters "ing" than have "n" in the next-to-last position, something has clearly gone wrong. If people think that some person Linda is more likely to be "a bank teller who is active in the feminist movement" than a "bank teller," there is an evident problem. If citizens of France or Germany think that New York University is more likely to have a good basketball team than St. Joseph's University, because they have not heard of the latter, then a simple examination of the record will show that they are wrong. In the moral domain, factual blunders and simple logic do not provide such a simple test.

3.2. Neutral benchmarks and weak consequentialism

My goal here is, therefore, not to show that common sense morality is a series of heuristics for the correct general theory (as suggested by Sidgwick and Mill), but more cautiously, that in many particular cases, moral heuristics are at work – and that this point can be accepted by people with diverse general theories, or with grave uncertainty about which general theory is correct. In the cases catalogued in Section 5, I contend that it is possible to conclude that a moral heuristic is at work without accepting any especially controversial normative claims. In several of the examples, that claim can be accepted without accepting any contestable normative theory at all. Other examples will require acceptance of what I shall call "weak consequentialism," in accordance with which the social consequences of the legal system are relevant, other things being equal, to what law ought to be doing.

Weak consequentialists need not be utilitarians; they do not have to believe that law and policy should attempt to maximize utility. They might agree that violations of rights count among the consequences that ought to matter, so that deontological considerations play a role in the overall assessment of what should be done. Consider Amartya Sen's frequent insistence that consequentialists can insist that consequences count without accepting utilitarianism and without denying that violations of rights are part of the set of relevant consequences (see Sen 1982; 1985). Thus Sen urges an approach that "shares with utilitarianism a consequentialist approach (but differs from it in not confining attention to utility consequences only)" while also attaching "intrinsic importance to rights (but . . . not giving them complete priority irrespective of other consequences)" (Sen 1996, p. 1038). Weak consequentialism is in line with this approach. In evaluating decisions and social states, weak consequentialists might well be willing to give a great deal of weight to nonconsequentialist considerations.

Of course, some deontologists will reject any form of consequentialism altogether. They might believe, for example, that retribution is the proper theory of punishment, and that the consequences of punishment are never relevant to the proper level of punishment. Some of my examples will be unpersuasive to deontologists who believe that consequences do not matter at all. But weak consequentialism seems to me sufficiently nonsectarian, and attractive to sufficiently diverse people, to make plausible the idea that in the cases at hand, moral heuristics are playing a significant role. And for those who reject weak consequentialism, it might nonetheless be productive to ask whether, from their own point of view, certain rules of morality and law are reflective of heuristics that sometimes produce serious errors.

3.3. Evolution and rule-utilitarianism: Simple heuristics that make us good?

Two clarifications before we proceed. First, some moral heuristics might well have an evolutionary foundation (de Waal 1996; Katz 2000; Sober & Wilson 1999). Perhaps natural selection accounts for automatic moral revulsion against incest or cannibalism, even if clever experiments, or life, can produce situations in which the revulsion is groundless. In the case of incest, the point is straightforward: The automatic revulsion might be far more useful, from the evolutionary perspective, than a more fine-grained evaluation of contexts (Stein 2001). In fact, an evolutionary account might be provided for most of the heuristics that I explore here. When someone has committed a harmful act, evolutionary pressures might well have inculcated a sharp sense of outrage and a propensity to react in proportion to it. As a response to wrongdoing, use of an *outrage heuristic* might well be much better than an attempt at any kind of consequentialist calculus, weak or strong. Of course, many moral commitments are a product not of evolution but of social learning and even cascade effects (see Sunstein 2003); individuals in a relevant society will inevitably be affected by a widespread belief that it is wrong to tamper with nature (discussed later), and evolutionary pressures need not have any role at all.

Second, and related, some or even most moral heuristics might have a rule-utilitarian or rule-consequentialist defense (see Hooker 2000). The reason is that in most cases they work well despite their simplicity, and if people attempted a more fine-grained assessment of the moral issues involved, they might make more moral mistakes rather than fewer (especially because their self-interest is frequently at stake). Simple but somewhat crude moral principles might lead to less frequent and less severe moral errors than com-

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

plex and fine-grained moral principles. Compare the availability heuristic. Much of the time, use of that heuristic produces speedy judgments that are fairly accurate, and those who attempt a statistical analysis might make more errors (and waste a lot of time in the process). If human beings use "simple heuristics that make us smart" (Gigerenzer et al. 1999), then they might also use "simple heuristics that make us good." I will offer some examples in which moral heuristics seem to me to produce significant errors for law and policy, but I do not contend that we would be better off without them. On the contrary, such heuristics might well produce better results, from the moral point of view, than the feasible alternatives – a possibility to which I will return.

4. The Asian disease problem and moral framing

In a finding closely related to their work on heuristics, Kahneman and Tversky find "moral framing" in the context of what has become known as "the Asian disease problem" (Kahneman & Tversky 1984). Framing effects do not involve heuristics, but because they raise obvious questions about the rationality of moral intuitions, they provide a valuable backdrop. Here is the first component of the problem:

Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences are as follows:

If Program A is adopted, 200 people will be saved.

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

Which of the two programs would you favor?

Most people choose Program A.

But now consider the second component of the problem, in which the same situation is given, but followed by this description of the alternative programs:

If Program C is adopted, 400 people will die.

If Program D is adopted, there is a one-third probability that nobody will die and a two-thirds probability that 600 people will die.

Most people choose Problem D. But a moment's reflection should be sufficient to show that Program A and Program C are identical, and so too for Program B and Program D. These are merely different descriptions of the same programs. The purely semantic shift in framing is sufficient to produce different outcomes. Apparently, people's moral judgments about appropriate programs depend on whether the results are described in terms of "lives saved" or in terms of "lives lost." What accounts for the difference? The most sensible answer begins with the fact that human beings are pervasively averse to losses (hence the robust cognitive finding of loss aversion, Tversky & Kahneman 1991). With respect to either self-interested gambles or fundamental moral judgments, loss aversion plays a large role in people's decisions. But what counts as a gain or a loss depends on the baseline from which measurements are made. Purely semantic reframing can alter the baseline and hence alter moral intuitions (for many examples involving fairness, see Kahneman et al. 1986).

This finding is usually taken to show a problem for standard accounts of rationality. Recently, however, it has been argued that subjects are rationally responding to the information provided, or "leaked," by the speaker's choice of frame (McKenzie 2004). Certainly, the speaker's choice might offer a clue about the desired response; some subjects in the Asian disease problem might be responding to that clue. But even if people are generally taking account of the speaker's clues,[2] that claim is consistent with the proposition that frames matter a great deal to moral intuitions, which is all I am stressing here.

Moral framing has been demonstrated in the important context of obligations to future generations (see Frederick 2003), a much-disputed question of morality, politics, and law (Morrison 1998; Revesz 1999). To say the least, the appropriate discount rate for those yet to be born is not a question that most people have pondered, and hence their judgments are highly susceptible to different frames. From a series of surveys, Maureen Cropper and her coauthors suggest that people are indifferent between saving one life today and saving 45 lives in 100 years (Cropper et al. 1994). They make this suggestion on the basis of questions asking people whether they would choose a program that saves "100 lives now" or a program that saves a substantially larger number "100 years from now." It is possible, however, that people's responses depend on uncertainty about whether people in the future will otherwise die (perhaps technological improvements will save them?); and other ways of framing the same problem yield radically different results (Frederick 2003). For example, most people consider "equally bad" a single death from pollution next year and a single death from pollution in 100 years. This finding implies no preference for members of the current generation. The simplest conclusion is that people's moral judgments about obligations to future generations are very much a product of framing effects (for a similar result, see Baron 2000a).[3]

The same point holds for the question of whether government should consider not only the number of "lives" but also the number of "life-years" saved by regulatory interventions. If the government focuses on life-years, a program that saves children will be worth far more attention that a similar program that saves senior citizens. Is this immoral? People's intuitions depend on how the question is framed (see Sunstein 2004). People will predictably reject an approach that would count every old person as worth "significantly less" than what every young person is worth. But if people are asked whether they would favor a policy that saves 105 old people or 100 young people, many will favor the latter, in a way that suggests a willingness to pay considerable attention to the number of life-years at stake.

At least for unfamiliar questions of morality, politics, and law, people's intuitions are very much affected by framing. Above all, it is effective to frame certain consequences as "losses" from a status quo; when so framed, moral concern becomes significantly elevated. It is for this reason that political actors, and those involved in law, often phrase one or another proposal as "turning back the clock" on some social advance. The problem is that for many social changes, the framing does not reflect social reality, but is simply a verbal manipulation.

Let us now turn to examples that are more controversial.

5. Moral heuristics: A catalogue

My principal interest here is the relationship between moral heuristics and questions of law and policy. I separate

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

the relevant heuristics into four categories: (1) those that involve morality and risk regulation; (2) those that involve punishment; (3) those that involve "playing God," particularly in the domains of reproduction and sex; and (4) those that involve the act-omission distinction. The catalogue is meant to be illustrative rather than exhaustive.

5.1. Morality and risk regulation

5.1.1. Cost–benefit analysis. An automobile company is deciding whether to take certain safety precautions for its cars. In deciding whether or not to do so, it conducts a cost–benefit analysis, in which it concludes that certain precautions are not justified – because, say, they would cost $100 million and save only four lives, and because the company has a "ceiling" of $10 million per life saved (a ceiling that is, by the way, significantly higher than the amount the United States Environmental Protection Agency uses for a statistical life). How will ordinary people react to this decision? The answer is that they will not react favorably (see Viscusi 2000, pp. 547, 558). In fact, they tend to punish companies that base their decisions on cost–benefit analysis, even if a high valuation is placed on human life. By contrast, they impose less severe punishment on companies that are willing to impose a "risk" on people, but that do not produce a formal risk analysis that measures lives lost and dollars, and trades one against another (see Tetlock 2000; Viscusi 2000). The oddity here is that under tort law, it is unclear that a company should be liable at all if it has acted on the basis of a competent cost–benefit analysis; such an analysis might even insulate a company from a claim of negligence. What underlies people's moral judgments, which are replicated in actual jury decisions (Viscusi 2000)?

It is possible that when people disapprove of trading money for lives, they are generalizing from a set of moral principles that are generally sound, and even quite useful, but that work poorly in some cases. Consider the following moral principle: *Do not knowingly cause a human death*. In ordinary life, you should not engage in conduct with the knowledge that several people will die as a result. If you are playing a sport or working on your yard, you ought not to continue if you believe that your actions will kill others. Invoking that idea, people disapprove of companies that fail to improve safety when they are fully aware that deaths will result. By contrast, people do not disapprove of those who fail to improve safety while believing that there is a "risk" but appearing not to know, for certain, that deaths will ensue. When people object to risky action taken after cost–benefit analysis, it seems to be partly because that very analysis puts the number of expected deaths squarely "on screen" (see Tetlock 2000).

Companies that fail to do such analysis, but that are aware that a "risk" exists, do not make clear, to themselves or to anyone else, that they caused deaths with full knowledge that this was what they were going to do. People disapprove, above all, of companies that cause death knowingly. There may be a kind of "cold-heart heuristic" here: Those who know they will cause a death, and do so anyway, are regarded as cold-hearted monsters.[4] In this view, critics of cost–benefit analysis should be seen as appealing to System I, and as speaking directly to the homunculus: "is a corporation or public agency that endangers us to be pardoned for its sins once it has spent $6.1 million per statistical life on risk reduction?" (Ackerman & Heinzerling 2004).

Note that it is easy to reframe a probability as a certainty and vice versa; if I am correct, the reframing is likely to have large effects. Consider two cases:

(a) Company A knows that its product will kill ten people. It markets the product to its ten million customers with that knowledge. The cost of eliminating the risk would have been $100 million.

(b) Company B knows that its product creates a one in one million risk of death. Its product is used by ten million people. The cost of eliminating the risk would have been $100 million.

I have not collected data, but I am willing to predict that Company A would be punished more severely than Company B, even though there is no difference between them.

I suggest, then, that a moral heuristic is at work, one that imposes moral condemnation on those who knowingly engage in acts that will result in human deaths. And of course this heuristic does a great deal of good. The problem is that it is not always unacceptable to cause death knowingly, at least if the deaths are relatively few and an unintended byproduct of generally desirable activity. When government allows new highways to be built, it knows that people will die on those highways; when government allows new coal-fired power plants to be built, it knows that some people will die from the resulting pollution; when companies produce tobacco products, and when government does not ban those products, hundreds of thousands of people will die; the same is true for alcohol. Of course it would make sense, in all of these domains, to take extra steps to reduce risks. But that proposition does not support the implausible claim that we should disapprove, from the moral point of view, of any action taken when deaths are foreseeable.

There is a complementary possibility, involving the confusion between the ex ante and ex post perspectives. If a life might have been saved by a $50 expenditure on a car, people are going to be outraged, and they will impose punishment. What they will not see or incorporate is the fact, easily perceived ex ante, that the $50-per-car expenditure would have been wasted on millions of other people. It is hardly clear that the ex ante perspective is always preferable. But something has gone badly wrong if the ex post perspective leads people to neglect the tradeoffs that are actually involved.

I believe that it is impossible to vindicate, in principle, the widespread social antipathy to cost–benefit balancing.[5] But here too, "a little homunculus in my head continues to jump up and down, shouting at me" that corporate cost–benefit analysis, trading dollars for a known number of deaths, is morally unacceptable. The voice of the homunculus, I am suggesting, is not reflective, but is instead a product of System I, and a crude but quite tenacious moral heuristic.

5.1.2. Emissions trading. In the last decades, those involved in enacting and implementing environmental law have experimented with systems of "emissions trading" (Sunstein 2002). In those systems, polluters are typically given a license to pollute a certain amount, and the licenses can be traded on the market. The advantage of emissions trading systems is that if they work well, they will ensure emissions reductions at the lowest possible cost.

Is emissions trading immoral? Many people believe so. Political theorist Michael Sandel, for example, urges that trading systems "undermine the ethic we should be trying

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

to foster on the environment" (Sandel 1997; see also Kelman 1981). Sandel contends:

> [T]urning pollution into a commodity to be bought and sold removes the moral stigma that is properly associated with it. If a company or a country is fined for spewing excessive pollutants into the air, the community conveys its judgment that the polluter has done something wrong. A fee, on the other hand, makes pollution just another cost of doing business, like wages, benefits and rent.

In the same vein, Sandel objects to proposals to open carpool lanes to drivers without passengers who are willing to pay a fee. Here, as in the environmental context, it seems unacceptable to permit people to do something that is morally wrong as long as they are willing to pay for the privilege.

I suggest that, like other critics of emissions trading programs, Sandel is using a moral heuristic; in fact, he has been fooled by his homunculus. The heuristic is this: *People should not be permitted to engage in moral wrongdoing for a fee*. You are not allowed to assault someone as long as you are willing to pay for the right to do so; there are no tradable licenses for rape, theft, or battery. The reason is that the appropriate level of these forms of wrongdoing is zero (putting to one side the fact that enforcement resources are limited; if they were unlimited, we would want to eliminate, not merely to reduce, these forms of illegality). But pollution is an altogether different matter. At least some level of pollution is a byproduct of desirable social activities and products, including automobiles and power plants. Of course, certain acts of pollution, including those that amount to the intentional or reckless infliction of harm, are morally wrong; but the same cannot be said of pollution as such. When Sandel objects to emissions trading, he is treating pollution as equivalent to a crime in a way that overgeneralizes a moral intuition that makes sense in other contexts. There is no moral problem with emissions trading as such. The insistent objection to emissions trading systems stems from a moral heuristic.

Unfortunately, that objection has appeared compelling to many people, so much as to delay and to reduce the use of a pollution reduction tool that is, in many contexts, the best available (Sunstein 2002). Here, then, is a case in which a moral heuristic has led to political blunders, in the form of policies that impose high costs for no real gain.

5.1.3. Betrayals. To say the least, people do not like to be betrayed. A betrayal of trust is likely to produce a great deal of outrage. If a babysitter neglects a child or if a security guard steals from his employer, people will be angrier than if the identical acts are performed by someone in whom trust has not been reposed. So far, perhaps, so good: When trust is betrayed, the damage is worse than when an otherwise identical act has been committed by someone who was not a beneficiary of trust. And it should not be surprising that people will favor greater punishment for betrayals than for otherwise identical crimes (see Koehler & Gershoff 2003). Perhaps the disparity can be justified on the ground that the betrayal of trust is an independent harm, one that warrants greater deterrence and retribution – a point that draws strength from the fact that trust, once lost, is not easily regained. A family robbed by its babysitter might well be more seriously injured than a family robbed by a thief. The loss of money is compounded and possibly dwarfed by the violation of a trusting relationship. The consequence of the violation might also be more serious. Will the family ever feel entirely comfortable with babysitters? It is bad to have an unfaithful spouse, but it is even worse if the infidelity occurred with your best friend, because that kind of infidelity makes it harder to have trusting relationships with friends in the future.

In this light, it is possible to understand why betrayals produce special moral opprobrium and (where the law has been violated) increased punishment. But consider a finding that is much harder to explain: *People are especially averse to risks of death that come from products (like airbags) designed to promote safety* (Koehler & Gershoff 2003). The aversion is so great that people have been found to prefer a higher chance of dying, as a result of accidents from a crash, to a significantly lower chance of dying in a crash as a result of a malfunctioning airbag. The relevant study involved two principal conditions. In the first, people were asked to choose between two equally priced cars, Car A and Car B. According to crash tests, there was a 2% chance that drivers of Car A, with Air Bag A, will die in serious accidents as a result of the impact of the crash. With Car B, and Air Bag B, there was a 1% chance of death, but also an additional chance of one in 10,000 (0.01%) of death as a result of deployment of the air bag. Similar studies involved vaccines and smoke alarms.

The result was that most participants (over two-thirds) chose the higher risk safety option when the less risky one carried a "betrayal risk." A control condition demonstrated that people were not confused about the numbers: when asked to choose between a 2% risk and a 1.01% risk, people selected the 1.01% risk so long as betrayal was not involved. In other words, people's aversion to betrayals is so great that they will increase their own risks rather than subject themselves to a (small) hazard that comes from a device that is supposed to increase safety. "Apparently, people are willing to incur greater risks of the very harm they seek protection from to avoid the mere possibility of betrayal" (Koehler & Gershoff 2003, p. 244). Remarkably, "betrayal risks appear to be so psychologically intolerable that people are willing to double their risk of death from automobile crashes, fires, and diseases to avoid incurring a small possibility of death by safety device betrayal."

What explains this seemingly bizarre and self-destructive preference? I suggest that a heuristic is at work: *Punish, and do not reward, betrayals of trust*. The heuristic generally works well. But it misfires in some cases, as when those who deploy it end up increasing the risks they themselves face. An airbag is not a security guard or a babysitter, endangering those whom they have been hired to protect. It is a product, to be chosen if and only if it decreases aggregate risks. If an airbag makes people safer on balance, it should be used, even if in a tiny percentage of cases it will create a risk that would not otherwise exist. People's unwillingness to subject themselves to betrayal risks, in circumstances in which products are involved and they are increasing their likelihood of death, is the moral cousin to the use of the representativeness heuristic in the Linda case. Both stem from a generally sound rule of thumb that leads to systematic errors.

In a sense, the special antipathy to betrayal risks might be seen to involve not a moral heuristic but a taste. In choosing products, people are not making purely moral judgments; they are choosing what they like best, and it just turns out that a moral judgment, involving antipathy to be-

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

trayals, is part of what they like best. It would be useful to design a purer test of moral judgments, one that would ask people not about their own safety but about that of others – for example, whether people are averse to betrayal risks when they are purchasing safety devices for their friends or family members. There is every reason to expect that it would produce substantially identical results to those in the experiments just described. Closely related experiments support that expectation (see Ritov & Baron 2002, p. 168). In deciding whether to vaccinate their children from risks for serious diseases, people show a form of "omission bias." Many people are more sensitive to the risk of the vaccination than to the risk from diseases – so much so that they will expose their children to a greater risk from "nature" than from the vaccine. (There is a clear connection between omission bias, trust in nature, and antipathy to "playing God," as discussed later.) The omission bias, I suggest, is closely related to people's special antipathy to betrayals. It leads to moral errors, in the form of vaccination judgments, and undoubtedly others, by which some parents increase the fatality risks faced by their own children.

5.2. Morality and punishment

5.2.1. Pointless punishment? In the context of punishment, moral intuitions are sometimes disconnected from the consequences of punishment, suggesting that a moral heuristic may well be at work (see Darley et al. 2000). Suppose, for example, that a corporation has engaged in serious wrongdoing. People are likely to want to punish the corporation as if it were a person (see Kahneman et al. 1998; Sunstein et al. 2002). They are unlikely to inquire into the possibility that the consequences of serious punishment (say, a stiff fine) will not be to "hurt" corporate wrongdoers, but instead to decrease wages, increase prices, or produce lost jobs. Punishment judgments are rooted in a simple heuristic, to the effect that penalties should be a proportional response to the outrageousness of the act. In thinking about punishment, people use an *outrage heuristic* (see Kahneman & Frederick 2002, pp. 49, 63). According to this heuristic, people's punishment judgments are a product of their outrage. This heuristic may produce reasonable results much of the time, but in some cases, it seems to lead to systematic errors – at least if we are willing to embrace weak consequentialism.

Consider, for example, an intriguing study of people's judgments about penalties in cases involving harms from vaccines and birth control pills (Baron & Ritov 1993). In one case, subjects were told that the result of a higher penalty would be to make companies try harder to make safer products. In an adjacent case, subjects were told that the consequence of a higher penalty would be to make the company more likely to stop making the product, with the result that less safe products would be on the market. Most subjects, including a group of judges, gave the same penalties in both cases. "Most of the respondents did not seem to notice the incentive issue" (see Baron 1993a, pp. 108, 123). In another study, people said that they would give the same punishment to a company that would respond with safer products and one that would be unaffected because the penalty would be secret and covered by insurance (whose price would not increase) (Baron 1993a). Here, too, the effects of the punishment did not affect judgments by a majority of respondents.

A similar result emerged from a test of punishment judgments that asked subjects, including judges and legislators, to choose penalties for dumping hazardous waste (Baron et al. 1993). In one case, the penalty would make companies try harder to avoid waste. In another, the penalty would lead companies to cease making a beneficial product. Most people did not penalize companies differently in the two cases. Most strikingly, people preferred to require companies to clean up their own waste, even if the waste did not threaten anyone, instead of spending the same amount to clean up far more dangerous waste produced by another, now-defunct company.

How could this preference make sense? Why should a company be asked to engage in a course of action that costs the same but that does much less good? In these cases, it is most sensible to think that people are operating under a heuristic, mandating punishment that is proportional to outrageousness, and requiring that punishment be based not at all on consequential considerations. As a general rule, of course, it is plausible to think that penalties should be proportional to the outrageousness of the act; utilitarians will accept the point as a first approximation, and retributivists will insist on it. But it seems excessively rigid to adopt this principle whether or not the consequence would be to make human beings safer and healthier. Weak consequentialists, while refusing to reject retributivism, will condemn this excessive rigidity. Those who seek proportional punishments might well disagree in principle. But it would be worthwhile for them to consider the possibility that they have been tricked by a heuristic – and that their reluctance to acknowledge the point is a product of the insistent voice of their own homunculus.

5.2.2. Probability of detection. Now we turn to some closely related examples from the domain of punishment. On the economic account, the state's goal, when imposing penalties for misconduct, is to ensure optimal deterrence (for this point and some complexities, see Polinsky & Shavell 1998). To increase deterrence, the law might increase the *severity* of punishment, or instead increase the *likelihood* of punishment. A government that lacks substantial enforcement resources might impose high penalties, thinking that it will produce the right deterrent "signal" in light of the fact that many people will escape punishment altogether. A government that has sufficient resources might impose a lower penalty, but enforce the law against all or almost all violators. These ideas lead to a simple theory in the context of punitive damages for wrongdoing: The purpose of such damages is to make up for the shortfall in enforcement. If injured people are 100% likely to receive compensation, there is no need for punitive damages. If injured people are 50% likely to receive compensation, those who bring suit should receive a punitive award that is twice the amount of the compensatory award. The simple exercise in multiplication will ensure optimal deterrence.

But there is a serious question of whether people accept this account, and if not, why not. (For the moment, let us put to one side the question of whether they should accept it in principle.) Experiments suggest that people reject optimal deterrence and that they do not believe that the probability of detection is relevant to punishment. The reason is that they use the outrage heuristic. I participated in two experiments designed to cast light on this question (Sunstein et al. 2000). In the first experiment, subjects were given cases of

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

wrongdoing, arguably calling for punitive damages, and also were provided with explicit information about the probability of detection. Different subjects saw the same case, with only one difference: the probability of detection was substantially varied. Subjects were asked about the amount of punitive damages that they would choose to award. The goal was to see if subjects would impose higher punishments when the probability of detection was low. In the second experiment, subjects were asked to evaluate judicial and executive decisions made to reduce penalties when the probability of detection was high, and to increase penalties when the probability of detection was low. Subjects were questioned about whether they approved or disapproved of varying the penalty with the probability of detection.

The findings were simple and straightforward. The first experiment found that varying the probability of detection had no effect on punitive awards. Even when people's attention was explicitly directed to the probability of detection, they were indifferent to it. The second experiment found that strong majorities of respondents rejected judicial decisions to reduce penalties because of a high probability of detection – and also rejected executive decisions to increase penalties because of a low probability of detection. In other words, people did not approve of an approach to punishment that would make the level of punishment vary with the probability of detection. What apparently concerned them was the extent of the wrongdoing and the right degree of moral outrage – not optimal deterrence.

Of course many people have principled reasons for embracing retributivism and for rejecting utilitarian accounts of punishment. And some such people are likely to believe, on reflection, that the moral intuitions just described are correct – that what matters is what the defendant did, not whether his action was likely to be detected. But if we embrace weak consequentialism, we will find it implausible to suggest that the aggregate level of misconduct is *entirely* irrelevant to punishment. We will be unwilling to ignore the fact that if a legal system refuses to impose enhanced punishment on hard-to-detect wrongdoing, it will end up with a great deal of wrongdoing. People's unwillingness to take any account of the probability of detection suggests the possibility that a moral heuristic is at work, one that leads to real errors. Because of the contested nature of the ethical issues involved, I cannot demonstrate this point; but those who refuse to consider the probability of detection might consider the possibility that System I has gotten the better of them.

5.3. Playing God: Reproduction, nature, and sex

Issues of reproduction and sexuality are prime candidates for the operation of moral heuristics. Consider human cloning, which most Americans reject and believe should be banned. Notwithstanding this consensus, the ethical and legal issues here are extremely difficult. To make progress, it is necessary to distinguish between reproductive and nonreproductive cloning; the first is designed to produce children, whereas the second is designed to produce cells for therapeutic use. Are the ethical issues different in the two cases? In any case, it is important to identify the particular grounds for moral concern. Do we fear that cloned children would be means to their parents' ends, and if so, why? Do we fear that they would suffer particular psychological harm? Do we fear that they would suffer from especially severe physical problems?

In a highly influential discussion of new reproductive technologies, above all cloning, ethicist Leon Kass (1998, pp. 17–19) points to the "wisdom in repugnance." Kass writes:

> People are repelled by many aspects of human cloning. They recoil from the prospect of mass production of human beings, with large clones of look-alikes, compromised in their individuality, the idea of father-son or mother-daughter twins; the bizarre prospects of a woman giving birth to and rearing a genetic copy of herself, her spouse or even her deceased father or mother; the grotesqueness of conceiving a child as an exact replacement for another who has died; the utilitarian creation of embryonic genetic duplicates of oneself, to be frozen away or created when necessary, in case of need for homologous tissues or organs for transplantation; the narcissism of those who would clone themselves and the arrogance of others who think they know who deserves to be cloned or which genotype any child-to-be should be thrilled to receive; the Frankensteinian hubris to create human life and increasingly to control its destiny; man playing God . . . We are repelled by the prospect of cloning human beings not because of the strangeness or novelty of the undertaking, but because we intuit and feel, immediately and without argument, the violation of things that we rightfully hold dear. . . . Shallow are the souls that have forgotten how to shudder.

Kass is correct to suggest that revulsion toward human cloning might be grounded in legitimate concerns, and I mean to be agnostic here on the question of whether human cloning is ethically defensible. But I want to suggest that moral heuristics, and System I, are responsible for what Kass seeks to celebrate as that which "we intuit and feel, immediately and without argument." In fact Kass's catalogue of alleged errors seems to me to be an extraordinary exercise in the use of such heuristics. Availability operates in this context, not to drive judgments about probability, but to call up instances of morally dubious behavior (e.g., "mass production of human beings, with large clones of look-alikes, compromised in their individuality"). The representativeness heuristic plays a similar role (e.g., "the Frankensteinian hubris to create human life and increasingly to control its destiny"). But I believe that Kass gets closest to the cognitive process here with three words: "man playing God."

In fact, we might well think that "do not play God" is the general heuristic here, with different societies specifying what falls in that category and with significant changes over time. Even in secular societies, a closely related heuristic plays a large role in judgments of fact and morality: *Do not tamper with nature*. This heuristic affects many moral judgments, though individuals and societies often become accustomed to various kinds of tampering (consider in vitro fertilization). An anti-tampering heuristic helps explain many risk-related judgments. For example, "[h]uman intervention seems to be an amplifier in judgments on food riskiness and contamination," even though "more lives are lost to natural than to man-made disasters in the world" (Rozin 2001, pp. 31, 38). Studies show that people overestimate the carcinogenic risk from pesticides and underestimate the risks of natural carcinogens (Rozin 2001). People also believe that nature implies safety, so much so that they will prefer natural water to processed water even if the two are chemically identical (Rozin 2001).

The moral injunction against tampering with nature plays a large role in public objections to genetic engineering of food, and hence legal regulation of such engineering is sometimes driven by that heuristic rather than by a de-

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

liberative, System II encounter with the substantive issues. For genetic engineering, the anti-tampering heuristic drives judgments even when the evidence of risk is slim (McHughen 2000). In fact, companies go to great lengths to get a "natural" stamp on their products (Schlosser 2002), even though the actual difference between what counts as a "natural additive" and an "artificial additive" bears little or no relation to potential harm to consumers. So too, in the domains of reproduction and sexuality, where a pervasive objection is that certain practices are "unnatural." And for cloning, there appears to be a particular heuristic at work: *Do not tamper with natural processes for human reproduction*. It is not clear that this heuristic works well; but it is clear that it systematically misfires.

Issues at the intersection of morality and sex provide an obvious place for the use of moral heuristics. Such heuristics are peculiarly likely to be at work in any area in which people are likely to think, "That's disgusting!" Any examples here will be contentious, but let us return to the incest taboo. We can easily imagine incestuous relationships, say between first cousins or second cousins, which ought not to give rise to social opprobrium but which might nonetheless run afoul of social norms or even the law (Haidt 2001). The incest taboo is best defended by reference to coercion, psychological harm, and risks to children who might result from incestuous relationships. But in many imaginable cases, these concrete harms are not involved.

Of course, it is plausible to say that the best way to defend against these harms is by a flat prohibition on incest, one that has the disadvantage of excessive generality but the advantage of easy application. Such a flat prohibition might have evolutionary origins (Stein 2001); it might also have strong rule-utilitarianism justifications. We would not like to have family members asking whether incest would be a good idea in individual cases, even if our underlying concern is limited to coercion and psychological harm. So defended, however, the taboo stands unmasked as a moral heuristic. In this vein, Haidt and his coauthors (Haidt et al. 2004) refer to "moral dumbfounding" – to the existence of moral judgments that people "feel" but are unable to justify. In the domain of sex and reproduction, many taboos can be analyzed in similar terms.

5.4. Acts and omissions

To say the least, there has been much discussion of whether and why the distinction between acts and omissions might matter for morality, law, and policy. In one case, for example, a patient might ask a doctor not to provide life-sustaining equipment, thus ensuring the patient's death. In another case, a patient might ask a doctor to inject a substance that will immediately end the patient's life. Many people seem to have a strong moral intuition that the failure to provide life-sustaining equipment, and even the withdrawal of such equipment, is acceptable and legitimate – but that the injection is morally abhorrent. And indeed, American constitutional law reflects judgments to exactly this effect: People have a constitutional right to withdraw equipment that is necessary to keep them alive, but they have no constitutional right to physician-assisted suicide (see Washington vs. Glucksberg 1997, pp. 724–25). But what is the morally relevant difference?

It is worth considering the possibility that the act–omission distinction operates as a heuristic for a more complex and difficult assessment of the moral issues at stake. From the moral point of view, harmful acts are generally worse than harmful omissions, in terms of both the state of mind of the wrongdoer and the likely consequences of the wrong. A murderer is typically more malicious than a bystander who refuses to come to the aid of someone who is drowning; the murderer wants his victim to die, whereas the bystander need have no such desire. In addition, a murderer typically guarantees death, whereas a bystander may do no such thing. (I put to one side some complexities about causation.) But in terms of either the wrongdoer's state of mind or the consequences, harmful acts are not *always* worse than harmful omissions. The moral puzzles arise when life, or a clever interlocutor, comes up with a case in which there is no morally relevant distinction between acts and omissions, but when moral intuitions (and the homunculus) strongly suggest that there must be such a difference. As an example, consider the vaccination question discussed earlier; many people show an omission bias, favoring inaction over statistically preferable action (Baron & Ritov 1993). Here an ordinarily sensible heuristic, favoring omissions over actions, appears to produce moral error.

In such cases, we might hypothesize that moral intuitions reflect an overgeneralization of principles that usually make sense – but that fail to make sense in the particular case (see Baron 1994a). Those principles condemn actions but permit omissions, a difference that is often plausible in light of relevant factors, but that, in hard cases, cannot be defended (but see Kamm 1998). I believe that the persistent acceptance of withdrawal of life-saving equipment, alongside persistent doubts about euthanasia, is a demonstration of the point. There is no morally relevant difference between the two; the act–omission distinction makes a difference apparent or even clear when it is not real (on some complications regarding this, see Sunstein 1999).

This point cannot be demonstrated here; further experiments on the nature of moral intuitions in this domain would be extremely valuable (for an illustration, see Haidt & Baron 1996). But compare the dispute over two well-known problems in moral philosophy (see Thomson 1986, pp. 94–116). These problems do not involve the act–omission distinction; no omission is involved. But the problems implicate closely related concerns. The first, called the trolley problem, asks people to suppose that a runaway trolley is headed for five people, who will be killed if the trolley continues on its current course. The question is whether you would throw a switch that would move the trolley onto another set of tracks, killing one person rather than five. Most people would throw the switch. The second, called the footbridge problem, is the same as that just given, but with one difference: the only way to save the five is to throw a stranger, now on a footbridge that spans the tracks, into the path of the trolley, killing that stranger but preventing the trolley from reaching the others. Most people will not kill the stranger. But what is the difference between the two cases, if any? A great deal of philosophical work has been done on this question, much of it trying to suggest that our firm intuitions can indeed be defended in principle.

Without engaging these arguments, let me suggest the possibility of a simpler answer. As a matter of principle, there is no difference between the two cases. People's different reactions are based on moral heuristics that condemn the throwing of the stranger but support the throwing of the switch. As a matter of principle, it is worse to

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

throw a human being in the path of a trolley than to throw a switch that (indirectly?) leads to a death. The relevant heuristics generally point in the right direction. To say the least, it is desirable for people to act on the basis of a moral heuristic that makes it extremely abhorrent to throw innocent people to their death. But the underlying heuristics misfire in drawing a distinction between the two cleverly devised cases. Hence, people struggle heroically to rescue their intuitions and to establish that the two cases are genuinely different in principle. But they aren't. In this sense, a moral heuristic, one that stems from System I and has "ecological rationality" (Gigerenzer 2000), leads to errors. And this objection does not bear only on ingeniously devised hypothetical cases. It suggests that a moral mistake pervades both commonsense morality and law, including constitutional law, by treating harmful omissions as morally unproblematic or categorically different from harmful actions.

Is there anything to be said to those who believe that their moral judgments, distinguishing the trolley and footbridge problems, are entirely reflective, and embody no heuristic at all? Consider a suggestive experiment designed to see how the human brain responds to the two problems (Greene et al. 2001). The authors do not attempt to answer the moral questions in principle, but they find "that there are systematic variations in the engagement of emotions in moral judgment," and that brain areas associated with emotion are far more active in contemplating the footbridge problem than in contemplating the trolley problem. An implication of Greene et al.'s finding is that human brains are hard-wired to distinguish between bringing about a death "up close and personal" and doing so at a distance. Of course, this experiment is far from decisive; emotions and cognition are not easily separable (Nussbaum 2002), and there may be good moral reasons why certain brain areas are activated by one problem and not by the other. Perhaps the brain is closely attuned to morally irrelevant differences. But consider the case of fear, where an identifiable region of the brain makes helpfully immediate but not entirely reliable judgments (Ledoux 1996), in a way that suggests a possible physical location for some of the operations of System I. The same may well be true in the context of morality, politics, and law (Greene & Haidt 2002).[6]

6. Exotic cases, moral judgments, and reflective equilibrium

Some of these examples will seem more contentious than others. But taken as a whole, they seem to me to raise serious doubts about the wide range of work that approaches moral and political dilemmas by attempting to uncover moral intuitions about exotic cases of the kind never or rarely encountered in ordinary life. Should you shoot an innocent person if that is the only way to save twenty innocent people (Williams 1973)? What is the appropriate moral evaluation of a case in which a woman accidentally puts cleaning fluid in her coffee, and her husband, wanting her dead, does not provide the antidote, which he happens to have handy (see Thomson 1986, p. 31)? If Martians arrived and told you that they would destroy the world unless you tortured a small child, should you torture a small child? Is there a difference between killing someone by throwing him into the path of a train and killing someone by diverting the train's path to send it in his direction?

I believe that in cases of this kind, the underlying moral intuitions ordinarily work well, but that when they are wrenched out of familiar contexts, their reliability, for purposes of moral and legal analysis, is unclear. Consider the following rule: *Do not kill an innocent person, even if this is necessary to save others*. (I put to one side the contexts of self-defense and war.) In all likelihood, a society does much better if most people have this intuition, if only because judgments about necessity are likely to be unreliable and self-serving. But in a hypothetical case, in which it really is necessary to kill an innocent person to save twenty others, our intuitions might well turn out to be unclear and contested – and if our intuitions about the hypothetical case turn out to be very firm (do not kill innocent people, ever!), they might not deserve to be so firm, simply because they have been wrenched out of the real-world context, which is where they need to be to make sense.

The use of exotic cases has been defended, not on the ground that they are guaranteed to be correct, but as a means of eliciting the structure of our moral judgments in a way that enables us to "isolate the reasons and principles" that underlie our responses (Kamm 1993, p. 8; see generally, Sorenson 1992). But if those responses are unreliable, they might not help to specify the structure of moral judgments, except when they are ill-informed and unreflective. For isolating reasons and principles that underlie our responses, exotic cases might be positively harmful (cf. Flanagan 1993).

In short, I believe that some philosophical analysis, based on exotic moral dilemmas, is inadvertently and even comically replicating the early work of Kahneman and Tversky by uncovering situations in which intuitions, normally quite sensible, turn out to misfire. The irony is that where Kahneman and Tversky meant to devise cases that would demonstrate the misfiring, some philosophers develop exotic cases with the thought that the intuitions are likely to be reliable and should form the building blocks for sound moral judgments. An understanding of the operation of heuristics offers reason to doubt the reliability of those intuitions, even when they are very firm (cf. the emphasis on moral learning from real-world situations in Churchland 1996).

Now, it is possible that the firmness of the underlying intuitions is actually desirable. Perhaps social life is better, not worse, because of the large number of people who treat heuristics as moral rules and who believe (for example) that innocent people should never be killed. If the heuristic is treated as a universal and freestanding principle, perhaps some mistakes will be made, but only in highly unusual cases, and perhaps people who accept the principle will avoid the temptation to depart from it when the justification for doing so appears sufficient, but really is not. In other words, a firm rule might misfire in some cases, but it might be better than a more fine-grained approach, which, in practice, would misfire even more. Those who believe that you should always tell the truth may do and be much better, all things considered, than those who believe that the truth should be told only on the basis of case-specific, all-things-considered judgments in its favor.

To the extent that moral heuristics operate as rules, they might be defended in the way that all rules are – as much better than the alternatives, even if they produce errors in some imaginable cases. I have noted that moral heuristics might show a kind of "ecological rationality," thus working well in most real-world contexts (Gigerenzer 2000); recall

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

the possibility that human beings live by simple heuristics that make us good. My suggestion is not that the moral heuristics, in their most rigid forms, are socially worse than the reasonable alternatives. It is hard to resolve that question in the abstract. I am claiming only that such heuristics lead to real errors and significant confusion. Of course, a great deal of experimental work remains to be done on this question; existing research has only scratched the surface.

Within philosophy, there is a large literature on the role of intuitions in moral argument, much of it devoted to their role in the search for reflective equilibrium (Hooker 2000; Raz 1994). In John Rawls' influential formulation, people's judgments about justice should be made via an effort to ensure principled consistency between their beliefs at all levels of generality (Rawls 1971). Rawls emphasizes that during the search for reflective equilibrium, all beliefs are revisable in principle. But as Rawls also emphasizes, some of our beliefs, about particular cases and more generally, seem to us especially fixed, and it will take a great deal to uproot them. It is tempting to use an understanding of moral heuristics as a basis for challenging the search for reflection equilibrium, but I do not believe that anything said here supports that challenge (see Pizarro & Bloom 2003, emphasizing the potential role of conscious deliberation in informing and reshaping our moral intuitions). Recall that in Rawls' formulation, all of our intuitions are potentially revisable, including those that are quite firm.

What I am adding here is that if moral heuristics are pervasive, then some of our apparently fixed beliefs might result from them. We should be aware of that fact in attempting to reach reflective equilibrium. Of course, some beliefs that are rooted in moral heuristics might turn out, on reflection, to be correct, perhaps for reasons that will not occur to people who use the heuristics mechanically. I am suggesting only that judgments that seem most insistent, or least revisable, may result from overgeneralizing intuitions that work well in many contexts, but that also misfire (see the discussion of wide and narrow reflective equilibrium in Stein 1996).

If this is harder to demonstrate in the domain of morality than in the domain of facts, it is largely because we are able to agree, in the relevant cases, about what constitutes factual error, and are often less able to agree about what constitutes moral error. With respect to the largest disputes about what morality requires, it may be too contentious to argue that one side is operating under a heuristic, whereas another side has it basically right. But I hope that I have said enough to show that in particular cases, sensible rules of thumb lead to demonstrable errors not merely in factual judgments, but in the domains of morality, politics, and law, as well.

ACKNOWLEDGMENTS

I am grateful to Daniel Kahneman and Martha Nussbaum for valuable discussions. For helpful comments on a previous draft, I also thank participants in a seminar at Cambridge University, Jonathan Baron, Mary Anne Case, Elizabeth Emens, Robert Frank, Robert Goodin, Jonathan Haidt, Steven Pinker, Peter Singer, Edward Stein, and a number of anonymous reviewers.

NOTES

1. In a widely held view, a primary task of ethics is to identify the proper general theory and to use it to correct intuitions in cases in which they go wrong (Hooker 2000). Consider here the provocative claim that much of everyday morality, nominally concerned with fairness, should be seen as a set of heuristics for the real issue, which is how to promote utility (see Baron 1998; to the same general effect, with numerous examples from law, see Kaplow & Shavell 2002).

2. Note also that loss aversion is quite robust in the real world (Benarzi & Thaler 2000; Camerer 2000), and it has not been shown to be solely or mostly a result of the speaker's clues. Also note that the nature of the clue, when there is one, depends on the speaker's appreciation of the existence of framing effects – otherwise the clue would be ineffective.

3. Here too the frame may indicate something about the speaker's intentions, and subjects may be sensitive to the degree of certainty in the scenario (assuming, for example, that future deaths may not actually occur). While strongly suspecting that these explanations are not complete (see Frederick 2003), I mean not to reject them, but only to suggest the susceptibility of intuitions to frames (for skeptical remarks, see Kamm 1998).

4. I am grateful to Jonathan Haidt for this suggestion.

5. I put to one side cases in which those who enjoy the benefits are wealthy and those who incur the costs are poor; in some situations, distributional considerations will justify a departure from what would otherwise be compelled by cost–benefit analysis (on this and other problems with cost–benefit analysis, see Sunstein 2002).

6. To see the implications, consider the controversial area of capital punishment, and let us simply assume that for each execution, at least five murders are prevented (see the Dezhbakhsh et al. [2004] finding that each execution prevents eighteen murders, with a margin of error of ten). If the assumption is correct, the refusal to impose capital punishment will effectively condemn numerous innocent people to death. Many people think that capital punishment counts as an "act," while to refuse to impose it counts as an "omission," and that the two are morally different. Many others point to differences in the nature of the causal chains, which might make capital punishment unacceptable even if the failure to impose it leads to the death of innocent people. I cannot resolve the moral issues in this space. But for weak consequentialists, it is at least worth considering the possibility that if capital punishment deters large numbers of murders, then it cannot be so easily condemned on moral grounds – at least if we do not employ an act–omission distinction in a context in which that distinction might be difficult to defend in principle.

Similar issues are raised by the debate over torture. If torture would prevent the death of many innocent people, or if torture would prevent many other tortures, might not a ban on torture be seen as a moral heuristic, one that misfires in imaginable cases?

Open Peer Commentary

Cognitivism, controversy, and moral heuristics

Matthew D. Adler

University of Pennsylvania Law School, Philadelphia, PA 19104.
madler@law.upenn.edu

Abstract: Sunstein aims to provide a nonsectarian account of moral heuristics, yet the account rests on a controversial meta-ethical view. Further, moral theorists who reject act consequentialism may deny that Sunstein's examples involve moral mistakes. But so what? Within a theory that counts consequences as a morally weighty feature of actions, the moral judgments that Sunstein points to are indeed mistaken, and the fact that governmental action at odds with these judgments will be controversial doesn't bar such action.

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

What are moral heuristics? Noncognitivists and cognitivists will answer this question differently. Noncognitivists deny that there are moral truths, or facts, or that moral statements express beliefs. Rather, these statements express the speaker's feelings, commitments, or other such conative attitudes. The debate about cognitivism remains a live one within meta-ethics, animated by two deep truisms about morality: that there are moral disagreements, and that moral judgments prompt action (Miller 2003). Noncognitivists have trouble with the first truism (How can there be disagreements about nonfactual matters?), but cognitivists have trouble with the second (Since beliefs, alone, don't prompt action, how can moral judgments?).

Sunstein's approach to moral heuristics is cognitivist. Indeed, he assumes that the construct of a moral heuristic entails cognitivism. "To show that heuristics operate in the moral domain, we have to specify some benchmark by which we can measure moral truth" (target article, sect. 3, para. 1). Yet, consider Sunstein's generic definition of a heuristic as an automatic, unreflective, decision-making process that substitutes a heuristic attribute for a target attribute. Judge J_i reaches a moral judgment about some action or some other object of moral assessment, focusing on heuristic attribute *H* rather than the target attribute *T* that (in some sense) is the "morally relevant" feature of the object. Does the characterization of *T* rather than *H* as morally relevant entail cognitivism? Maybe not. It would seem that the noncognitivist can say something like this: *T* is the feature that J_i would focus on after full, "System II" processing; *H* is the feature that J_i attends to as a result of quick, "System I" processing.

So Sunstein seems to be wrong to think that the general concept of a moral heuristic entails cognitivism; but he is right that his particular conception does so. His cognitivist take on moral heuristics is this: *H* is the feature that J_i focuses on after quick, System I processing, whereas *T* is the feature that is truly morally relevant.

Let us now assume moral cognitivism. This position, albeit meta-ethically controversial, is one that I (and Sunstein) believe to be correct. If there are indeed moral facts and truths, does it follow that there are moral heuristics? It is important to distinguish, here, between intratheoretical and intertheoretical claims about moral heuristics. Different moral theories, such as act consequentialism, or rule consequentialism, or rule contractarianism, or Kantianism, offer different substantive accounts of moral truths. The intratheoretical claim is that, for any plausible moral theory, there can be cases where some J_i as a result of quick processing focuses on a heuristic feature *H* of some assessment object rather than the truly relevant feature *T* of that object identified by the theory. This intratheoretical claim is surely correct. The very point of cognitivism, and particular cognitivist theories, is to delineate moral truths that can depart from, and provide critical purchase on, the actual judgments that an individual engaged in moral assessment happens to reach.

Sunstein argues, however, not merely for the intratheoretical claim, but for the much stronger, intertheoretical claim. His examples (involving framing, risk regulation, punishment, playing God, and the act/omission distinction) are offered as cases where individual judgments are incorrect across a wide range of moral theories – all plausible theories, in the case of framing, and all "weakly consequentialist" theories, in the case of the other heuristics. This claim underestimates, I think, just how diverse moral theories are (Kagan 1998). Moral theorists continue to disagree about the foundations of moral truth. Is this ultimately a matter of producing good consequences (where the notion of "good consequences" might itself be more or less welfarist and more or less egalitarian); of implementing some hypothetical contract; of following universalizable principles; or of reflecting human nature? Moral theorists also continue to disagree about the primary object of moral assessment: acts, rules, virtues, motives, or something else. For example, the act consequentialist thinks that the right action is the action with the best consequences; the rule consequentialist thinks that the right action is the action conforming to the rule with the best consequences.

These intertheoretical disagreements reflect the wide range of preliminary moral judgments that the theorist will have and that any theory attempts to fit: judgments about particular cases, and more systemic judgments about the connection of "morality" to concepts like "welfare," "rights," "autonomy," "equality," "rationality," and so on. The threshold theoretical choice between cognitivism and noncognitivism, and the subsequent adoption of a particular cognitivist theory, is a matter of appropriately synthesizing these preliminary judgments, accepting some and rejecting others. (That is one way of understanding Rawls' claims about "reflective equilibrium"). Given the multiplicity of preliminary judgments, and the fuzziness of the notion of appropriate synthesis, the plurality of moral theories is not surprising. This plurality, in turn, makes Sunstein's intertheoretical claims quite ambitious.

Consider framing. Couldn't framing be the upshot of rule consequentialism, or indeed of any other theory that makes rules rather than actions the primary object of moral assessment? If individuals naturally frame effects as losses or gains from some baseline, then moral loss aversion may economize on deliberation costs relative to full-blown moral deliberation about actions. If the economies are substantial enough, and if the error rate of framing relative to full-blown deliberation is not too high, framing would presumably be required by the optimal rule.

As for the risk regulation, punishment, playing God, and act/omission cases: Sunstein suggests that individual judgments in these instances depart from "weak consequentialism," offered as a "nonsectarian" principle endorsable by a wide range of moral theories (see sects. 3.2 and 5). But there is an ambiguity here. "Weak consequentialism" might mean: (1) the principle that consequences are a morally relevant feature of actions, perhaps lexically subordinate to other, nonconsequentialist factors; or (2) the principle that the consequences of actions are a morally weighty feature, with sufficient force to override nonconsequentialist factors in a significant range of choice situations. Sunstein defines "weak consequentialism" in the first way, as the view that "the social consequences of the legal system are relevant, other things being equal, to what law ought to be doing" (sect. 3.2). Two paragraphs later, he contrasts deontologists "who believe that consequences do not matter at all" with weak consequentialists, reinforcing the first definition of weak consequentialism. But the risk regulation, punishment, playing God, and act/omission cases are all *consistent* with weak consequentialism in this sense. For these are all cases in which putative nonconsequentialist factors, such as imposing proportionate penalties, or expressing respect for various goods, or not tampering with nature, or not actively causing harm, are in play. Only if consequences can sometimes override such factors in determining morally appropriate actions do the cases illustrate moral mistakes. And this second variant of weak consequentialism is substantially more "sectarian" than the first.

Does this matter? Sunstein is willing to bite the bullet of moral controversy at the meta-ethical level. Why not bite it at the substantive level too? Cost–benefit analysis implements weak consequentialism in the second, more sectarian sense: the view that consequences (in particular, overall welfare) are a morally weighty feature of government choices (Adler & Posner 1999). Given a sufficiently robust commitment to cost–benefit analysis, one can say that the risk regulation, punishment, playing God, and act/omission cases illustrate moral mistakes. Such a commitment would be controversial, but that doesn't mean that it is incorrect, or that government shouldn't act on it. Morally conscientious government actors must ultimately settle on a moral theory (or a probability distribution across theories) and choose. The fact that a particular choice will be controversial, flying in the face of other theories and ordinary moral judgments, will itself be handled differently by different theories. Cost–benefit analysis understands controversy as a potential source of discontent with governmental choice (a hedonic cost) and perhaps an obstacle to implementation (an enforcement cost), but these must be balanced against the benefits of controversial choices.

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

Alternative perspectives on omission bias

Christopher J. Anderson

Department of Psychology, Temple University, Philadelphia, PA 19123.
chris.anderson@temple.edu

Abstract: The act/omission distinction is likely to lead to biases and be used as a moral heuristic. However, it is frequently difficult to determine whether this act/omission distinction is responsible for a judgment outside the lab. Further, more encompassing theories of omission bias are needed to make progress in dealing with its harmful consequences. One such theory is briefly presented.

The distinction between actions and omission, Sunstein argues, is used as a heuristic that is often appropriate, but also leads to systematic moral errors. Regarding this, several issues must be considered. First, note that it can be difficult to evaluate the presence of an omission bias from the type of examples often used to elicit moral judgments. Attitudes towards action and omission are not easy to disentangle from status quo versus change, natural versus unnatural, and direct versus indirect cause distinctions; further complicating matters, people sometimes show a bias towards action (Baron & Ritov 2004; Patt & Zeckhauser 2000; Ritov & Baron 1990).

Progress has been made in elucidating these issues about omission bias. However, the point needs to be raised here, because extending omission bias findings to moral, political, and legal issues in which other "confounded" distinctions cannot be controlled, means that it is difficult for us to know whether the act/omission distinction is pertinent to the types of moral intuitions discussed by Sunstein. For example, the aversion to euthanizing by action compared to relative acceptance of euthanizing by omission could easily be driven by feelings about direct causation of harm or by distinguishing between not impeding natural death and causing an unnatural one. Even in the artificial case of the trolley scenario, there are problems, namely, that outcomes are not equal from the active and passive killing, as people are likely to experience more difficulty afterward with their choice if they kill by throwing someone off of a footbridge than if they kill by throwing a switch. The action choice would make a person susceptible to post-traumatic stress disorder (Grossman 1996), so one can hardly equate the outcomes in the trolley problem. Sunstein seems to implicitly recognize this problem with the trolley scenario, noting that humans may be genetically programmed to distinguish between causing death from near or from far.[1]

Regardless of these and other subtleties, which are important to consider when attempting to apply findings of omission bias, let us grant that an omission bias exists in at least some contexts, which the literature suggests (Anderson 2003; Baron & Ritov 2004). If we grant that it is a bias, that this distinction is taken into consideration when it is irrelevant, we must concede that it then causes real harm by leading to moral judgments that, upon further reflection, are considered erroneous.

An important question then becomes: where does this distinction originate and how do preferences become skewed and applied when they are irrelevant? In terms of reducing the harm caused by moral heuristics, understanding the source of the heuristic may prove to be crucial, for "debiasing" by education about a heuristic or bias is not always effective (e.g., Wilson et al. 2002).

Several explanations for omission bias have been proposed. One type of proposal is to embed the distinction between action and omission within another distinction and claim that the bias is subordinate to another bias (e.g., towards the status quo, normality, or indirectly caused over directly caused harm). These proposals, even if they have validity, do not so much address our need for a source, as change the locus of the search for a source.

Another type of proposal suggests that people have different emotional reactions to actions and omissions, and that the moral judgment is a consequence of a more basic emotional response.[2] It has been suggested that enhanced regret for actions over omissions could be the emotional source.[3] I have also suggested that even in the absence of feeling this distinction, the knowledge that others make the distinction would lead one to make the same distinction that others do, to avoid the blame of others for harm from actions. Again, although these explanations may have validity, this shifts the locus for explaining the distinction elsewhere, rather than explaining it in a more satisfying way.

Thus, I wish to suggest a more encompassing explanation which, while speculative, indicates the kind of theory that is needed to make further progress in omission bias research.

The concept behind this proposal is that there is a structural, environmental problem that distorts people's perceptions of actions and omissions. That problem is an asymmetry in the collection and processing of information relating to actions and omissions. In everyday experience, action and omission decisions happen on an ongoing basis. However, people are not aware of the consequences of all of their choices. Perhaps actions tend to focus our attention on consequences, so that consequences that result from action are likely to undergo more elaborated processing and are more likely to be processed at all. Although there is no reason to suppose actions or omissions generally result in more harm, people may come to be of the opinion that actions are more risky because they are more aware of the losses that they have incurred in the past as a result of action. Relative ignorance of those results for inaction might lead to a perception that omission options are actually somewhat safer, leading to a small bias in their favor. If there is validity to this explanation, the omission bias is likely to be recalcitrant, as it would result from an attitude accrued over a large number of experiences.

In summary, I agree with Sunstein that the act/omission distinction is likely used as a moral heuristic in important areas of applied decision-making. However, we must be cautious in applying this research to real-world problems that could have multiple sources of justification and/or intuition feeding a judgment. Furthermore, we do not yet have a firm idea of what causes omission bias, or how to debias it, which will have important implications for how we view and deal with this as a "moral heuristic." The primary challenge facing researchers in this area is to develop and test larger-scale theories that do more than shift the locus of the distinction, explain omission bias in a more complete manner, and lend themselves to effective debiasing strategies.

NOTES

1. If correct, this is likely just an innate preference to avoid killing conspecifics, which is not flexible enough to handle our recently developed ability to kill at a distance (Grossman 1996).

2. For more on this general type of explanation, see Haidt (2001) and Pizarro and Bloom (2003), who provide perspectives on this position.

3. Note that although there is evidence for this emotional effect, it occurs only provisionally (reviewed in Anderson 2003).

Moral heuristics: Rigid rules or flexible inputs in moral deliberation?

Elizabeth Anderson

Department of Philosophy, University of Michigan, Ann Arbor, MI 48109-1003.
eandersn@umich.edu **http://www-personal.umich.edu/~eandersn/**

Abstract: Sunstein represents moral heuristics as rigid rules that lead us to jump to moral conclusions, and contrasts them with reflective moral deliberation, which he represents as independent of heuristics and capable of supplanting them. Following John Dewey's psychology of moral judgment, I argue that successful moral deliberation does not supplant moral heuristics but uses them flexibly as inputs to deliberation. Many of the flaws in moral judgment that Sunstein attributes to heuristics reflect instead the limitations of the deliberative context in which people are asked to render judgments.

Sunstein's theory of moral heuristics continues a tradition of understanding moral judgment advanced by John Dewey. Dewey

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

(1922) argued that habits – what Sunstein calls heuristics and what philosophers call intuitions – govern moral thought in the immediate, nondeliberative, emotionally engaged way Sunstein describes. Like Sunstein, he argued that habits may make sense in the contexts in which they evolved, but misfire when placed in novel contexts. Then we need to consider the consequences of their operation by engaging what Sunstein calls "System II," and thereby arrive at more reflective judgments of what we ought to do. However, Dewey and Sunstein part ways regarding the relationship of reflection to habit. Sunstein, recapitulating a standard reason/emotion dichotomy, suggests that reflection is independent of habit and should simply supplant, without modifying, habit, when the latter misfires. Dewey argued that reflection depends on habit. It is always predicated on "the brute act of holding something dear" (1915, p. 46) – on intuitive valuations provisionally accepted as given, although they may have been modified by earlier reflections entrenched in habit, and may be called into question and further modified at later times. Moral intuitions or habits properly function as inputs to moral reflection rather than as fixed, rigid conclusions.

Dewey's theory suggests an alternative interpretation of the flaws Sunstein identifies in our moral judgments. Sunstein locates these flaws in inherent defects or limitations in our heuristics. Dewey would locate them in the limitations of our deliberative context. Following Dewey's cue, I suggest that Sunstein's evidence for our innate judgmental defects often reflects limitations built into the reflective contexts into which subjects have been placed. These limitations are of two sorts: (1) the deliberative context may not be the right one for solving the problem given to it; or (2) it may lack background information needed to make sense of the other information subjects have. These problems arise most evidently in Sunstein's punishment cases.

Sunstein claims that judges and juries should consider the consequences of punishment in determining how much guilty defendants should be punished. He attributes people's resistance to doing this to the outrage heuristic, which insists that punishment be determined by the gravity of the wrongdoing. I suggest that this resistance is rather a product of the limited role of courtroom deliberation in the division of moral labor. People already know that judges and jurors are assigned a limited task – determining the culpability of this defendant, the punishment this defendant deserves, and the problems this defendant is responsible for correcting. They know that decision-makers in the courtroom setting are deliberately denied information about the consequences of punishment (e.g., whether defendants are insured) so as not to distract them with information irrelevant to the limited task they are supposed to solve. Moreover, courtroom settings are not equipped to handle isolated bits of information about the consequences of punishment. What is a courtroom decision-maker to do with information about the probability of detection, without any information about how much the wrongdoing would decline in the face of increased penalties? Knowing the limited role of courtroom decision-makers and the lack of background information needed to deal with information about consequences, people placed in courtroom roles are rational to disregard this information.

We could imagine courtrooms set up differently, to take account of such information. Jurors in ancient Athens routinely considered the consequences of punishment on innocent third parties when deciding punishments. Defendants found guilty of a crime would drag in their wailing wives, children, and other dependents, who would beg the jurors for mercy so that they would not suffer for the crimes of the household head. If this were the norm for U.S. courtrooms, no doubt jurors would consider the consequences of punishment on innocents in their deliberations. Such consideration would not replace the emotionally grounded outrage heuristic with affectless rational deliberation, but rather enter an additional heuristic, grounded in pity, as an input in juror deliberation.

It is not evident that consequentialists should approve such a modification of our practices of courtroom deliberation. It would dramatically reduce the deterrent power of punishment on people with dependents. It would enable corporations to get away with wrongdoing by choosing to shift the costs of punishment to their employees, or by credibly threatening to leave the market to even more negligent producers. The courtroom is not the right place to deal with such consequences.

Instead, we charge this task to legislatures and administrative agencies, who have broad investigative powers, access to background information needed to put data about consequences to effective use, and facility with sophisticated deliberative tools, including not just cost–benefit analysis but processes for soliciting responses to proposed policies from stakeholders and citizens at large. If holding each firm responsible for its own waste cleanup is inefficient, we can levy a general tax on toxic chemical production and devote the revenues to cleaning up polluted sites. This administrative solution is preferable to relying on courts to decide who should clean up which sites. If safer vaccine makers would exit the market under certain liability rules, the state can indemnify them from lawsuits, provided they meet specified safety standards. Such "System II" deliberative contexts do not override, but rather incorporate, the force of emotionally engaged moral heuristics. The high crime and low enforcement rates of the 1970s created widespread public outrage at lawlessness and disorder. Legislatures responded to this outrage by drastically increasing penalties on relatively minor crimes, such as drug possession – just as Sunstein thinks ought to happen when enforcement levels are low.

Dewey argued that, given the inescapability and indispensability of evaluative habits to decision making, our task is to design decision-making institutions so that our heuristics function flexibly and in response to the widest view of the consequences of conduct for all. Rigidity and narrowness are not inherent features of our evaluative habits, but reflect defects in the social contexts that inculcate and trigger them. Properly structured reflection does not override our heuristics, but incorporates them as flexible inputs to deliberation.

Biting the utilitarian bullet

Jonathan Baron

Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104-6241. **baron@psych.upenn.edu**
http://www.sas.upenn.edu/~baron

Abstract: The heuristics-and-biases approach requires a clear separation of normative and descriptive models. Normative models cannot be justified by intuition, or by consensus. The lack of consensus on normative theory is a problem for prescriptive approaches. One solution to the prescriptive problem is to argue contingently: if you are concerned about consequences, here is a way to make them better.

Of course I agree, in spades, with the gist of Sunstein's argument. The extension of the heuristics-and-biases approach to the realms of morality and public policy is natural, and the target article presents this approach in a balanced way. I suspect, though, that most of the criticisms of this approach, and of the article, will say that he goes too far in making assumptions about normative models. So I want to provide a counter-weight by arguing that he does not go far enough.

The idea of a moral heuristic arises within an approach to the study of judgment that relies on three types of models: normative, descriptive, and prescriptive (Baron 1985; 2004). The normative model is the standard by which we evaluate a judgment as being better or worse. Heuristics are part of descriptive models, by which we explain systematic departures from normative models. Prescriptive models are suggestions about what someone should do, all things considered.

I have argued that normative theory must be separated from in-

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

tuitive judgment, lest we lose it as a tool for criticizing and improving judgment. We cannot base normative theory on intuition, or on what intelligent people can be convinced of. Thus, in order to say that some judgment is biased, we need to do more than show that it is not reflectively endorsed, or even that it is "inconsistent." The definition of consistency itself often presumes a normative theory (Baron 1994a).

Sunstein argues that "any ambitious theory is likely to be too contentious to serve as the benchmark for measuring moral truth" (sect. 3, para. 1). Why? Must we assume that consensus must be achieved in order to make progress? Must we accept the argument that "many people disagree" (or even "most scholars disagree") as a killer argument against an ambitious normative theory such as utilitarianism, without examining the reasons for disagreement and whether they are responsive to the best arguments in favor of the theory? Some proponents of utilitarianism (an ambitious theory, for sure) think that some form of it can be derived logically from a useful analytic framework, such as the analysis of decisions into acts, states, and outcomes, with beliefs depending on states, and values depending on outcomes (e.g., Baron 2004; Broome 1991; Hare 1981; Kaplow & Shavell 2002). When opponents neglect our arguments, are we to simply give in? Give in to what?

Consider the problem that non-utilitarians have with distribution. Sunstein argues, "It is far from clear that a moderate utility loss to those at the bottom can be justified by a larger utility gain for many at the top" (sect. 3.1, para. 3). It is indeed far from clear intuitively, but the intuition that makes it unclear seems to be an overextension of a good utilitarian heuristic – that, other things being equal, the poor benefit more than the rich from a given good – to utility itself. Greene and Baron (2001) asked their subjects to evaluate distributions of utility, after making utility ratings of other goods (so that they had some idea of what utility was). The subjects showed declining marginal utility for the goods, as we would expect, but they showed just as much "declining marginal utility" for utility! This made them internally inconsistent. Greene and Baron argued that Rawls's objection to the distributional consequences of utilitarianism is based on this overextension. Sunstein admits that such basic principles as Rawls's difference principle could result from overextension of intuitive heuristics. My point here is that this kind of overextension may account for much of the difficulty of reaching consensus.

Some biases can be demonstrated by showing that they are not reflectively endorsed. They result from System I. Yet, many demonstrations of biases, such as omission bias, present the two cases to be compared (act and omission) adjacently, so that subjects have a chance to reflect. Many biases typically demonstrated with separated examples are also found with adjacent presentation (Frisch 1993). Some non-moral biases seem to resist even extensive argumentation, although they are clearly biases, such as Ellsberg's ambiguity effect (Baron & Frisch 1994; see also Baron 2000b, pp. 268–73).

Possibly the most serious question that results from lack of consensus about normative theory is what prescriptive implications can be drawn from heuristics-and-biases research on policy judgments. It is difficult to impose utilitarianism on law and public policy when most people do not accept utilitarianism. Sunstein himself has faced this problem repeatedly and dealt with it creatively (Sunstein 2002; Sunstein & Thaler 2003).

Perhaps one other way to move forward without requiring consensus on utilitarianism (or any normative theory) is to focus on utilitarianism's main feature, its focus on consequences. Sunstein comes close to this in his emphasis on "weak consequentialism." Assume for a moment that the way to bring about the best consequences on the whole, to maximize utility, is to try to maximize utility. Then, any biases or heuristics that lead to different policies will make outcomes worse. The argument then becomes a conditional: If you are concerned about policies leading to consequences that are less good than they could be, then try to correct, or work around, the heuristics and biases that lead to suboptimal consequences (as argued in Baron 1998).

A possible problem with this argument is that the assumption it requires may be incorrect. It may be that we maximize utility only by trying to do something else, as Sunstein argues in the section on exotic cases. But the examples that make this argument plausible come mostly from personal behavior. In the domain of judgments about public policy, many other examples (such as those cited by Baron 1998) argue that the assumption is approximately correct: If we try harder to bring about good consequences, putting aside our nonconsequentialist intuitions, we might actually succeed.

Towards an intuitionist account of moral development

Karen Bartsch and Jennifer Cole Wright

Psychology Department, University of Wyoming, Laramie, WY 82071-3415.
bartsch@uwyo.edu **narvik@uwyo.edu**

Abstract: Sunstein's characterization of moral blunders jointly indicts an intuitive process and the structure of heuristics. But intuitions need not lead to error, and the problems with moral heuristics apply also to moral principles. Accordingly, moral development may well involve more, rather than less, intuitive responsiveness. This suggests a novel trajectory for future research into the development of appropriate moral judgments.

Sunstein argues that, like other types of judgment, our moral judgments often employ heuristics (i.e., "mental short-cuts" or "rules-of-thumb") that lead to blunders in the moral, legal, and political domains. Though we generally agree with his discussion, Sunstein's intriguing portrayal of moral decision-making fails to adequately distinguish between two distinct aspects of the phenomenon. The first is the *process* by which moral heuristics are employed; the second is the *structure* of moral heuristics themselves.

Concerning *process*, Sunstein claims that moral heuristics are employed by the "intuitive system," which is known for being "rapid, automatic, and effortless" (sect. 2.2, para. 3). We believe that Sunstein's indictment of moral heuristics relies on a complaint against the intuitive system that may be unwarranted. First, intuitions are not necessarily grounded in heuristics. Second, we see no reason why "intuitive responsiveness" must lead to error. In fact, it may *protect* us from error; for, when adequately developed, intuitive responsiveness may reliably lead to appropriate moral judgments – a possibility we will expand on shortly.

Concerning *structure*, Sunstein's discussion oscillates between at least two distinct types of moral heuristics: those of the traditional "rule-of-thumb" type (e.g., "do not tamper with nature"), and others more instinctive/affective in nature (e.g., the "outrage heuristic"). We will focus on the former. Sunstein argues that these heuristics are context-insensitive in a way that leads to unjustified, and sometimes dangerous, over-generalizations. We suggest that the structure of such heuristics is indistinguishable from the structure of moral principles. That is, we see no relevant structural difference between *heuristics* like "punish, and do not reward, betrayals of trust" and *principles* like "do not knowingly cause human death." Consequently, *pace* Sunstein, the problem with heuristics is a problem with principles, as well.

Consider two of Sunstein's moral heuristics: "people should not be permitted to engage in moral wrongdoing for a fee" and "do not tamper with nature." As Sunstein points out, these heuristics are structurally blind to the many morally relevant details present in particular situations. But so are moral principles. Consider two well-known principles: "always keep your promise" (Kant 1948/1964) and "maximize utility" (Mill 1861/1957). Using these principles to guide one's judgments can lead to moral blunders. For instance, keeping one's promise is problematic in situations where it is morally appropriate to break the promise. Thus, the problems with applying a maxim to a complicated, contextualized problem

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

exist regardless of whether the maxim is a principle or a heuristic. Moreover, this is so whether such heuristics/principles are utilized rapidly, automatically, and effortlessly, or in deliberative reasoning. By themselves, they do not specify how/when they should be employed, how/when they admit of exceptions, how/when they ought to be used in conjunction with other heuristics/principles, and so on. In other words, there is a *gap* between how far reliance on moral heuristics/principles takes us and where we need to be in order to achieve appropriate moral judgments.

Some moral philosophers have answered this dilemma by positing a moral "sensitivity" (e.g., McDowell 1998; Railton 2000; Wiggins 1987/2002). But, what could this "sensitivity" be? We think that it may be the product of a well-functioning intuitive system, one that allows for rapid, automatic, effortless responsiveness without heuristics/principles. The question for moral psychology, then, becomes: how does one develop a well-functioning intuitive system, and thus moral maturity. If Sunstein's critique of "rule-of-thumb" moral heuristics is correct (which we think it is), then the fact that heuristics and principles have identical structures suggests that this development cannot occur through the internalization of heuristics/principles alone.

Hubert and Stuart Dreyfus's model of expertise (Dreyfus & Dreyfus 1986; 1991) might provide insight into the development of moral maturity. Their model suggests that heuristics/principles play a circumscribed role in moral development. If we consider moral maturity akin to other forms of expertise, then its development might be best characterized as a movement *away from*, rather than *towards*, moral judgments guided by heuristics/principles. In the expertise model, heuristics/principles are introduced early in development as basic rules that identify features recognizable without the benefit of experience (e.g., when learning chess, each piece is assigned a value and one is taught the rule "always exchange if the total value of pieces captured exceeds the value of pieces lost"). Reliance on such heuristics/principles is gradually replaced with procedural knowledge (i.e., know-how) gained through experience. Such knowledge leads to intuitive responsiveness. Intuitive responsiveness is the hallmark of expertise generally, because it enables rapid, automatic, effortless judgments in response to particular environmental contingencies. Importantly, such responsiveness is also reliably *appropriate*: this is what makes an expert an expert.

Examples might help. Just as the professional ski racer knows precisely how to adjust her posture to bring herself quickly around a steep turn; just as the concert pianist's fingers move skillfully across the keys; just as the master chess player can play 5–10 second/move games without significant degradation in her performance – so, too, might the morally mature person simply *see* the moral relevance of particular situations and evaluate accordingly. Of course, we do not contend that intuitive responses are always correct, anymore than Sunstein maintains that they are always wrong. We simply wish to point out the need to treat the intuitive aspect of decision-making as a matter orthogonal to the issue of how heuristics/principles are applied, and to recognize that an intuitive response may in fact be characteristic of moral maturity.

In order to test the adequacy of the expertise model, researchers must gain insight into the development of moral maturity. This suggests a shift in emphasis from a focus on moral reasoning to the following sorts of questions: What kinds of activities lead to the development of moral know-how? What kinds of instruction/modeling are children morally responsive to? What kind of feedback best engenders moral sensitivity? Sunstein states that a primary goal of his article is to stimulate future research. We hope such research will include an exploration of these developmental issues, examining the potentially independent roles of intuitive processing and the application of heuristics/principles.

Neurobiology supports virtue theory on the role of heuristics in moral cognition

William D. Casebeer

National Security Affairs, Naval Postgraduate School, Monterey, CA 93943.
wdcasebe@nps.edu
http://www.usafa.af.mil/dfpfa/CVs/Casebeer.html

Abstract: Sunstein is right that poorly informed heuristics can influence moral judgment. His case could be strengthened by tightening neurobiologically plausible working definitions regarding what a heuristic is, considering a background moral theory that has more strength in wide reflective equilibrium than "weak consequentialism," and systematically examining what naturalized virtue theory has to say about the role of heuristics in moral reasoning.

I agree with much of what Sunstein says about the role heuristics play in moral judgment and applaud his effort to make moral theorizing responsive to what is known about how human beings reason in concrete circumstances. The case for the existence of moral heuristics can be strengthened, however, by: (1) tightening working definitions regarding what constitutes a heuristic, with requisite sensitivity to their neurobiological underpinnings, (2) pressing for wide reflective equilibrium as we formulate our background "most plausible moral theory" so as to avoid charges of circularity, and (3) systematically considering what one major moral theory, a naturalized virtue theory, would say about the role heuristics play in moral cognition.

As we triangulate on a good theory about what a heuristic is, we should keep in mind the neurobiological substrates that constitute them. Despite the advanced state of play in the study of heuristics, it is difficult to articulate a framework that tells us with rigor just what they are. As Gigerenzer notes, one-word explanations can become surrogates for what should be richer psychological/neurobiological theories, saying of the representativeness, availability, and anchoring heuristics that "thirty years and many experiments later these three 'heuristics' remain vague and undefined, unspecified both with respect to the antecedent conditions that elicit (or suppress) them and also to the cognitive processes that underlie them" (Gigerenzer 2000, p. 290). In like vein, I worry whether candidate moral heuristics offered by Sunstein (e.g., "people should not be allowed to engage in moral wrongdoing for a fee," "condemn as morally wrong things that outrage you," or "do not tamper with natural processes for human reproduction") really constitute *heuristics*. They sound much like candidates for potential moral *principles* (though not very promising ones). If I have independent reason (let's stipulate this for the moment) for believing people ought not to be treated as a mere means, in what sense am I bringing a "heuristic" to bear on moral problem-solving when I apply Kant's categorical imperative? Are heuristics present only in System I? Kantians would insist that they (not weak consequentialists) are in fact responding to the demands of the thoughtful, more detached, System II with their principles, as the deliverances of the categorical imperative are not automatized, influenced by emotion, subject to framing effects, and the like (or so they might maintain – this is probably false when we examine the neurobiological evidence). Allowing neurobiology to upwardly constrain theorizing about what is a heuristic will be useful, as minds and brains are always token-identical and perhaps even type-identical in some circumstances. If upon empirical investigation a heuristic has no plausible neurobiological substrate, nor any law-like connection to activation of evolved brain systems and architecture, that makes it a bad candidate for election to office. For reviews of the neurobiology of moral cognition, see Greene and Haidt (2002), Casebeer (2003a), and Casebeer and Churchland (2003).

My second concern is closely related. As a general methodological principle, theorizing in all domains should strive for consilience: at the very least, domains should be consistent, and in exemplary cases one domain might even be reduced to another (much scientific progress had been made in this way). "Wide," rather than "narrow," reflective equilibrium should be the norm.

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

Weak consequentialism fares better on this score card than the neo-Luddite natural law theory that would (for example) follow the prime directive with regard to natural human reproduction. But there may be other moral theories that fare even better than weak consequentialism on this score. If so, Sunstein's own background theory will be another species of heuristic. If not, Sunstein may have to spend more time "in the weeds" with regard to the difficult task of moral justification, otherwise friends of natural law or unrepentant deontologists will say that Sunstein himself is offering only a heuristic and not a truth yardstick (is weak consequentialism justified primarily by intuitions that are themselves pumped by ecologically invalid heuristics?). I don't think Sunstein's argument is viciously circular, but others may. In any case, investigation of moral cognition always involves a background normative moral theory, itself being justified well or poorly, and that justification should involve wide reflective equilibrium (indeed, this is one method for successfully bridging the is/ought gap that purportedly threatens to make the study of moral psychology irrelevant to moral theorizing).

My third concern is that the best candidate for the "big tent" in which other moral theories are seen as (sometimes praiseworthy) heuristics was not mentioned: namely, a fully naturalized neo-Aristotelian virtue theory. There's more to the moral life than rights and consequences. Indeed, consideration of only these two components of "moral ecosystems" tends to de-emphasize the cognitive work Sunstein rightly sees as not informing some contemporary moral theorizing. Virtue theory tends to require richer moral psychology, more awareness of the importance of ecological validity, and a willingness to recognize the limits of theory. Virtue theory comes down "in the middle" with regard to whether morality is universal or particular. (This will affect whether or not we view moral theories as being merely heuristics; for an introduction to particularism, see Hooker & Little 2001.) A scientifically burnished Aristotle will treat moral statements as being statements about what maximizes human flourishing *cum* functionality. Cognitive acts that enable us to maximize our proper functioning (by whatever proximate mechanism) are not merely heuristics; instead, we can to a first approximation "read off" moral theories from the cognitive models we build in various environments (most of them *social*) to enable us to effectively confront mismatches between our functional demands and our environment. In some cases, those moral theories will be deontic, in others, utilitarian. Weak consequentialism becomes a heuristic itself, given normative backbone by virtue theory. A highly consilient virtue theory would thus become the yardstick against which we could call some varieties of moral coping heuristics, but at least heuristics with sometime ecological validity. For more detail on this approach, see Casebeer (2003b), Churchland (1998), or Arnhart (1998). All told, I compliment Sunstein for accomplishing the difficult integrative work required if we are to improve moral judgment in actual practice.

About emotional intelligence and moral decisions

Pablo Fernandez-Berrocal and Natalio Extremera

Facultad Psicologia, University of Malaga, Malaga, 29071 Spain.
berrocal@uma.es **nextremera@uma.es**
http://campusvirtual.uma.es/intemo

Abstract: This commentary explores the use of interaction between moral heuristics and emotional intelligence (EI). The main insight presented is that the quality of moral decisions is very sensitive to emotions, and hence this may lead us to a better understanding of the role of emotional abilities in moral choices. In doing so, we consider how individual differences (specifically, EI) are related to moral decisions. We summarize evidence bearing on some of the ways in which EI might moderate framing effects in different moral tasks such as "the Asian disease problem" and other more real-life problems like "a divorce decision."

In their initial articles on heuristic and biases, Tversky and Kahneman used examples to illustrate heuristics that did not differ in their affective valence. Thus, to explain the availability heuristic they gave the example of how participants overestimated the number of words that begin with the letter r, but underestimated the number of words that have r as the third letter. Then they gave an example about risk perception as to assess our vulnerability to sexual assault. Although the difference may be obvious to the intelligent reader, research ignored the emotional component for a long time. Fortunately, research and conceptualization of heuristics has progressed to include important aspects such as emotions and individual differences (Kahneman & Frederick 2002; Schwarz & Vaughn 2002; for a review, see Gilovich et al. 2002).

In the sphere of moral reasoning, the concurrence of the rational and the affective element when making a moral decision has been emphasized (Greene & Haidt 2002). In this sense, we would like to point out the importance of Emotional Intelligence (EI) in the resolve of moral decisions. Why EI? Because EI involves striking a balance between emotion and reason in which neither is completely in control.

We will focus on two examples to show the influence of EI in moral decisions: The Asian disease problem (Tversky & Kahneman 1981, p. 453) and a divorce decision.

The Asian disease problem. It is true that within this kind of problem people's intuitions depend on how the question is framed (for a review, see Dawes 1998). However, previous studies have shown individual differences on a variety of framing problems. People of higher cognitive ability (i.e., individuals with higher need for cognition or verbal and mathematical SAT scores) were disproportionately likely to avoid fallacy (Smith & Levin 1996; Stanovich & West 1998).

On the other hand, studies have shown the influence of emotion on risk perception. Lerner and Keltner (2001) showed the general tendency for angry and happy individuals to seek risks and for fearful individuals to avoid them, and these patterns were held independent of framing. But, how do people's emotional abilities influence their moral decisions? Fernandez-Berrocal and Extremera (in preparation, Study 1) have shown, using "the Asian disease problem," the influence of emotional intelligence (EI) on risk decisions. To evaluate EI, subjects completed an abridged version of Trait Meta-Mood Scale (TMMS-24) a month before taking the task (Fernandez-Berrocal et al. 2004; Fernandez-Berrocal et al. 2005; Salovey et al. 1995). TMMS-24 is a measure of what Salovey's research group has termed Perceived Emotional Intelligence (PEI), or the knowledge individuals have about their own emotional abilities (Salovey et al. 2002). This scale addresses three key aspects of PEI: Attention conveys the degree to which individuals tend to observe and think about their feelings and moods; Clarity evaluates the tendency to discriminate between emotions and moods; Repair refers to the subject's tendency to regulate his/her feelings.

As previous research found, our results showed that 80% of a sample of university students ($N = 189$) became risk averse (i.e., choose the certain outcome) when identical choices were framed as gains. But when the results are analysed considering individuals scores on Repair, we find a different pattern. Specifically, 85% of low Repair individuals chose the certain outcome, but only 57% of individuals with high Repair chose this option ($z = 1.78$; $p < .05$). This preference of high Repair by risk seeking is similar to that reported in studies with happy or optimistic individuals (Lerner & Keltner 2001).

A divorce decision. Fernandez-Berrocal and Extremera (in preparation, Study 2) studied people's reactions towards an emotional dilemma closer to decisions people make in everyday life: a divorce decision. Two groups of participants were studied ($N = 142$): high school students ($N = 63$) and university students ($N = 79$). Participants completed the TMMS-24, and one month later they watched a fragment of the movie "The Bridges of Madison County." In this film, the main character, played by Meryl Streep, is a married woman, mother of two children, whose relationship

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

with her husband is very apathetic. She falls in love with a photographer, played by Clint Eastwood, who visits the town. In a very emotionally intense moment of the film, she has to make the decision of whether to stay with her husband or run away with her lover. This decision is visually represented in the movie when she is inside her husband's truck and she has to decide whether to open the door to get out of his truck and get in Clint Eastwood's car, or to stay in her husband's truck. Participants watch this fragment of the movie and then they are asked to write what they would do if they were in the same situation, and to justify their choice.

Results showed that 73% of high school students choose the option "go with him." In contrast, only 54% of university students prefer this option ($z = 1.66$; $p < .05$). If we examine the differences on EI scores measured with TMMS-24 between these two groups of students, significant differences on Clarity are found. Specifically, university students understand their emotions better than high school students ($F(1, 139) = 11.69$). If we consider the relation between the score on Clarity and the decision made by participants in each group, we find that high school students who chose "go with him" obtained lower scores on Clarity ($M = 2.61$; $SD = .78$). In contrast, the highest scores on Clarity were obtained by those university students who chose to stay with their family ($M = 3.47$; $SD = .89$).

The moral and emotional dilemma presented to the protagonist does not have one unique solution, and it is impossible to assess objectively which of the options is the better one. However, if we ask different people what they would do, we find that moral and emotional understanding of the situation is influenced by age (meaning life experience) and by their EI, specifically by their level of understanding emotions. These findings suggest that the quality of moral decisions is very sensitive to emotions, and that EI might determine decisions in different moral tasks.

Sunstein's promising proposal about moral heuristics should take into account these results to avoid errors committed by initial studies on heuristics in cognition missing the influence of emotion and of individual differences in decision-making processes.

Moral heuristics and the means/end distinction

Barbara H. Fried

Law School, Stanford University, Stanford, CA 94305. **Bfried@Stanford.edu**
http://www.Stanford.edu

Abstract: A mental heuristic is a shortcut (means) to a desired end. In the moral (as opposed to factual) realm, the means/end distinction is not self-evident: How do we decide whether a given moral intuition is a mere heuristic to achieve some freestanding moral principle, or instead a freestanding moral principle in its own right? I discuss Sunstein's solution to that threshold difficulty in translating "heuristics" to the moral realm.

Sunstein's suggestion that many of our most tenacious moral instincts may simply be moral heuristics that have outlived or outreached their usefulness helps illuminate some otherwise inexplicable features of our common-sense morality. At the same time, transposing the notion of a "heuristic" from the factual to the moral realm poses some difficulties. I want to press a bit on the central difficulty here, that is, the distinction between a moral heuristic and a freestanding moral principle.

A moral heuristic, by analogy to heuristics in the factual arena, is defined as a mental shortcut we employ to get us to what we think "morality" requires. In other words, it is not itself a freestanding moral principle, but instead just a means to advance some other, often unstated, moral principle. That definition opens up the possibility, seized on by Sunstein, that we can demonstrate "error" without judging the moral truth of the underlying moral principles themselves: A moral heuristic misfires if, adopted in order to advance a given freestanding moral principle (whether good or bad), it turns out not to advance it at all. So far, so good. But how do we tell whether a given moral intuition is a freestanding moral principle, or instead a moral heuristic in service to some other moral principle? Here, the analogy to factual heuristics runs into some trouble. In the factual context, the means/end distinction is self-evident. To use one of Sunstein's examples, if we are trying to guess how many words in four pages of a novel have "n" as the next-to-last letter, the desired end is a correct estimate; the particular illustrations we conjure up to answer that question, in response to the availability heuristic or other rules of thumb, are the means. But when someone says, "A company should never knowingly manufacture a product that will foreseeably kill 10 people," how do we tell whether this is a moral heuristic in service of some other moral principle, or instead a moral principle in its own right?

The answer Sunstein gives is, in effect, a procedural one: a moral intuition counts as a freestanding moral principle only if the holder judges it, upon System II reflection, to be coherent with all other moral principles he or she holds. I'm not sure anyone can ever do better than this, but there are some difficulties lurking here.

First, the requirement of "moral coherence" built into Sunstein's version of reflective equilibrium seems too stringent. Consider Sunstein's suggestion that a "cold heart heuristic" may be at work in our response to risk regulation: "Those who know they will cause a death, and do so anyway, are regarded as cold-hearted monsters" (sect. 5.1.1, para. 3). Most people would agree, on reflection, that the intuition misfires when (in Sunstein's example of Companies A and B) it causes people to judge identical conduct differently based on mere verbal differences. But consider another case of the System I "cold heart" moral intuition at work that is much harder to write off as mere moral error: the standard heroic rescue cases. Baby Jessica falls down a well. With all the world watching her plight on the evening news, we commit millions of dollars of society's resources to rescue the victim, putting the rescuers in physical peril. If you asked citizens whether they would be willing to commit one-tenth of that amount to safety measures that would save 100 lives, almost all will refuse. That we feel far more empathy for identifiable victims than statistical ones may be highly regrettable (Loewenstein et al. 2005). But can we dismiss it as simply the product of a moral ("cold heart") heuristic that has misfired in service of some freestanding moral principle (e.g., save lives where possible)? Why is it not a moral principle in its own right? Consider, in this regard, Allan Gibbard's (1986) thoughtful suggestion that, even if the fewest lives will be lost by allocating the entire safety budget to prevention and none to costly, heroic rescues, "[i]t may nevertheless be dehumanizing to stand idly by when strenuous, expensive effort has a substantial chance of saving lives." Clearly, a public that simultaneously wishes to maximize the number of lives saved *and* not to feel it has "stood idly by" while recognizable people die, is going to be torn between two contradictory impulses that are hard to reconcile into one coherent moral scheme. But surely it misses something to write off the latter impulse as the product of a "cold heart heuristic" – with the implication that *everyone* would produce a better world by their *own* lights if their System II self could only train their System I self to stand idly by when the costs of rescue become too great.

Second, although Sunstein clearly intends his criterion for smoking out "moral heuristics" to be neutral, as among different moral principles, I don't think it is. The requirement that a moral principle on reflection must "cohere[] . . . at all levels of generality" (sect. 1, para. 3) with all other moral principles one holds, if it has any constraining force at all, seems clearly biased in favor of certain moral systems, in particular welfarism. This is so, because the commitments of welfarism to commensurability between different values, indifference to the identity of persons, and the absence of agent-relative obligations, produce a set of working principles that (whatever their other virtues or drawbacks) tend to

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

cohere at all levels of generality. Deontological principles, in contrast, do not – at least once one descends from broad injunctions like "treat others as ends in themselves," to the incommensurate set of rights and duties such injunctions are typically taken to imply. Thus, the answer to Sunstein's sly rhetorical question, "Is Kantianism a series of cognitive errors?" is probably yes, at least as judged by his criterion.

A more neutral criterion, I think, would have to shed the substantive requirement of "moral coherence," leaving something closer to a pure procedural requirement: A moral intuition gets to be called a moral principle in its own right only if, after hard scrutiny alongside other principles one holds, one still holds to it as an end in itself, and not an uncertain means to some other end. Although that test may seem too toothless to compel any familiar moral intuition to be re-characterized as a mere heuristic, I share Sunstein's optimistic belief that it might suffice, at least for some of the more dubious intuitions he catalogues here.

Moral judgments in narrative contexts

Richard J. Gerrig

Department of Psychology, Stony Brook University, Stony Brook, NY 11794-2500. **rgerrig@notes.cc.sunysb.edu**

Abstract: In narrative contexts, people often find themselves mentally rooting for "bad guys." These circumstances lead to questions about how Sunstein's moral heuristics function during narrative experiences. In particular, must people undertake explicit moral analysis for the heuristics to apply?

At the outset of the movie "Matchstick Men," a character named Frank Mercer is on the telephone trying to complete a con job. Although we don't see the person on the other end of the phone, her voice and utterances identify her as a rather helpless elderly woman. Even so, it is hard to watch the scene without rooting that Frank's con will succeed. Although his actions are far from heroic, he is momentarily the hero of the tale and so his goals are the viewers' goals – however immoral those goals might be. A movie critic offered a similar analysis of moral disengagement in narrative experiences: "Narrative art forms like novels and movies are governed by certain mysterious but implacable laws, and one of them is that when people are in danger of being caught – even if they are doing something awful – we root for them to get away. Our identification overcomes our scruples" (Denby 1991, p. 32).

These anecdotes of narrative experiences provide interesting cases for Sunstein's account of moral heuristics. In Frank Mercer's case, it seems clear that he will profit from his immoral action. As such, viewers' tacit approval of his behavior suggests that the heuristic *Punish, and do not reward, betrayals of trust* does not govern responses in this situation. Similarly, we might expect viewers to be outraged by the way in which Frank victimizes the elderly woman, so that the *outrage heuristic* would assert itself. This does not appear to be the case. Why not?

Consider Denby's assertion that "identification overcomes our scruples." Perhaps we can encapsulate this insight in the heuristic *The hero should succeed* where "hero" refers to the character or group whose goals viewers have (locally) come to embrace. We could give the same gloss for this putative heuristic as Tversky and Kahneman (1974), and Sunstein, in turn, have given for the ones they have articulated. Specifically, for most of the narrative situations people face, it seems likely that rooting for the hero will be an entirely moral response – one that rises above external criticism. However, the heuristic would leave viewers vulnerable to unfortunate occasions upon which writers and directors arrange for viewers to identify with the wrong individuals (or individuals in the wrong). Then, the heuristic would lead to moral lapses. Still, it should be the case that were we to tally up the situations in which viewers mentally root for moral outcomes (as a consequence of characters accomplishing their goals) those situations would outnumber those in which they root for characters such as Frank Mercer to succeed.

Suppose that a heuristic such as *The hero should succeed* does, in fact, play a role in narrative experiences. Then, it also seems to be the case that it takes precedence over other heuristics such as *Punish, and do not reward, betrayals of trust* – judging, at least, by the responses that reach the viewers' consciousness. During the moment-by-moment experience of the scene in which Frank Mercer attempts to hustle the helpless elderly woman, there's little hint that viewers examine the scene with sufficient rigor to realize that Frank is betraying the woman's trust.

This observation leads to the broader issue of when and how it is that moral heuristics operate. We typically think of heuristics as being automatic – availability or representativeness affect judgments without any particular entry conditions. The putative heuristic *The hero should succeed* has the same feel to it. That is, viewers do not need to make a conscious identification with a particular character before they start to embrace that character's goals. The question with respect to moral heuristics is whether people need to make an overt analysis of a situation as one in which moral judgments are relevant, before those moral heuristics come into play. With respect to Frank Mercer, it seems quite likely that one could get most viewers to apply *Punish, and do not reward, betrayals of trust* once they began to align themselves with the victim rather than with the "hero." Similarly, suppose viewers were rooting for a bank robber to escape the clutches of the police. If they took a moment for moral reflection, they might feel chastened and root instead for the police. The issue, once again, is why reflection appears to be required. Do other forces take precedence (e.g., *The hero should succeed*)? Do aspects of narrative experiences suppress or attenuate moral responses? Do moral judgments (driven by heuristics) only occur when viewers expend strategic effort?

Although the focus here has been on anecdotes from movies, there's every reason to believe that people have the same responses to narratives in other media (Gerrig 1993). In addition, it probably doesn't much matter that "Matchstick Men" is a fictional narrative. Theorists sometimes seize upon Coleridge's (1817/1907) phrase "the willing suspension of disbelief" (p. 6) as a way of conceptualizing how it is that people experience fictional narratives. In that context, we might imagine that part of what gets willingly suspended in narrative contexts would be the impulse to make moral judgments. However, "the willing suspension of disbelief" does not survive either philosophical or empirical scrutiny (e.g., Carroll 1990; Prentice & Gerrig 1999). Rather, it seems that people must effortfully encode experiences as fictional – they construct disbelief rather than suspend it. If moral judgments are affected by concomitants of narrative experiences, that ought to be equally true for nonfictional as for fictional narratives. The challenge, therefore, is to specify under what general circumstances moral heuristics are able to have an impact on covert or overt moral judgments.

ACKNOWLEDGMENTS

This material is based upon work supported by National Science Foundation Grant No. 0325188. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

Heuristics, moral imagination, and the future of technology

Michael E. Gorman

Department of Science, Technology, and Society, School of Engineering and Applied Science, University of Virginia, Charlottesville, VA 22904-4744.
meg3c@virginia.edu

Abstract: Successful application of heuristics depends on how a problem is represented, mentally. Moral imagination is a good technique for reflecting on, and sharing, mental representations of ethical dilemmas, including those involving emerging technologies. Future research on moral heuristics should use more ecologically valid problems and combine quantitative and qualitative methods.

Linking moral reasoning to heuristics is a useful approach that connects the process of ethical decision-making with the problem-solving literature. As Sunstein points out, ethical frameworks like utilitarianism can even be turned into heuristics. When I teach ethics, I often ask my students to apply different ethical approaches to a problem as if they were heuristics, designed to provoke considerations of alternatives.

The effective use of all but the most general heuristics depends on how the problem is represented (Gorman 1992). Sunstein raises the issue of representation indirectly in his brief discussion of moral framing, but misses the way in which stories and metaphors create mental models that guide thinking about moral dilemmas outside of the psychology laboratory. For example, consider the trolley task, in which switching tracks to save five people by killing one is compared with throwing one stranger in front of the tracks to save five. To understand why practical reasoners might consider these two situations different, it is necessary to understand their mental models (Johnson-Laird 1983). The action of throwing a stranger in front of a train is likely to result both in the death of the stranger and the five people. Why would the trolley stop for a single person and not five? This possible mental model converts the abstract decision into a story. Others might have different mental models of this problem, but that is the point.

Consider another example from environmental ethics. A heuristic such as "do not tamper with nature" depends on your mental model of nature, a construct that varies across cultures (Gorman & Mehalik 2002).

The problem is that people's mental models are often implicit; these need to be inferred, which is the classic problem of representation in cognitive psychology. The solution in the ethics literature is to encourage people to engage in moral imagination (Werhane 1999) – the ethical application of mental models (Johnson 1993). The first step in moral imagination is similar to the first step in innovative problem-solving, that is,to become aware of one's own mental model of a situation, so that one can explore a space of alternate solutions. This awareness step is particularly difficult, because many people reason from deeply held ideological frameworks that they represent as truths, as reality; therefore, the idea of exploring alternatives is heresy. Seeing these "realities" as views is a critical first step in listening to other views and considering alternatives.

Moral imagination is particularly critical as emerging technologies create new moral dilemmas never experienced in human history. Sunstein mentions the issue of cloning, and Kass's reliance on "moral repugnance" to decide what is ethical. Repugnance is based on deeply held unquestioned beliefs; moral imagination asks one to examine the basis for those beliefs.

Cloning and other technologies give human beings abilities reserved for the Gods in most traditional stories that serve as the basis for morality. Consider convergent technologies (Nano, Bio, Info, Cogno) that include possibilities such as interfaces that will allow for neural control of devices, wearable sensors that will extend human capabilities beyond the traditional five senses, and the development of "intelligent" military vehicles (Roco & Bainbridge 2002). These developments will create dilemmas that will require the modification of existing mental models and moral heuristics – without, however, abandoning higher-level ethical principles like Kant's "never use human beings merely as a means."

Future research on moral heuristics should use more ecologically valid problems (see Gorman et al. [2000] for examples), employing methodologies like protocol analysis (Ericsson & Simon 1984) and reflective diaries (Shrager 2004) to track the reasoning process as it occurs, supplemented afterwards by the sorts of probing questions asked by Kohlberg (Rest & Narvaez 1994). The moral problems studied by these methods should include those raised by emerging technologies, which will help anticipate public reaction before technological options have been "locked in" and human choices are constrained.

What's in a heuristic?

Ulrike Hahn, John-Mark Frost, and Greg Maio

School of Psychology, Cardiff University, Cardiff CF10 3AT, United Kingdom.
hahnu@cardiff.ac.uk frostjm@cardiff.ac.uk maio@cardiff.ac.uk
http://www.cardiff.ac.uk/psych/home/hahnu
http://www.cardiff.ac.uk/psych/home/maio

Abstract: The term "moral heuristic" as used by Sunstein seeks to bring together various traditions. However, there are significant differences between uses of the term "heuristic" in the cognitive and the social psychological research, and these differences are accompanied by very distinct evidential criteria. We suggest the term "moral heuristic" should refer to processes, which means that further evidence is required.

The target article presents an exciting synthesis of interdisciplinary research. Ideas from cognitive and social psychology are brought together with moral and legal philosophy. A common difficulty in interdisciplinary work, however, is that terms are used in subtly different ways across disciplines: the same term across disciplines might not only alert to common themes, but also obscure substantive differences. The notion of "heuristic" now possesses this ambiguity and has become a set of loosely correlated concepts, rather than a well-defined technical term. In cognitive psychology, this term was first used to denote problem-solving procedures that are easier to use than more complex algorithms, but were not guaranteed to provide a solution (Newell & Simon 1972). Famously, Tversky and Kahneman (1974) used "heuristic" to refer to procedures for probability judgment or "intuitive statistics," which contrast with a more cumbersome, correctness guaranteeing procedure. For others, "heuristics" denote simple, special purpose strategies that are *adaptations* to the environment. These strategies will not always be right, but there need be no other strategy that would be, and the key to identification of putative heuristics is showing their ecological rationality, typically through computational demonstration of how they exploit patterns of information in the environment so as to make accurate inferences (see e.g., Goldstein & Gigerenzer 2002). All of these uses refer to *procedures* that can be applied to many problems of widely varying content; heuristics are not declarative content statements such as "dogs bark." A rather distinct use of the term heuristic is found in the cognitive literature, which seeks to clarify the nature of our lay theories about the world, such as our naïve physics (see e.g., Profitt & Kaiser 2002). Here, the term heuristic refers to an implicit *theory*, that is, typically content statements, though much of the interest in this area stems from systematic errors relative to scientific theories about the world.

Within social psychology, models of persuasion treat heuristics as rules of thumb that are quick and easy guides for evaluating the validity of persuasive messages, such as "if an expert says it, it must be true." In this literature, the features that make such statements heuristics are how and when they are used. Contemporary models of persuasion indicate that people tend to use "heuristics" when they are unmotivated and unable to process the issue deeply

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

(Maio & Haddock, in press). When motivation and ability are higher, people tend to scrutinize the relevant arguments more carefully and disregard any heuristics that are unreliable or irrelevant. In this literature, "correctness," though relevant, is downplayed. There is no algorithm for deciding whether someone spoke the truth – and the relevant standard by which something is deemed to be a heuristic is typically *other reasoning by the same user*, in situations of high engagement. This means it might even be possible for a particular statement to function as a heuristic on one occasion and as a valid premise on another (Kruglanski et al. 2004). In contrast, and more akin to the terminology of the naïve physics literature, Baron (e.g., 1993a; 1994a) introduces the term "moral heuristic" for the rules that constitute our "naïve morality" (e.g., Baron 1993a). Examples include "it is wrong to hurt some people for the benefit of others" or "harmful commissions are worse than harmful omissions." Though similar in appearance to persuasion heuristics, the status of these rules as heuristics is not determined by processing context. Another perspective on moral reasoning emphasises that moral judgment might be achieved through two separate cognitive systems: an intuitive system and a reasoning system (Haidt 2001). The "intuitive system," which Sunstein equates with heuristics, is characterised as fast and effortless; its processing is unintentional, typically inaccessible to awareness, and involves parallel processing and pattern matching. For moral judgment, this intuitive system additionally involves emotion (Greene & Haidt 2002; Haidt 2001).

Crucially, these related but distinct notions of the term "heuristic" all require different kinds of evidence. Evidence for "cognitive" heuristics in the Tversky and Kahneman sense requires patterns of judgment that deviate in the predicted fashion from some standard of correctness (see also, Kahneman & Tversky 1996). By contrast, the Gigerenzer sense requires evidence of the opposite, namely, accuracy. Whether processing was deliberate or automatic, conscious or unconscious, or involves affect, is, at least in the first instance, unimportant for both (though see now, Kahneman 2002; Kahneman & Frederick 2002) and evidence for heuristics in problem-solving was even derived largely from verbal protocols of reasoners describing their thinking out loud. By contrast, evidence for "persuasion" heuristics requires demonstration that their use is influenced by motivation and ability. Finally, evidence for the "intuitive system" is virtually orthogonal to that required for "cognitive" heuristics: standards of correctness are irrelevant, and processing characteristics are all important.

Sunstein's article seems to simultaneously endorse all of the above uses, in that "deviations from correctness" and "output from System 1" and "adaptiveness" are variously emphasised. However, most of the examples given are content rules, part of our naïve morals in Baron's sense. That is, they are "moral principles that are generally sound, and even quite useful, but that work poorly in some cases" (sect. 5.1.1, para. 2). Evidence for these principles as heuristics is then supplied by describing a case for which they seemingly lead to an "incorrect" answer.

In order to evaluate Sunstein's proposal we turn to consideration of legal systems as complex systems explicating our sense of right and wrong. Setting aside the vexed issue of absolute standards of correctness, one finds that it is a property of *all* legal rules and principles that eventually cases will emerge for which their application suggests an undesired outcome. Real legal systems try to minimize this problem through a proliferation of rules of different scope, whereby the system is supplemented with further rules defining exceptions and exceptions to exceptions. Unanticipated exceptions will nevertheless arise. In other words, legal norms are inherently *defeasible* (see e.g., Bankowski et al. 1995). Seen in this light, there is little point in calling a moral content statement a "heuristic" simply because it can and eventually will give rise to an unwanted "overgeneralization."

This suggests to us that the term "moral heuristic" would better be limited to *processes.* The target article provides no real evidence to this effect. However, we have, for example, recently found intriguing effects of typicality. In several experiments (Frost et al., in preparation), we asked participants to analyse their reasons for the value of equality in a typical context (gender discrimination) and an atypical context (handedness discrimination). Results indicated that participants who generated reasons in a typical context later acted in a more egalitarian manner than participants who generated reasons in the atypical context, despite listing similar numbers of reasons for the value and being equally confident in their reasons. Here, issues of correctness seem unproblematic, as participants see no reason why their behaviour on the same task should differ according to exposure to previous material. At the same time, it seems safe to assume that the workings of this particular typicality effect are entirely opaque to participants. In short, the concept of a moral heuristic might yet prove useful in explaining moral judgment and behaviour, but only if it is about more than particular content rules or principles, which are prone to exception.

Invisible fences of the moral domain

Jonathan Haidt

Department of Psychology, University of Virginia, Charlottesville, VA 22904.
haidt@virginia.edu **http://www.people.virginia.edu/~jdh6n/**

Abstract: Crossing the border into the moral domain changes moral thinking in two ways: (1) the facts at hand become "anthropocentric" facts not easily open to revision, and (2) moral reasoning is often the servant of moral intuitions, making it difficult for people to challenge their own intuitions. Sunstein's argument is sound, but policy makers are likely to resist.

Look at it from Bin Laden's point of view. For years the United States had been. . . . don't worry, I'm not going to finish the sentence. I can't. I study morality and I know that terrorism is driven largely by moral commitments. Yet, every time I try to understand Bin Laden, or Hitler, or political leaders with whom I strongly disagree, I feel a kind of invisible fence (the kind used for suburban dogs) giving me a warning shock, saying "don't go there, don't even think about empathizing." In contrast, I can roam freely around the Linda problem, the Asian Disease problem, and the visual illusions that I use to show my Psych 101 students how perceptual heuristics can sometimes misfire. It can be difficult to look at a probability problem or a perceptual illusion in a different way, but it is never dangerous or painful.

Sunstein's effort to bring the well-developed tools of research on heuristics into moral psychology is welcome and well done. His emphasis on "System I" processes in the moral homunculus is consistent with recent emphases on the role of emotion and intuition in moral judgment (Damasio 1994; Greene et al. 2001; Haidt 2001). However, the moral domain is a weird and treacherous world in which objects change their weights and rivers flow uphill. Or at very least, minds that worked in one way on non-moral problems suddenly start working differently when moral concerns are introduced. Here I discuss two such differences which I believe can be integrated into Sunstein's approach, giving us a fuller and more social picture of the workings of moral heuristics.

1. Moral truths are anthropocentric truths. Sunstein contrasts the moral domain with the "domain of facts," suggesting that moral truths are not facts, but this is not quite right. A useful distinction can be made between two kinds of facts – anthropocentric and non-anthropocentric (Wiggins 1987). Non-anthropocentric facts are those that do not depend for their truth on the way the human mind is constituted. Facts about the physical world and mathematical truths are true regardless of what we happen to think about them, and they would presumably be true for any intelligent species that came to our solar system to inspect them. But our judgments about beauty, humor, and morality are factual judgments too. They are judgments about anthropocentric truths – truths that are true only because of the kinds of minds that we

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

happen to have, and the cultural worlds within which our minds developed. When we give an A+ to one paper and a D to another we are asserting that one paper really is better than the other, within our academic community, although we might not expect intelligent extraterrestrials to agree with us.

Anthropocentric truths arise within communities, and they then do much of the work of marking out the limits of those communities. But even within the realm of anthropocentric truths, moral facts are especially potent. Groups can usually tolerate a diversity of beliefs about beauty and comedy, but moral diversity is much more damaging (Haidt et al. 2003). One cannot even coherently want moral diversity. For example, if a person says, "I believe that women should have the right to choose, but I would prefer that there be a diversity of opinions on that matter," then that person treats abortion rights as a taste, not as a moral issue. Foundational beliefs, such as taking the bible as the literal word of God, or the idea that the world is full of victims of oppression who must not be blamed for their fate, become sacralized, and those who question them risk becoming pariahs. Many moral heuristics may have this sacred character for some groups (e.g., don't play God, don't knowingly kill anyone, don't have sex with your family members, don't blame victims), so questioning them, even in special cases where they don't really apply, is likely to meet with resistance and even outrage. The problem with moral heuristics is not that there is no fact of the matter with which to compare them; rather, it is that there are many (anthropocentric) facts of the matter, and it is hard to get people to question their anthropocentric moral facts.

2. In the moral domain, System II is often a slave. In Sunstein's analysis, System II (reasoning) either opposes System I (by reaching a conclusion that the homunculus opposes) or it sits back and does nothing while System I spits out its heuristic conclusion. But, in any domain in which strong motivations are at work, reasoning often becomes the "slave of the passions," as David Hume put it. We can sometimes see this process at work for non-anthropocentric facts, as when students struggle to find reasons to explain how Linda is more likely to be a bank teller active in the feminist movement than to be a bank teller. But many people are able to reason their way to a solution, and there is often a moment of insight in which System II triumphs and people understand their error. Not so for moral disputes. I have now interviewed several hundred people about taboo violations such as consensual safe sex between an adult brother and sister, and I have never yet seen a person say "Oh, I see! I had this strong gut feeling that it was wrong, but now that I understand that no child can result from the union, I realize that I was mistaken." More typically, people struggle valiantly to find some reason why even in this special case the brother and sister should not have sex. We can therefore expect a lot less help from System II in challenging moral heuristics than we get from it in challenging non-moral heuristics. In fact, whenever moral emotions are engaged, as they often are when anthropocentric facts are challenged, we can expect to find the System I homunculus ordering System II to man the ramparts and fight off persuasion.

I think these two differences make a difference. I applaud Sunstein's call for distrust of moral heuristics when considering unusual or difficult cases. And I expect that many people will agree with him, as I do, in the abstract. But when it comes time to make policy decisions about abortion, euthanasia, cloning, or any other difficult issue, don't be surprised when politicians and policy makers refuse to cross their invisible fences, or when they attack those who ask them to do so.

ACKNOWLEDGMENTS

I thank the John Templeton Foundation for its generous support of my work.

Sunstein's heuristics provide insufficient descriptive and explanatory adequacy

Marc D. Hauser

Departments of Psychology, Organismic and Evolutionary Biology, and Biological Anthropology, Harvard University, Cambridge, MA 02138.
Mdhauser@Wjh.Harvard.edu **http://www.Wjh.Harvard.edu/~Mnkylab**

Abstract: In considering a domain of knowledge – language, music, mathematics, or morality – it is necessary to derive principles that can describe the mature state and explain how an individual reaches this state. Although Sunstein's heuristics go some way toward a description of our moral sense, it is not clear that they are at the right level of description, and as stated, they provide no guidelines for looking at the acquisition process – the problem of explanatory adequacy.

Consider the human language faculty. When we generate sentences, or comprehend them, we do so effortlessly. Our capacity to both understand what others say and to generate new prose is boundless. The way to make sense of this capacity is by appealing to a dedicated faculty of the mind, a system that contains a repository of computational resources for building an externalized language. For each individual, the language they construct, both over their lifetime, as well as on a moment to moment basis, represents the output of a complicated series of interfaces between the computational resources dedicated to language, on the one hand, and interactions with other mind internal–external factors, on the other hand. Linguists interested in the underlying principles that can account for what a mature speaker of a language knows are studying the descriptive principles of the system.

One of the early mysteries surrounding this approach to language was the observation that young children are able to both generate surprisingly sophisticated sentences and comprehend them in the absence of relevant input. This observation led in part to the hypothesis that our species is innately equipped with a universal grammar, a set of principles and parameters that not only enables the capacity to build a natural language, but also constrains the range of possible languages. The now rich description of the principles and parameters in play early on in development provides a sense of the explanatory adequacy of this field.

In this commentary, I make use of the importance of descriptive and explanatory adequacy in characterizing a domain of knowledge, as well as the tie in to language, to evaluate Sunstein's discussion of our moral psychology. I first describe the shortcomings of the moral heuristics position and then provide a sketch of an alternative which builds on an analogy with language (Dwyer 1999; 2004; Harman 1999; Hauser, in press; Hauser et al., in press; Jackendoff 2004; Mikhail 2000; Mikhail et al. 2002; Rawls 1971; Smith 1759/1976).

Sunstein wants to show that heuristics play a significant role in moral, legal, and political spheres, and that sometimes they generate inappropriate judgments. As stated, it is hard to imagine that anyone would disagree with these claims. Those who thought hard about common sense morality, beginning with Hutcheson and Shaftesbury, recognized that we often apply general rules of thumb in cases of moral conflict and, as Hume importantly recognized, funnel these rules through an emotional filter that guides our actions. What have always been the primary challenges to these views include our ability to understand where our common sense intuitions come from, what their representational content is, the extent to which they are consciously available principles as opposed to unconscious and inaccessible, how children alight upon them in the course of attaining a mature moral faculty, and the degree to which they facilitate or detract from our interests in normative or prescriptive principles aimed at a just world. Concerning the latter, the interest has always been a concern with how our intuitions or heuristics about right and wrong interface with more formal and explicit policies, whether they are the unstated social norms of a hunter-gatherer society or the legal doctrine of our founding fathers. So, on a general level, there is not much new in

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

terms of Sunstein's general framework, nor does his description illuminate these age-old questions.

Sunstein goes on to state that moral heuristics are different from Kahneman and Tversky-esque heuristics in that the latter are based on factual problems. But this strikes me as an inaccurate reading. Certainly, much of Kahneman & Tversky's (KT's) work has been based on how we judge the market, and what dictates our views of fairness and subjective utility. Both play critical roles in delivering moral verdicts. For the utilitarian, there is much to gain from KT's work because we now have a better sense of the currency over which individuals may seek to maximize overall well-being. For the deontologically inclined, we gain a better sense of how individuals compute fairness by appreciating that they unconsciously appeal to the principle of a reference transaction. Although it is true that much of what KT had to say about these heuristics were more readily identified as logical flaws that led to objective errors, and that the moral sphere is undeniably more subjective, it is not the case that this work falls squarely outside issues of moral concern.

Overall, then, though I am sympathetic to the general framework that Sunstein articulates, I do not think that there is much new here, and nor do I believe that his framing of the problem significantly advances how one goes about doing the science of moral psychology; of course, if the message is largely targeted at lawyers or policy makers, who may either fail to recognize the importance of heuristics in our common sense morality, or assume that such heuristics are unambiguous determiners of what we ought to do, then I couldn't agree more.

An alternative, by no means incompatible with Sunstein's moral heuristics, draws on an analogy with the language faculty. If there are either strong or weak analogies with the language faculty, then we might expect to find the following design features:

1. A universal moral grammar [UMG] that represents a theory of the initial state.
2. The UMG consists of a set of principles that provides a toolkit for building possible – external – moral systems.
3. These principles are based on combinations of actions and action sequences ("phrases") into events, anchored by the psychological processes of intentionality, motivation, cause, and consequence.
4. The judgments and actions that young children make in the moral domain cannot be accounted for by the input. As such, there is a poverty of the stimulus-type argument, which requires the inference that the initial state consists of largely content-free, abstract, and innate principles. What experience does, under this kind of model, is set the parameters, and thereby dictate which particular moral system is acquired.
5. There is a moral organ – dedicated circuitry that consists of principles for deciding whether actions are permissible, obligatory, forbidden, and/or punishable. This circuitry must interface with both other mind-internal processes, as well as mind-external ones.

When this faculty breaks down, there will be specific deficits in our moral judgments, as opposed to more general cognitive deficits.

This is an extremely rough sketch, explicated in greater detail in the references cited earlier. These ideas gain support in that they have generated both new empirical findings and have also helped to set up new research problems. For example, in a large-scale study of moral judgments using the internet, results show that, across considerable demographic and cultural variation, people converge on a set of common judgments concerning permissible harm, while having no access to the underlying principles (Hauser et al., in press). This dissociation between judgment and justification is similar, at some level, to evidence in linguistics of grammaticality judgments, and highlights the distinction between operative and expressed principles. It also suggests that some aspects of our moral judgments may well be universal. This work has led to ongoing studies of patient populations in which the relative contribution of unconscious emotions and principles of action interface with our moral judgments. These patient studies will help us to understand how the moral faculty is fractionated into different component processes, and to decide which are specific to our moral psychology as well as uniquely human.

The next frontier: Moral heuristics and the treatment of animals

Harold A. Herzog[a] and Gordon M. Burghardt[b]

[a]*Department of Psychology, Western Carolina University, Cullowhee, NC 28723;* [b]*Department of Psychology, University of Tennessee, Knoxville, TN 37996.* **herzog@email.wcu.edu gburghar@utk.edu**
http://wcuvax1.wcu.edu/%7Eherzog/ http://web.utk.edu/~gburghar/

Abstract: Heuristics provide insight into the inconsistencies that characterize thinking related to the use of nonhuman animals. We examine paradoxes in judgments and policy related to the treatment of animals in science from a moral intuition perspective. Sunstein's ideas are consistent with a model of animal-related ethical evaluation we developed twenty-five years ago and which appear readily formulated as moral heuristics.

Sunstein's argument is simple yet powerful – moral thinking, like other forms of human cognition, is frequently thrown awry by simple cognitive heuristics. This insight sheds considerable light on a topic we have long been interested in – the fact that ethical thinking about animals is rife with inconsistency and paradox (Burghardt & Herzog 1980; Herzog 1993). Indeed, for several reasons, Sunstein's heuristics may be illuminated even more when applied to understanding contradictions in how we think about animals than it is to human-focused moral quandaries. First, with animals there is ambiguity over the existence and moral relevance of mental capacities of different species (e.g., consciousness, intelligence, emotions, and the experience of pain). Second, these considerations reflect subtle and often unrecognized ethical rules of thumb. There is no shortage of examples where moral heuristics interfere with clear thinking about the use of animals. Here we briefly discuss two situations of interest to scientists.

The first is the comparative status of rats and dogs under the Animal Welfare Act (AWA). Although they make up the majority of animals used in biomedical and behavioral research, rats (along with lab-bred mice and all birds) are denied coverage under the AWA because they are not considered "animals" under the provisions of the statutes.[1] Dogs, in contrast, not only are covered by the AWA, they are the only species that the act specifies must be given daily exercise. Indeed, because the AWA applies to deceased as well as living animals, a dead dog actually has legal status not afforded a living laboratory rat.[2] There was only minor public outcry about the exclusion of rats, either when the act was written or several years ago when Congress enacted legislation permanently excluding rats from AWA coverage. Why? We suspect the rat exclusion reflects the operation of a heuristic along the lines of "Rats are pests: pests are bad."

Dogs, on the other hand, are treated differently. One reason is that rats are perceived as far less intelligent and sentient than dogs (Herzog & Galvin 1997). More importantly, dogs live in 40% of American households. For most owners, dogs assume the role of friend or even family member (Serpell 1989). The specter of one's pet splayed on the dissection table evokes a particularly powerful moral heuristic – "Don't betray friends and family." The inclusion of dogs in and the exclusion of rats from AWA coverage are consistent with most people's moral intuition.[3] The rat exclusion rule, however, is increasingly viewed as an embarrassment by regulators, and surveys indicate that most researchers now advocate coverage of rats and mice under the AWA (Plous & Herzog 2000).

Our second example concerns the role of heuristics in approval/disapproval decisions of Institutional Animal Care and Use Committees (IACUCs). As Sunstein indicates, it is rarely possible to assess the validity of ethical judgments by holding them to some sort

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

of "correct" moral yardstick. Reliability, however, is a different matter, and inconsistency of ethical decisions precludes their validity. Two studies have examined the consistency of IACUC decision-making procedures by having different IACUCs evaluate the same protocols (Dresser 1989; Plous & Herzog 2001). Both arrived at the same conclusion – more often than not, different committees make different decisions. Plous and Herzog found that even members of the same IACUC were inconsistent in their evaluations of dimensions of protocols (e.g., clinical significance, clarity). Interestingly, when IACUC members were provided with specific guidelines, such as a detailed pain scale, the role of intuitive appraisals (heuristics) seemed to decrease, and inter-rater reliability substantially increased.

Before the emergence of the animal rights movement as a political force and the enactment of important 1985 amendments to the Animal Welfare Act, we attempted to make sense of inconsistencies that we observed in ourselves and others when it came to moral judgments pertaining to other species (Burghardt & Herzog 1980; 1989). In order to systematize discussion in this area, we constructed a typology of factors that influence thinking about the ethical use of nonhuman animals in general, not just in research. We identified 26 "ethical considerations" under four major headings: human benefits (and costs), anthropomorphism, ecological, and psychological. In retrospect, we believe many of these factors function as moral heuristics (e.g., cuteness of the species, similarity in appearance to humans, status as a pest or competitor, rarity, domestication). And these 26 could be added to, subdivided, and extended today. In 1980 we concluded: "We suspect that currently it is impossible to derive from science, theology, philosophy, or any conceivable source a consistent universal set of principles to guide humans in dealing with members of other species" (Burghardt & Herzog 1980, p. 767).

Sadly, despite the growth of a veritable cottage industry of professionals in many fields and numerous journals, books, conferences, and organizations, we think that little progress has been made on general principles outside of the acceptance of some regulations and greater scientific understanding of animals. Some scholars in this area focus on narrow issues, while others adopt their own simple set of heuristics or insulated philosophical stance (utilitarianism, deontology) and ignore or remain blind to their problematic aspects. Others simply revel in the dilemmas as an enduring contradiction of the human drama, one best minimized by good intentions and modest melioration. Perhaps more formally embedding animal issues into work on moral heuristics will help clarify and resolve issues too often approached with feelings divorced from thought. Research in the cognitive sciences along the lines suggested by Sunstein may provide insights into the psychological processes that underlie differences in opinion related to human–animal interactions. This message is certainly not lost on Sunstein, who has contributed elsewhere to legal thinking about the status of animals (see Sunstein & Nussbaum 2004).

NOTES

1. Although they are excluded under the AWA, rats, mice, and birds do fall under NIH guidelines.

2. A footnote in the regulations, however, exempts dead dogs from AWA canine cage size requirements.

3. Some moral intuitions are culture-specific; whereas common sense may tell most North American pet lovers that dogs are family members, in some Asian cultures puppies are dinner.

A selectionist approach integrates moral heuristics

Robert A. Hinde

St. John's College, Cambridge University, Cambridge CB2 1TP, United Kingdom. **rah15@cam.ac.uk**

Abstract: The nature and diversity of moral codes can be understood in terms of a few basic propensities honed by diachronic dialectics between what people do and what they are supposed to do in the culture in question. Many of the moral heuristics presented by Sunstein can be seen as by-products of these processes.

In his important contribution, Sunstein shows successfully that we sometimes use "heuristics" or "short-cuts" in making moral judgements, applying principles that usually work well in instances where they are inappropriate. We must ask, however, where these heuristics come from. Sunstein uses descriptive categorisations of the heuristics as if they were causal principles, referring, for example, to "a process of attribute substitution" or an "outrage heuristic." My claim is that most of the instances of moral heuristics cited by Sunstein are compatible with, and perhaps could have been predicted from, a more interdisciplinary approach (Hinde 2002).

Such an approach indicates that moral codes stem from certain pan-cultural propensities, notably to look after one's own interests (selfish assertiveness) and to be cooperative and kind to others (prosociality), especially to close kin and in-group members. These propensities are present even in very young babies (Kagan 2000; Rheingold & Hay 1980), but are honed in development by parenting, relationships with peers, charismatic figures, and so on. These relationships have themselves been affected by the precepts to which they have been exposed and the physical environment. Individuals incorporate moral precepts into the way in which they see themselves, and experience pangs of conscience when they behave contrary to their own standards. Some individuals seem to behave morally without thinking. In other words, individuals differ in what Sunstein refers to as System 1.

Morality is concerned with maintaining a balance between the basic propensities such that group living is possible in the circumstances prevailing. The resulting moral precepts are reified somewhat differently between cultures. Often the processes involved depend on diachronic dialectical relations between what people do and what they are supposed to do. For example, the respectability of divorce in western countries has changed through dialectics between what people do and what they are supposed to do.

This is essentially an evolutionary approach (mentioned but not exploited by Sunstein), but does not try to explain everything by natural selection. There is no implication that what is natural is right. Moral judgements change somewhat with time and circumstances: Cultural selection over prehistorical and historical time is crucial. Moral precepts therefore differ somewhat between cultures, but the basic principles on which they are based (selfish assertiveness; prosociality to in-group members) appear to be pancultural. Variants of the Golden Rule, Do-as-you-would-be-done-by, are shared by all moral codes. Most of the Ten Commandments are compatible with the Golden Rule, and, not surprisingly, the commandment not to kill has special potency. This is compatible with the judgements made in, for instance, the trolley problem (e.g., stealing to save an in-group member; answering A or D in the Asian Disease problem). However, the basic propensity is limited to in-group members. Thus, killing out-group members may be permissible, and the death of contemporaries is more salient than that of remote descendants, who are seen as more distantly associated.

Again, exchange theories, invoking reciprocity (Kelley & Thibaut 1978; review in Hinde 1997) explain many aspects of human relationships, and reciprocity accompanied by prosociality is compatible with selectionist theory (Boyd & Richerson 1991). Because reciprocity often involves delay, trust in the partner, honesty, and

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

commitment have come to be esteemed highly. The emphasis on trust is entirely compatible with the cases of aversion presented by Sunstein, including those that involve a reduction in safety from the use of a device that is supposed to increase it.

The desire for reciprocal revenge is an extension of positive reciprocity, and formed the initial basis of Anglo-Saxon law (Adams 1876). Hence the preference for "pointless punishment," the desire to penalise companies in accordance with their offence, and the perceived irrelevance of the probability of detection to the penalty. Permitting wrong-doing for a fee contradicts the principle of reciprocity, but the view that companies should clear up their own waste is entirely compatible with it.

An individual may be treated prosocially even though he or she is unlikely to be met again, and will therefore not be able to reciprocate. This appears to be contrary to the principle of reciprocity. However, several processes encourage prosociality to strangers. One is moral outrage, mentioned several times by Sunstein. Computor simulation shows that, if the costs of being punished are high enough, strategies involving cooperating, punishing non-cooperators, and punishing those who do not punish non-cooperators, can be stable (Boyd & Richerson 1992). Another issue is that prosociality brings prestige, which can bring further rewards (Zahavi 2000). Thus, even this apparent exception to the principle of reciprocity can be understood in terms of selection.

The dialectical relations between what people do and what they are supposed to do are not the only issue shaping moral precepts. Influential individuals or groups can propagate precepts that are to their own benefit. Thus, the Christian emphases on humility as a virtue, and on respect for priests, are probable examples. Incest rules, referred to twice by Sunstein, depend both on a biological basis that regulates the degree of in-breeding, and on culture specific rules that regulate inter-group relations or favour influential groups, like the Church (Goody 1997). In general, precepts are likely to carry an historical legacy, and western morality has been heavily influenced, as well as purveyed, by the Christian Church.

So what is being claimed? The present approach argues that moral precepts have a biological basis, and that their precise form can be understood in terms acceptable to the behavioural sciences. The dogma that moral precepts are in some sense absolute, perhaps carved in stone, cannot be disproved but is unnecessary. This selectionist approach to morality is able to integrate many of the examples given by Sunstein, even though the foundations of moral codes referred to interactions between individuals and Sunstein's examples mostly involve applying precepts initially evolved for reactions between individuals to larger social issues.

Betrayal aversion is reasonable

Jonathan J. Koehler[a] and Andrew D. Gershoff[b]

[a]*Behavioral Decision Making Faculty, McCombs School of Business, The University of Texas at Austin, Austin, TX 78712-0212;* [b]*Department of Marketing, Michigan Business School, University of Michigan, Ann Arbor, MI 48109-1234.* **koehler@mail.utexas.edu agershof@umich.edu http://www.mccombs.utexas.edu/faculty/jonathan.koehler/**

Abstract: We accept Sunstein's claim that people often use moral heuristics to make judgments and decisions. However, in situations that include a risk of betrayal, we disagree with Sunstein about when the relevant moral heuristic may be said to "misfire." We suggest that the moral heuristic people apply to avoid the possibility of safety-product betrayal may be reasonable.

We accept Sunstein's premise that people often rely on broad, simple, moral intuitions for making judgments. Indeed, given people's desire for social goals such as fairness, justice, and trustworthiness, it would be strange if moral intuitions did *not* impact the decisions people make.

However, it is less clear that these moral intuitions – or moral heuristics – are as prone to systematic error as the classic heuristics (availability, representativeness, and anchoring and adjustment) described by Amos Tversky and Daniel Kahneman (Tversky & Kahneman 1974). Tversky and Kahneman appealed to logic and probability theory – often in combination with some controversial assumptions about how people represent decision tasks – to illustrate the occasional failures of their heuristics. The normative status of moral heuristics is less grounded. Although this shortcoming does not make moral heuristics any less real or important than other heuristics, it does mean that there is room to challenge Sunstein's judgments about when moral heuristics have misfired.

Consider the betrayal aversion phenomenon Sunstein discusses in section 5.1.3. The research we performed on this phenomenon, and which Sunstein reviews, suggests that people are willing to incur great costs to avoid betrayals and seek to punish betrayals severely when they arise (Koehler & Gershoff 2003). We provided experimental support for these phenomena in contexts where humans betrayed and in contexts where safety products "betrayed" (or threatened to betray) by causing the very harm that they were employed to prevent. Sunstein states that the morality heuristic behind betrayal aversion is "Punish, and do not reward, betrayals of trust" (sect. 5.1.3, para. 4). We agree that a rule along these lines operates, though we would describe the heuristic as "Avoid and punish betrayals of trust."

We also agree with Sunstein that this moral heuristic appears to work well in cases involving betrayals by human actors. A security guard who commits an act of betrayal by robbing the store he is paid to protect deserves the tough punishment he will no doubt receive because his crime causes multiple harms. His betrayal not only causes the focal harm to the business, but it also damages the victims' ability to trust other security officers and undermines their sense of the social order.

We are less inclined to agree with Sunstein that the moral heuristic necessarily "misfires" when people use it in situations involving betrayals by safety products rather than by people. Most participants in our safety-product study indicated that they were willing to double their risk of dying (from 1% to 2%) to eliminate an even smaller risk of dying as a result of a betrayal (Koehler & Gershoff 2003, Study 5). For example, most people preferred an airbag that carried with it a 2% risk of dying in a serious automobile crash to one that carried with it a 1% risk plus an additional 0.01% risk due to fatal deployment of the airbag. Sunstein thinks this is an example where the moral heuristic "punish, and do not reward, betrayals of trust" misfires and leads to error. A safety product should be chosen, Sunstein says, "if and only if it decreases aggregate risks."

However, we are not persuaded that the risk of betrayal is an irrelevant consideration in the safety-product context. Although a safety product lacks the intentionality of a human actor, the negative consequences of a safety-product betrayal may be as varied, severe, and protracted as other types of betrayals. An airbag that kills drivers who would otherwise survive the car accident can instill a deep mistrust of car manufacturers and government safety agencies among the victims' families and friends. Safety products that betray people in this manner – by causing the very harm we trust them to prevent – may also increase our sense of vulnerability in the world and arouse a variety of negative emotions. Indeed, we show that the negative emotions associated with feelings of broken trust from an exploding airbag are mediated by perceptions of breakdown in social order (Koehler & Gershoff 2003, Study 4).

If the negative consequences of safety-product betrayals reach beyond the immediate harm – that is, if these betrayals produce multiple harms similar to those that arise when intentional human actors betray – then it is not clear that people's safety-product preferences should be judged against a benchmark that only considers aggregate risks of the immediate harm. In fact, we thought that our empirical results were striking not because they showed how an otherwise reasonable heuristic could lead to absurd preferences, but because they indicated that the consequences of var-

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

ious types of betrayal are so unbearable that people are willing to incur substantial costs to eliminate them. Certainly it would be unreasonable if people chose to avoid betrayal risks at *all* costs. But we are not persuaded that a finding that people are willing to incur *some* additional cost to avoid betrayal provides sufficient evidence of a moral heuristic gone awry.

Moral heuristics or moral competence? Reflections on Sunstein

John Mikhail

Georgetown University Law Center, Washington, DC 20001.
jm455@law.georgetown.edu http://www.law.georgetown.edu/curriculum/tab_faculty.cfm?Status=Faculty&Detail=2065

Abstract: By focusing on mistaken judgments, Sunstein provides a theory of performance errors without a theory of moral competence. Additionally, Sunstein's objections to thought experiments like the footbridge and trolley problems are unsound. Exotic and unfamiliar stimuli are used in theory construction throughout the cognitive sciences, and these problems enable us to uncover the implicit structure of our moral intuitions.

After a period of neglect, philosophers and psychologists have begun to focus on the problems of descriptive and explanatory adequacy in the moral domain. These problems can be represented by perceptual and acquisition models of moral intuition similar to those utilized in the study of language (Chomsky 1964). The problem of descriptive adequacy in the theory of moral cognition seeks to explain the human perceptual ability to individuate and interpret novel acts and omissions and to recognize their moral properties – for example, whether or not they are permissible. The problem of explanatory adequacy, in turn, seeks to explain how this ability is acquired (Mikhail 2000; Dwyer 1999).

Sunstein does not take up ontogeny, but he does seek to explain certain acquired patterns of moral intuition. However, unlike Rawls, whose class of "considered judgments" denotes "those judgments in which our moral capacities are most likely to be displayed without distortion" (Rawls 1971, p. 47), Sunstein focuses attention on judgments he thinks are distorted or mistaken. This reverses the normal order of inquiry, which seeks to understand the ideal operations of a cognitive system before explaining its occasional pathologies and disorders. What Sunstein gives us, in effect, is a theory of performance errors without a corresponding theory of moral competence.

Sunstein recognizes the need for some benchmark to demarcate the set of judgments caused by heuristics. Only then can they be characterized as unsound, unreliable, or mistaken. However, Sunstein's definitions of "weak consequentialism" – his proposed benchmark – are too vague and uncontroversial to do the work required of them in this context. We are left without a clear sense of whether Sunstein thinks moral competence even exists and, if so, which theory adequately describes it.

Echoing a common refrain (e.g., Kaplow & Shavell 2002), Sunstein questions the use of "exotic cases of the kind never or rarely encountered in ordinary life" to reveal the structure of our moral intuitions. The objection sounds plausible, but on reflection seems difficult to understand. Exotic and unfamiliar stimuli are used in theory construction throughout the cognitive sciences. For example, the discovery that infants perceive objects in accordance with principles of cohesion, contact, and continuity utilizes novel displays that depart from previous experience and violate ordinary expectations (e.g., Spelke et al. 1992). Or consider such contrivances as "blicket detectors" (Gopnik & Sobel, 2000), rotating three-dimensional line drawings (Shepard & Metzler 1971), or nonsense expressions like "colorless green ideas sleep furiously" (Chomsky 1957). In any psychology experiment, the decisive question is not whether the stimulus is unfamiliar or artificial, but whether it reveals something interesting about how the mind works. Although some moral dilemmas may indeed be too outlandish to qualify, many of the specific examples Sunstein criticizes appear to satisfy that test.

To take a concrete example, consider the footbridge and trolley problems, which can readily be shown to elicit common deontic intuitions among demographically diverse populations, including young children (Mikhail et al. 1998; Mikhail 2000; cf. Hauser et al., under review). Following Greene et al. (2001), Sunstein suggests that these intuitions cannot be given a principled explanation. Here I think a computational theory of moral competence has been ruled out too soon. In fact, the two cases trigger distinct mental representations whose relevant temporal, causal, intentional, and moral properties can be exhibited in a two-dimensional tree diagram, successive nodes of which bear a generation relation to one another that is asymmetric, irreflexive, and transitive (Goldman 1970; Mikhail 2000). As these diagrams reveal, the key structural difference between the two cases is that, in the footbridge condition, the agent commits a series of distinct trespasses prior to and as a means of achieving his good end (Fig. 1), whereas in the trolley condition, these violations are subsequent and foreseen side effects (Fig. 2).

The computational hypothesis holds that when people encounter the footbridge and trolley problems, they spontaneously compute unconscious representations like those in Figures 1 and 2 (Mikhail, in press). Note that in addition to explaining the relevant intuitions, this hypothesis has further testable implications. For example, we can investigate the structural properties of the underlying representations by asking subjects to evaluate certain probative descriptions of the relevant actions. Descriptions using the word "by" to connect individual nodes of the tree in the downward direction (e.g., "D turned the train by throwing the switch," "D killed the man by turning the train") will be deemed accept-

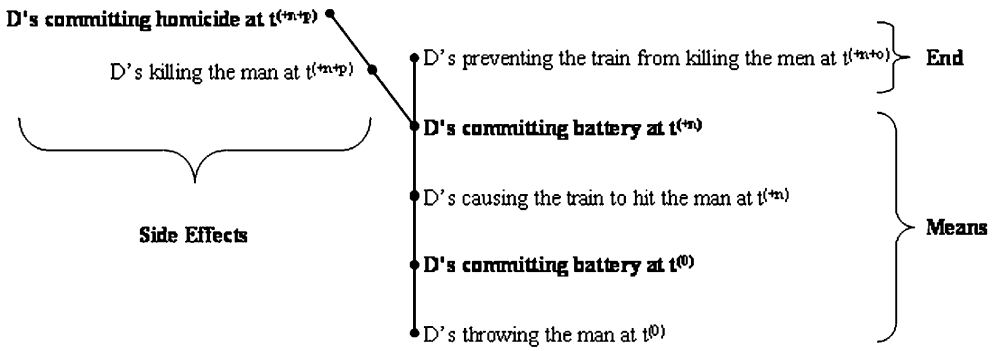


Figure 1 (Mikhail). Mental representation of footbridge problem (Mikhail, in press).

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

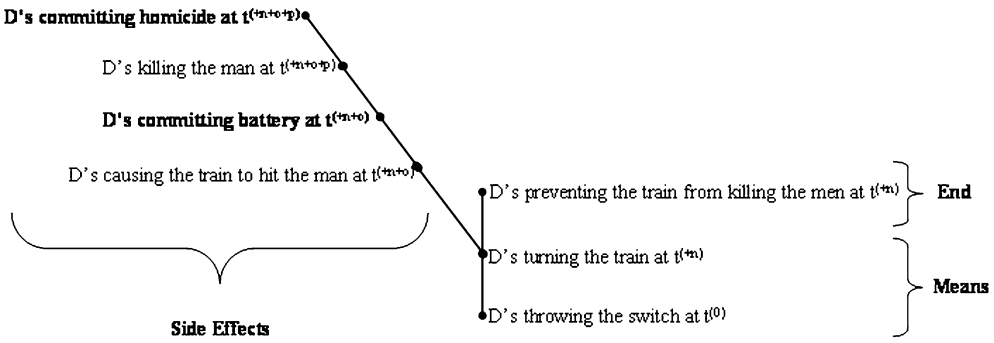


Figure 2 (Mikhail). Mental representation trolley problem (Mikhail, in press).

able; by contrast, causal reversals using "by" to connect nodes in the upward direction ("D threw the switch by turning the train," "D turned the train by killing the man") will be deemed unacceptable. Likewise, descriptions using the phrase "in order to" to connect nodes in the upward direction along the vertical chain of means and ends ("D threw the switch in order to turn the train") will be deemed acceptable. By contrast, descriptions of this type linking means with side effects ("D threw the switch in order to kill the man") will be deemed unacceptable. In short, there is an implicit geometry to these representations, which Sunstein neglects but an adequate theory can and must account for.

"The law has long used actors' intent or purpose to distinguish between two acts that may have the same result" (Vacco vs. Quill 1997, p. 802). Simple but revealing thought experiments like the footbridge and trolley problems suggest that ordinary mortals do so as well. Perhaps this explains why so many legal doctrines turn on an analysis of purpose and on the distinction between intended and foreseen effects (Mikhail 2002). Of course, some of these doctrines may constitute the kind of overgeneralization Sunstein usefully warns against. But many others presumably do not. Consider the norms of proportionality and noncombatant immunity in the law of armed conflict, which limit the permissibility of harming civilians as a side effect of an otherwise justifiable military operation and categorically prohibit directly targeting them. Are these norms the product of heuristics, or of shared principles of moral competence? The fact that we can seriously contemplate the latter alternative – that cognitive science and human rights can be linked in this manner – is significant and worth reflecting upon. In the final analysis, Sunstein's insistent homunculus may be the human sense of justice, which behaviorism in all its varieties leads us to ignore, but which we persistently disregard at our own peril.

Do normative standards advance our understanding of moral judgment?

David A. Pizarro[a] and Eric Luis Uhlmann[b]

[a]*Department of Psychology and Social Behavior, University of California – Irvine, Irvine, CA 92697-7085;* [b]*Department of Psychology, Yale University, New Haven, CT 06520.* **dpizarro@uci.edu** **eric.uhlmann@yale.edu**

Abstract: Sunstein's review of research on moral heuristics is rich and informative – even without his central claim that individuals often commit moral errors. We question the value of positing such a normative moral framework for the study of moral judgment. We also propose an alternative standard for evaluating moral judgments – that of *subjective rationality*.

Sunstein wants to extend Kahneman et al.'s (1982) thesis that generally adaptive cognitive heuristics also lead to systematic and predictable errors in judgment, and makes the provocative argument that moral heuristics can "lead to mistaken and even absurd moral judgments" (target article, Abstract). Sunstein makes an important contribution to the literature on moral judgment by highlighting the role of intuitions in everyday moral thinking (see also Haidt 2001). Although Sunstein does not endorse any grand moral theory explicitly (e.g., Utilitarianism or Kantianism), he agrees that the very concept of a "moral error" requires a normative benchmark, and endorses "weak consequentialism" as being, in his view, a relatively uncontroversial standard by which to judge the successes and failures of various moral judgments.

We do not wish to debate the virtues and vices of any normative moral theory – this is a task best left to philosophers. However, we do question the necessity of positing a normative framework for understanding the psychology of moral judgment. Does a good theory of moral judgment require an objectively "right" set of moral criteria with which to compare lay judgments? Perhaps not. We believe that the research reviewed by Sunstein is extremely informative without the additional claim that individuals are making mistakes. For example, knowing and predicting the conditions under which individuals rigidly adhere to principles despite consequences is important for any successful moral theory. So the fact that individuals are willing to accept a (slightly) increased risk of dying in order to punish a betrayal is quite provocative – but does it add more value to claim that this is an error?

One possible downside of such an approach is a proliferation of error-focused work in the moral domain – a domain in which claiming an objective standard may simply lead to a whole lot of argument about which standard is right, at the expense of paying attention to the data. In our opinion, this was equally problematic with the approach of Kohlberg and his colleagues (cf. Kohlberg 1969) – a willingness to embrace a Kantian/Rawlsian theory of justice led to the questionable claim that certain individuals were at a "lower stage" of moral reasoning. Much like focusing on Kantian justice, focusing on moral errors may divert attention away from more fruitful areas of inquiry, such as (for example) cross-cultural differences in moral judgment (e.g., Haidt et al. 1993), or the emotional processes that underlie moral judgments (e.g., Pizarro 2000).

This does not mean that psychologists must abandon all talk of error in moral judgment – there is one sense of the word "error" that may still be useful in this domain. To the extent that people's moral judgments are influenced by factors that *even they perceive as irrational*, their judgments may be said to be in error (Kruglanski 1989). Empirical examples of this *subjective irrationality* in moral judgment are already available. For example, people believe that they punish to deter future criminals, yet their judgments are driven by the severity of the crime, not deterrence-related variables (Carlsmith et al. 2002; Sunstein refers to this as the "moral outrage" heuristic). Presumably, if a participant in this research was aware of this influence she would revise her judgment, as it fails to match her own standard.

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

In another study, Pizarro et al. (2003) found that participants discounted blame for intentional actions that were not carried out quite as intended (i.e., acts that lacked "intentions-in-action"; Searle 1983). For example, when a murderer tripped and accidentally stabbed his victim in the process of attempting to kill him, he was perceived as less blameworthy. Interestingly, when asked to give their most rational response, participants judged acts that did and did not possess intention-in-action to be equally blameworthy. This suggests that, at least for some, discounting blame for acts that lacked intention-in-action was subjectively irrational.

In another example, Tetlock et al. (under review) examined conservative and liberal managers' reactions to a hypothetical employee error (failure to mail a package on time) with either mild or severe consequences. Both conservative and liberal managers judged the employee more harshly when the consequences of the error were severe (this has been referred to as an "outcome bias" and "moral luck"; Baron & Hershey 1988). Liberals viewed this outcome bias as an error, and reduced their recommended punishment in the severe consequences case when asked to consider how they would have reacted had the consequences been mild. In contrast, conservatives saw it as perfectly appropriate to determine the employee's punishment based on the consequences of his or her actions.

Liberals and conservatives also disagree regarding whether certain socialized intuitions are rational. Ingenious studies by Jonathan Haidt and his colleagues demonstrate that most people find it intuitively wrong to wash one's toilet with the American flag, eat one's recently expired pet, or masturbate into a dead chicken (Haidt 2001; Haidt et al. 2003). When asked to make the most rational judgment possible, liberals appear to correct for their intuitions – reducing blame for eating Fido, for example (Uhlmann et al., in preparation; see also Haidt & Hersh 2001). In contrast, conservatives provide essentially the same judgments when asked to respond rationally versus intuitively. For liberals, the judgments identified by Haidt exert a subjectively irrational influence on their judgments. But for conservatives, who place a high priority on traditional values, such judgments may seem perfectly well-grounded.

If people are indeed exhibiting "absurd moral judgments" (target article, Abstract), we suggest that this is not because heuristics lead individuals' moral judgments to diverge from some objective standard of morality (such as weak consequentialism), but because these judgments would be deemed irrational by the participant himself upon reflection. Perhaps this sense of the term "error" may be the best way to avoid the morass of subjectivity inherent in studying the moral judgments of other people, and may also keep researchers from hurling insults at each other's normative theories of choice.

ACKNOWLEDGMENT

We thank Andy Poehlman for his comments on an earlier draft of this article.

Cognitive heuristics and deontological rules

Ilana Ritov

School of Education, Hebrew University, 91905 Jerusalem, Israel.
msiritov@mscc.huji.ac.il

Abstract: Preferences for options that do not secure optimal outcomes, like the ones catalogued by Sunstein, derive from two sources: cognitive heuristics and deontological rules. Although rules may stem from automatic affective reactions, they are deliberately maintained. Because strongly held convictions have important behavioral implications, it may be useful to regard cognitive heuristics and deontological rules as separate sources of nonconsequential judgment in the moral domain.

The idea of error-prone heuristics is especially controversial in the moral domain, as Sunstein notes, although examples of choices that violate consequential principles are abundant. Among those examples are the "punishment" of companies for cost–benefit analyses to determine their investment in safety, the betrayal aversion, the resistance to "tampering with nature," and the rejection of probability of detection as a normative factor in determining punitive damages. These choices have grave consequences for the lives and well-being of many people, and the contribution of this article in drawing attention to these problems is highly important.

To ascertain that those nonconsequential judgments result from the application of mental heuristics, it is necessary to address the question of what a heuristic is. The notion of a heuristic is not well defined in the psychological literature. As Sunstein notes, Tversky and Kahneman (1984) used the term heuristic to refer to a strategy that "relies on a natural assessment to produce an estimation or a prediction." These strategies take on the form of mental shortcuts, or general purpose rules, often applied without consciousness, in judgmental tasks requiring assessment of unknown values. More recently, the evolving research on dual process theories led to a broader view of the nature of heuristics. Heuristics have come to be equated with processes of System I. This system, also referred to as the experiential system, operates automatically and effortlessly, is oriented to concrete images, and responds affectively. By contrast, the rational system, or System II, operates consciously and effortfully, and is deliberate and reason-oriented (Epstein & Pacini 1999).

In the current broad view of heuristics, not only are estimates of quantities by rules of thumb seen as the products of heuristics, but any expression of preference derived through the experiential system is regarded as such, as well. Although the boundaries of the set have not been explicitly delineated, the most notable feature of a heuristic process that distinguishes it from the cognitive processes classified as reasoning or rational is its nondeliberative nature. Although the outcomes of a heuristic can be deliberately adopted by System II, judgment by heuristic is typically an intuitive and unintentional process (Kahneman & Frederick 2002). It is usually passive and preconscious.

Returning to the examples discussed by Sunstein in the present article, these can arguably be roughly classified into two kinds: the ones that reflect the use of general cognitive heuristics in judgments (applied in the moral domain), and others that deliberate application of rules. The clearest example of a non-deliberative heuristic is the outrage heuristic in punishment. Although people are certainly aware of their outrage, they are most likely not aware of using this emotional reaction as the primary, or even the sole determinant of the punishment they set.

The resistance to cloning, stemming from the conviction that one should not "play god," or "tamper with nature," is an example of the second kind. Although the belief itself may stem from an emotional reaction, it is explicitly adopted by the rational system. The principle is held consciously and deliberately. It is relatively abstract and context-general. Similarly, the rejection of the role of probability of detection in setting punitive damages is the result of deliberate processing, often by expert and sophisticated respondents (Sunstein et al. 2000). In both of these examples, as well as in other ones, the judgment is determined by a deontological rule.

Deontological rules are rules that concern actions rather than consequences. These rules are often associated with values that people think of as absolute, not to be traded off for anything else (Baron & Spranca 1997). These protected values, compared to values that are not absolute in this way, have various predicted properties, such as insensitivity to quantity: The amount of the harm done when they are violated does not matter as much as for other values. Furthermore, in judgments involving a deontological rule or a protected value, the participation of the actor is crucial, even when the consequences are the same. The tendency to punish companies that base their decisions on cost–benefit analysis, even if a high valuation is placed on human life, may reflect the agent relativity characteristic of the rule "do not trade human life for money."

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

As protected values are related to deontological rules against action ("do not play god," "do not tamper with nature," "do not cause death," etc.) they tend to amplify omission bias (Ritov & Baron 1999). If a person has a protected value against, for example, destroying species, this value seems to apply to action rather than inaction. That person might be unwilling to take an action that would cause the extinction of one species, in order to save five, even when the relative outcomes are fully spelled out. By contrast, another person, who cares just as much about preventing the extinction of species as the first person, but not as a protected value, would prefer that action be taken in order to achieve better consequences as a whole. Although the values people hold protected vary considerably, the basic finding of greater omission bias for protected values holds across a wide array of issues, ranging from endangered species, to withdrawal from occupied territories (for Israeli respondents). In all those cases, people holding protected values deliberately preferred omission, despite the fact that they knew explicitly that action would yield better consequences with respect to the specific problem.

The origin of deontological rules is the subject of much research. They may be the result of generalization from a range of problems. Deontological rules are undoubtedly closely linked with affect, but it remains an open question whether their impact is fully mediated by emotions. Even if espousing that deontological rules are primarily an expression of extreme affect, the judgmental process is different from other experiential processes in its explicit and deliberate nature. Until further research provides better understanding of those processes, it may be more useful to regard cognitive heuristics and deontological rules as separate sources of nonconsequential judgment in the moral domain.

Intuitions, heuristics, and utilitarianism

Peter Singer

University Center for Human Values, Princeton University, Princeton, NJ 08544; and Centre for Applied Philosophy and Public Ethics, University of Melbourne, Victoria 3010, Australia. **psinger@princeton.edu**

Abstract: A common objection to utilitarianism is that it clashes with our common moral intuitions. Understanding the role that heuristics play in moral judgments undermines this objection. It also indicates why we should not use John Rawls' model of reflective equilibrium as the basis for testing normative moral theories.

At one point Cass Sunstein suggests that his assertion that heuristics play an important role in our moral judgments does not really favor one side or the other in the debate between utilitarians and deontologists:

> If moral heuristics are in fact pervasive, then people with diverse foundational commitments should be able to agree, not that their own preferred theories are wrong, but that they are often applied in a way that reflects the use of heuristics. Utilitarians ought to be able to identify heuristics for the maximization of utility; deontologists should be able to point to heuristics for the proper discharge of moral responsibilities; those uncommitted to any large-scale theory should be able to specify heuristics for their own more modest normative commitments. (target article, sect. 1, para. 4)

This seems to me to lean too far towards normative neutrality. Seen against the background of a long-running debate in normative ethics, Sunstein's illuminating essay gives support to utilitarians, and not to deontologists.

A major theme in normative ethics for the past two centuries has been the debate between those who support a utilitarian, or more broadly consequentialist, normative ethical theory and those who ground their normative ethics on our common moral judgments or intuitions. In this debate, the standard strategy employed by deontologists has been to present examples intended to show that the dictates of utilitarianism clash with moral intuitions that we all share – and that fit with deontological views of ethics. A famous literary instance occurs in Dostoyevsky's *The Karamazov Brothers*. Ivan challenges Alyosha to say whether he would consent to build a world in which people were happy and at peace, if this ideal world could be achieved only by torturing "that same little child beating her chest with her little fists." Alyosha says that he would not consent to build such a world on those terms (Dostoyevsky 1879). Hastings Rashdall purported to refute hedonistic utilitarianism by arguing that it cannot explain the value of sexual purity (Rashdall 1907, p. 197). H. J. McCloskey, writing at a time when lynchings in the American South were still a possibility, thought it a decisive objection to utilitarianism that the theory might direct a sheriff to frame an innocent man in order to prevent a white mob from lynching half a dozen innocents in revenge for a rape (McCloskey 1957). Bernard Williams invited utilitarians to ponder a similar example, of a botanist who wanders into a village in the jungle where twenty innocent people are about to be shot. He is told that nineteen of them will be spared, if only he will himself shoot the twentieth (Williams 1973).

Initially, the use of such examples to appeal to our common moral intuitions against consequentialist theories was an ad hoc device lacking meta-ethical foundations. It was simply a way of saying: "If Theory U is true, then in situation X you should do Y. But we know that it would always be wrong to do Y, therefore U cannot be true." This is an effective argument against U, as long as the judgment that it would always be wrong to do Y is not challenged. But the argument does nothing to establish that it is always wrong to do Y, nor what a sounder theory than U would be like.

John Rawls took the crucial step towards fusing this argument with an ethical methodology when he argued that the test of a sound moral theory is that it can achieve a "reflective equilibrium" with our considered moral judgments. By "reflective equilibrium" Rawls meant that, where there is no inherently plausible theory that perfectly matches our initial moral judgments, we should modify either the theory, or the judgments, until we have an equilibrium between the two.

The model here is the testing of a scientific theory. In science, we generally accept the theory that best fits the data, but sometimes, if the theory is inherently plausible and fits some of the data, we may be prepared to accept it despite its failure to fit all the data. We assume, perhaps, that the outlying data are erroneous, or that there are undiscovered factors at work in that particular situation. In the case of a normative theory of ethics, Rawls assumes that the raw data are our prior moral judgments. We try to match them with a plausible theory, but if we cannot, we reject some of the judgments, and modify the theory so that it matches others. Eventually the plausibility of the theory and of the surviving judgments reach an equilibrium, and we then have the best possible theory. In this view, the acceptability of a moral theory is not determined by the internal coherence and plausibility of the theory itself, but rather, to a significant extent, by its agreement with those of our prior moral judgments that we are unwilling to revise or abandon. In *A Theory of Justice*, Rawls uses this model to justify tinkering with his original idea of a choice arising from a hypothetical contract, until he is able to produce results that are not too much at odds with our ordinary ideas of justice (Rawls 1951; 1971, p. 48).

The model of reflective equilibrium has always struck me as dubious. The analogy between the role of a normative moral theory and a scientific theory is fundamentally misconceived (Singer 1974). Our common moral intuitions are not "data" in the sense that a series of measurements of the positions of electrons may be data that any credible scientific theory must explain. A scientific theory seeks to explain the existence of data that are about a world "out there" that we are trying to explain. Granted, the data may have been affected by errors in measurement or interpretation, but unless we can give some account of what the errors might have been, it is not up to us to choose or reject the observations. A normative ethical theory, however, is not trying to explain our common moral intuitions. It might reject all of them, and still be su-

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

perior to other normative theories that better matched our moral judgments. For a normative moral theory is not an attempt to answer the question "Why do we think as we do about moral questions?" Obviously, that question may require a historical, rather than a philosophical, investigation. On abortion, suicide, and voluntary euthanasia, for instance, we may think as we do because we have grown up in a society that was, for two thousand years, dominated by the Christian religion. We may no longer believe in Christianity as a moral authority, but we may find it difficult to rid ourselves of moral intuitions shaped by our parents and our teachers, who were either themselves believers, or were shaped by others who were.

Similarly – to come at last to Sunstein's article – an understanding of the way in which we tend to use heuristics may cast doubt on the value of our moral intuitions as a test for the acceptability of a normative theory, even when, as Sunstein says, "they are very firm." We may not appreciate that these intuitions are heuristics, nor that in the special situation in which we find ourselves, the heuristic does not give us the right answer. The case against these intuitions gains further support from recent work exploring what is actually going on in the brain when people are considering moral dilemmas. This work enables us to see differences between intuitive, more or less automatic, responses, and those involving cognitive processes (Greene et al. 2001).

Whenever it is suggested that normative ethics should disregard our common moral intuitions, the objection is made that without intuitions, we can go nowhere. There have been many attempts, over the centuries, to find proofs of first principles in ethics, but most philosophers consider that they have all failed. Even a radical ethical theory like utilitarianism must rest on a fundamental intuition about what is good. So we appear to be left with our intuitions, and nothing more. If we reject them all, we must become ethical skeptics. If some of our moral intuitions are heuristics, however, it isn't hard to see how we can criticize them without ending up as skeptics. We need to think about what our underlying values are, and then distinguish these values from the moral intuitions that merely have a heuristic role in furthering them.

A defender of the idea of reflective equilibrium might say that the knowledge that some of our moral intuitions are heuristics can itself be part of the process of achieving a wider equilibrium between a theory and our considered moral judgments (Daniels 1996). That approach renders the model of "reflective equilibrium" relatively innocuous by making it so all-embracing that it can include any grounds for rejecting intuitions, even, in the limiting case, grounds for rejecting all of them. Now, the "data" that a sound moral theory is supposed to match have become so changeable that they are no longer a barrier to the acceptability of utilitarianism. In that form, there is no need to object to reflective equilibrium.

As Sunstein notes, utilitarians from Mill and Sidgwick onwards have discussed the role of rules and intuitions in moral thinking, sorting through which intuitions we ought to preserve because they are conducive to the larger utilitarian goal, and which we should reject. Sunstein has added a more sophisticated understanding of heuristics to this tradition, thereby helping to refute the objection most commonly invoked against utilitarianism as a normative ethical theory. Admittedly, as Sunstein himself notes, there are many further objections, not based on an appeal to common moral intuitions, which can be asked about utilitarianism. Debate in normative ethical theory will continue. Perhaps, however, one chapter is drawing to a close.

Wide reflective equilibrium as an answer to an objection to moral heuristics

Edward Stein

Benjamin N. Cardozo School of Law, New York, NY 10003.
ed@edstein.com **http://www.edstein.com**

Abstract: If, as is not implausible, the correct moral theory is indexed to human capacity for moral reasoning, then the thesis that moral heuristics exist faces a serious objection. This objection can be answered by embracing a wide reflective equilibrium account of the origins of our normative principles of morality.

Sunstein's central thesis is that moral heuristics, shortcuts that sometimes lead to serious errors, exist. This comment considers a serious conceptual objection to this thesis and provides an answer to it.

It is important to distinguish between determining what the right moral principles are and determining how humans reason about morality. Sunstein focuses primarily on the second project; for example, he shows we favor inaction over a statistically preferable action. Sunstein's thesis has implications for the first project: that is, by identifying rules of thumb that lead us to make moral mistakes, he is implicitly suggesting what the correct moral principles are. Making the distinction between moral norms and human moral competence does not entail that the two must diverge or that they are unrelated. For comparison, consider researching whether humans have a good sense of humor. Imagine researchers telling subjects jokes (some funny and some not), then asking them whether each joke was funny, and, on the basis of their assessments, concluding that humans have (or lack) a good sense of humor. Something seems seriously amiss about this research project, in part because (a) there might be no objective standards of funniness, perhaps because funniness is relative to individual tastes; (b) there might be objective standards of funniness, but they might be indexed to human faculties; and (c) there might be objective, human-independent standards of funniness, but we might lack access to them (Stein 1996, pp. 29–34). In the context of morality, these positions are: (a) relativism about morality, (b) norms of morality are indexed to human moral competence, and (c) moral norms are epistemologically inaccessible. The significance of moral heuristics would change dramatically if any of these positions were true. Sunstein implicitly rejects (a) and (c) by appealing to and identifying objective moral truths, but he does not discuss (b). If norms of morality are indexed to moral capacity, then it is impossible that we make systematic moral mistakes; if our moral competence sets the standards of morality, then systematic moral errors are not possible. This would completely undercut Sunstein's central thesis.

The comparison to sense of humor is just one reason for thinking that this conceptual argument against moral heuristics warrants examination. A similar objection arises in the context of human reasoning more generally. Our use of the availability heuristic and other cognitive shortcuts suggests that human reasoning competence diverges from the norms of reasoning. Despite psychological evidence that cognitive heuristics exist, some have argued that human reasoning competence necessarily matches the norms of reasoning because such norms are indexed to our cognitive capacities (Cohen 1981; Macnamara 1986). This argument is even more plausible in the context of morality, because, as Sunstein admits, it is harder to demonstrate that an intuitively plausible reasoning practice leads to serious errors in the context of morality than in the context of probability or logic.

This conceptual argument against moral heuristics can be refuted by focusing on the origins of moral norms. Sunstein briefly mentions narrow and wide reflective equilibrium, but a more detailed discussion is necessary. Narrow reflective equilibrium is achieved when a set of first-order judgments is coherently systematized by (that is, brought into balance with) a set of general principles. This would be accomplished in the moral realm by ar-

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

ticulating a set of ethical principles from which all and only our somewhat altered and refined moral judgments follow. Wide reflective equilibrium is achieved when a set of first-order judgments, a set of general principles, and a set of philosophical theories (e.g., theories of personal identity, metaphysics, and the social role of moral and political theory) are brought into agreement. The process of wide reflective equilibrium begins with the search for a narrow reflective equilibrium, but once first-order judgments and general principles are brought into agreement, alternative pairings of judgments and general principles are considered with an eye towards bringing them into balance with philosophical theories through the reflective equilibrium process at the more abstract level. This process, rather than producing a systematization of our revised first-order judgments, is more revisionary. Also, the results of wide reflective equilibrium have a broader network of support, including significant philosophical backing. There is, thus, a greater likelihood that wide reflective equilibrium will produce a theory that diverges from intuitions (Daniels 1979; Stein 1996, Ch. 5).

If, as some have argued, our moral norms result from wide reflective equilibrium (Daniels 1979; Rawls 1974–1975), insofar as moral heuristics exist, they are likely to be rejected by a wide reflective equilibrium process because of its revisionary character. Consider, for example, Derek Parfit's discussion of utilitarianism (Parfit 1984). A standard objection to utilitarianism is that it requires balancing losses and gains between people. Such interpersonal balancing is seen as problematic because individual persons are the relevant units for moral theory. According to this standard objection, utilitarianism should be rejected because it requires interpersonal balancing. Parfit tries to show that persons are not the relevant units for moral theory. Rather, according to Parfit's metaphysical arguments, psychological continuity and connectedness are what matters to moral theory. Since one may be psychologically connected to other people, benefits and harms, pleasures and pains can be balanced among various people. If Parfit's metaphysical argument is right, then the original objection to utilitarianism is not strong. Even though utilitarianism might be rejected in narrow reflective equilibirum, according to Parfit, in light of metaphysical theories about personhood and theories of value, utilitarianism will be supported by a wide reflective equilibrium process in the moral context.

Sunstein needs to defend his thesis that moral heuristics exist against the objection that the norms of morality are indexed to human moral competence. Wide reflective equilibrium provides an answer to the conceptual argument that humans must be rational, even though psychological evidence suggests that general reasoning competence diverges from the norms (Stein 1996, Ch. 5; Stich 1990, Ch. 4). A wide reflective equilibrium account of the norms of morality similarly provides a strong answer to this objection to Sunstein's thesis.

Gauging the heuristic value of heuristics

Philip E. Tetlock

Haas School of Business, University of California, Berkeley, CA 94720-1900.
tetlock@haas.berkeley.edu
http://www.haas.berkeley.edu/faculty/tetlock.html

Abstract: Heuristics are necessary but far from sufficient explanations for moral judgment. This commentary stresses: (a) the need to complement cold, cognitive-economizing functionalist accounts with hot, value-expressive, social-identity-affirming accounts; and (b) the importance of conducting reflective-equilibrium thought and laboratory experiments that explore the permeability of the boundaries people place on the "thinkable."

I appreciate – as much as any good Lakatosian – the importance of pushing powerful research programs to their limits (Tetlock 2002). Sometimes the surest way to discover when we have had enough of an approach is to help ourselves to more than enough. I also appreciate how tricky it is to ascertain when we have reached the point of diminishing marginal epistemic returns. Nonetheless, I am prepared to speculate that Sunstein's argument brings us right up to the inflection point, and perhaps beyond.

This may seem a dubious compliment. But, as Lakatosians well know, it is high praise. The heuristics-and-biases research program is one of the most phenomenally successful programs in the behavioral and social sciences. I agree with Sunstein that there is a great deal of suggestive evidence – laboratory and field – that people often rely on simple moral heuristics in reaching conclusions on complex issues. And I agree that, as a result, there is obvious potential for rhetorical and policy mischief.

But there are also good reasons for wariness, some of which Sunstein himself endorses. Certain of these objections have been trotted out so many times that they feel a tad hackneyed. The list of heuristics is indeed endless, ranging from the idiosyncratic (the Scalia heuristic) to the extremely general (equity, equality, procedural, and retribution heuristics). From a theory-building perspective, it is necessary to advance beyond lists and articulate: (1) a taxonomy of heuristics so that we do not come quite so close to a one-to-one correspondence between explanans and explanandum; and (2) a set of hypotheses that specify when particular types of heuristics are likely to be activated.

From a theory-building perspective, it will also eventually be necessary to circumscribe the explanatory applicability of heuristic-based explanations. I do not challenge the cognitive economy gained from heuristics, but I wonder whether the language of heuristics is too coldly cognitive to capture the red-hot passions implied by betrayal, revenge, and disgust. When people categorically declare certain transactions or even forms of thought (such as cost–benefit analysis) off-limits, they are not just simplifying their decision task. They are rising to the defense of sacred values by refusing even to consider taboo trade-offs. Perhaps the most intriguing property of taboo trade-offs is that they are morally corrosive. Decent people do not contemplate the advantages of "playing God," or of "improving returns to shareholders by killing a handful of customers," or of "shortening airport security lines by ethnic-racial profiling." Mere contemplation of such possibilities (failure to tell those who float such ideas to "go to hell") is enough to undercut one's standing in one's moral community (Tetlock et al. 2000).

The moral outrage triggered by such policy proposals is much more than a heuristic; it is a powerful emotion-laden statement about who we collectively are and the types of conduct we countenance (McGraw & Tetlock 2005; McGraw et al. 2003). This makes it easy for moral arguments to degenerate into name-calling. Leon Kass's prose turns progressively brighter shades of purple as he reaches for terms to characterize those "shallow souls" who have the temerity to think beyond the boundaries he wants to place on stem-cell research and genetic engineering: bizarre, grotesque narcissism, and Frankensteinian hubris. It is small wonder that when ordinary folks are caught straying into the territory of the unthinkable, they rush to repair the damage to their social identities through symbolic acts of moral cleansing (Tetlock 2003).

My deepest point of agreement with Sunstein is that he – unlike some cognitive psychologists whose skepticism of human intuition runs truly deep – has not abandoned all hope in the "method of reflective equilibrium." Imperfect though it may be, reflective equilibrium is the best tool at our disposal for encouraging thoughtful responses to cognitive dissonance. Dissonance theory predicts that people gravitate toward low-effort strategies of resolving inconsistency, such as denial of the weaker cognition and bolstering of the stronger, thereby producing a spreading of alternatives. But what happens when the clashing cognitions are equally powerful and neither denial nor bolstering is feasible. In the value-pluralism model (Tetlock 1986), I sketched the conditions under which people will respond to value conflict by engaging in more effort-demanding, System II cognitive maneuvers,

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

such as delineating boundary conditions for competing principles and merging those principles into higher-order composites.

From this standpoint, there is value in conducting reflective-equilibrium experiments even in "exotic" cases. I have done some small-scale experiments along these lines and they reveal how flummoxed people become when researchers design head-on collisions between powerful moral intuitions. Pilot work has shown that, although most people initially agree with Kass's arguments against biotechnology, their opposition even to currently far-out proposals, such as designer babies, is not absolute. Opposition significantly tapers off when we pit Kassian categorical imperatives against countervailing pragmatic pressure (e.g., a major international competitor, China, moving ahead with "modifying the genome of its population" and raising its average IQ to 165, thereby dominating high technology, and sweeping both the Nobel prizes and the Olympics). And most of the remaining opposition is confined to resisting the premise ("that just is not possible"), raising the possibility that if the "impossible" proved possible, they too might change their minds. Few feel comfortable consigning their descendants to perpetual inferiority. History, it is useful to remember, offers many precedents for overwhelming majorities turning into eccentric minorities.

In sum, Sunstein is right that much moral reasoning is more rigid and simplistic than we academics like. But heuristics are but one component of a comprehensive explanation. "Rigidity" also serves valuable self-control and social-solidarity functions. And people are far from hopelessly rigid; they can be quite flexible when reality demands it and politicians obligingly provide the right rhetorical framing.

Towards a taxonomy of modes of moral decision-making

Elke U. Weber[a] and Jessica S. Ancker[b]

[a]*Center for the Decision Sciences, Columbia University, New York, NY 10027;* [b]*Department of Sociomedical Sciences, Mailman School of Public Health, Columbia University, New York, NY 10032.*

euw2@columbia.edu **jsa2002@columbia.edu**

http://www0.gsb.columbia.edu/whoswho/full.cfm?id=55663

Abstract: Sunstein advocates a more systematic approach to the study of moral decision-making, namely the heuristics-and-biases paradigm. We offer two concerns and suggest that a focus on decision processes can add value. Recent research on decision modes suggest that it is useful to distinguish between the qualitative differences in the ways in which moral decisions can be made when they are not made by reflective, consequentialist reasoning.

Because psychological and economic decision researchers have tended to focus on content-independent aspects of judgment and choice (Goldstein & Weber 1995), they have only occasionally discussed decisions with moral implications. Even when such topics have been considered, their treatment has been unsystematic. Sunstein advocates a more systematic approach, namely, to apply Tversky and Kahneman's (1974) heuristics-and-biases research paradigm to moral decision-making. We endorse his goal of a systematic research program, offer two concerns, and suggest fruitful research extensions to Sunstein's call for action.

As Sunstein acknowledges, the heuristics-and-biases research program examined how individuals thought about questions of fact, such as event probabilities. Judgments and decisions could be compared to objective facts, and systematic deviations from normatively correct answer were dubbed *biases*. In the moral domain, such a research program cannot be pursued without a consensus on the normatively correct answer. Sunstein suggests that virtually everyone would agree on the moral superiority of the *weak consequentialist* perspective, which should thus be treated as the normatively correct moral model. Yet he is forced to acknowledge that the weak consequentialist model is not uncontroversial; for example, strong deontologists, such as religious conservatives, might disagree that negative consequences of a moral choice should carry moral weight. Consequently, Sunstein's suggested research program cannot be considered analogous to Kahneman and Tversky's original work in probabilistic reasoning. It must instead be considered the product of a particular ethical worldview. Baron has explicitly acknowledged this in *Thinking and Deciding* (1994b) and *Morality and Rational Choice* (1993b). In both books, he prefaces his discussion of moral decision-making with arguments in favor of utilitarianism as the normatively correct moral framework. Only if utilitarianism is accepted can a heuristics-and-biases interpretation be applied.

Our second concern relates to Sunstein's loose definition of the term "heuristic." He uses the term to denote decisions made by attribute substitution, those made by consulting an authority figure, and those made by recognizing the similarity between the current situation and another for which the decision-maker already has determined the best course of action. Elsewhere, he defines a heuristic as being any form of reasoning other than reflective reasoning. Thus, in Sunstein's article, heuristic reasoning is any decision process that is less cognitively effortful than reflective, consequentialist reasoning and that produces a different outcome (if it produced the same decision, it would not be detected). Simply put, Sunstein seems to blame "heuristics" for all instances in which moral decisions deviate from the weak consequentialist perspective.

These two concerns about Sunstein's arguments are offered in a constructive spirit. The second, in particular, suggests directions for research that will further Sunstein's goal of a better and more systematic understanding of moral decision-making. We review recent research on modes of decision-making and outline implications for the study of moral decisions. A decision-modes approach would subsume the heuristics-and-biases approach into a broader and perhaps less judgmental framework.

Several taxonomies of the variety of processes used to arrive at decisions have recently been suggested (Hammond 1996; Weber & Hsee 2000; Yates & Lee 1996). The modes include the following: (a) reflective, consequentialist reasoning such as utilitarianism (often referred to as calculation-based decision-making, with evaluation of component outcomes and their likelihoods, and integration of such information into a judgment), (b) recognition-and-rule-based decisions, where the situation is recognized as a member of a category or schema for which a judgment or best action has already been stored and behavior is triggered as a production rule (schema-based reasoning has been claimed to be an important component of moral decisions; Narvaez 1999; Rest et al. 1999); (c) story-based decisions, where people construct and evaluate alternative "stories" of what might happen under different courses of action; and (d) affect-based decisions, where people base their decisions on holistic affective reactions to choice alternatives.

Decision modes often operate in parallel and at different speeds, and different modes often (though not always) lead to different decisions. We tend to become aware of the operation of different decision modes when our heads point us in one direction (by calculation-based decision-making), but our hearts point us in another (by affect-based processing). Using the very broad taxonomy offered by Kahneman and Frederick (2002) and others before them (see Table 3 of Stanovich & West 2000), we can identify this situation as a conflict between one of the modes from the fast, associative, and intuitive System I, and another from the analytic System II. Confidence in a decision is inversely related to the degree of conflict experienced as the result of parallel decision processes (Weber et al. 2000). Preference for different decision modes appears to be related to decision domain (e.g., social vs. financial vs. ethical decisions), culture, and goal (e.g., the maximization of material well-being vs. social needs) (Weber et al. 2004). There seems to be social consensus about the desirability of certain modes for specific types of decisions (Ames et al. 2004).

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

What are the implications of decision-mode research for moral decision-making? We suggest that deviations from reflective, consequentialist reasoning should not always be considered errors, but also that decision-mode research can help in the design of interventions or decision aids in situations in which such answers are considered suboptimal. It matters, for example, whether the outrage heuristic cited by Sunstein is an affective response, or the implicit or explicit application of a rule.

A decision-mode approach to the study of moral decision-making would determine the variety of processes by which decision-makers arrive at moral decisions and study how these processes result in different choices. How many different decision modes are there? How do decision-makers select a decision mode? When do decision-makers reason in a consequentialist way, and when do they apply deontological rules? Can choice of decision mode be influenced? What is the role of culture, religion, or political affiliation in determining decision mode? Lumping all of these different modes of decision-making into a single "heuristic" category fails to take advantage of the knowledge conveyed by a process-level analysis of decision-making.

Regulation of risks

Paul Weirich

Philosophy Department, University of Missouri, Columbia, MO 65211.
weirichp@missouri.edu **http://www.missouri.edu/~weirichp/**

Abstract: Sunstein argues that heuristics misguide moral judgments. Principles that are normally sound falter in unusual cases. In particular, heuristics generate erroneous judgments about regulation of risks. Sunstein's map of moral reasoning omits some prominent contours. The simple heuristics he suggests neglect a reasoner's attempt to balance the pros and cons of regulating a risk.

Prejudice, bias, and unreliable general reasoning heuristics yield mistaken moral judgments. Sunstein shows that, in addition, unreliable moral heuristics generate errors. He presents heuristics to explain bad moral judgments about regulation of risks. The heuristics he suggests ignore considerations that many reasoners recognize as relevant. I sketch an alternative, more fine-grained account of the reasoning behind their judgments. However, in agreement with Sunstein, I acknowledge a need for additional psychological studies of moral reasoning.

A heuristic is a principle. Its application to a case may be unreliable and so yield an inaccurate judgment about the case. Sunstein offers several illustrations concerning risk regulation. For various regulatory issues, he suggests a heuristic and points out its unreliability in reaching a judgment about the issue. Do people use the heuristic suggested to reach the judgment presented? A heuristic may yield a judgment and yet not guide the reasoning that people use to reach the judgment. Also, a heuristic may guide some populations but not other populations. Consequently, the suggested heuristics are not full explanations of judgments.

People condemn failures to make cars safer even when the costs of additional safety devices are very high. Sunstein suggests that they follow the moral principle: Do not knowingly cause a human death. He takes this usually reliable heuristic to yield bad judgments about risks. He contends that it is not morally wrong to hold down the production costs of cars by forgoing expensive safety devices that will save only a few lives.

Conflicting moral principles apply to risk regulation. Auto safety triggers, besides principles concerning lives, principles concerning efficient use of resources to improve the standard of living. Judgments may follow one principle to the exclusion of others, but they may also seek a balance between considerations the conflicting principles express.

Typical reasoners do not use the simple heuristic Sunstein suggests. They do not conclude that an auto company knowingly causes highway deaths. Rather, they object to profiting from a disregard for life. Their judgments about safety therefore balance considerations and do not narrowly attend to just one consideration.

A second example considers emissions trading. People condemn the practice despite its effectiveness in reducing pollutants. Sunstein suggests that they misapply the heuristic: People should not be permitted to engage in moral wrongdoing for a fee. The heuristic reliably applies to only immoral acts. Emissions are not immoral when justified by the products whose manufacture generates the emissions. The heuristic falls outside its range of reliability.

Another explanation of the judgment against emissions trading is that people regard pollution as non-cooperative behavior. A balance of reasons supporting cooperation leads them to favor a ban on pollution instead of emissions trading, just as it leads them to favor a no-parking zone in front of a hospital instead of high fees for parking there.

A third example concerns fatalities caused by safety measures such as air bags. Sunstein suggests that people use the heuristic: Punish, and do not reward, betrayals of trust. The heuristic is out of its element because the safety measures, not being agents, do not literally betray anyone. Do people follow a heuristic that fits the case so loosely? Its not applying well is evidence that people do not use it. In fact, they do not apply it to other risky interventions, for example, anesthesia during surgery.

A fourth example comes from the section on playing god (sect. 5.3). It concerns food from genetically engineered crops. Sunstein presents the heuristic: Do not tamper with nature. He uses it to explain public support for regulation of genetic engineering.

Are there other explanations for the public's resistance to genetically modified food? Some come to mind readily. For example, the public may not trust scientific assessments of risks. People may believe that the assessments are unreliable because they are sponsored by industries heavily invested in agbiotechnology. The heuristic suggested identifies a source of caution, but does not fully explain judgments about regulations. People do not mind tampering with nature to halt the spread of tooth decay, for example. Perhaps they use the milder maxim: Tampering with nature is risky. A majority of people may favor regulation to reduce risks they perceive, even when evidence about those risks is incomplete, as Weirich (2001, Ch. 7) explains.

Moral rules of thumb such as "Be honest" acknowledge exceptions such as harmless lies to spare another's feelings. Such maxims have two interpretations. They may express heuristics that are occasionally unreliable. Or, they may express nondecisive reasons for acts. Taking them to express reasons yields a better account of their role in moral deliberations. A moral heuristic attends to a single reason, and exclusive reliance on the heuristic makes that reason decisive. Such narrow-mindedness is unreliable. In the examples concerning risk regulation, errors may arise not from unreliable heuristics, but from overlooking or poorly balancing reasons.

People who consider regulatory issues recognize the complexity of the issues. Sunstein's heuristics oversimplify their reasoning. Their deliberations weigh pros and cons and seek a judgment that is best supported by the reasons behind simple maxims.

A convenient deliberational heuristic makes decisive the reason that looms the largest. Gigerenzer (2000, p. 125) proposes this heuristic, and it reconciles Sunstein's moral heuristics with the multiplicity of reasons concerning regulations. It yields Sunstein's heuristics, given that in his examples they identify the weightiest reasons. A supplementary account of the framing of the regulatory issues may explain the salience of those reasons. The reconciliation just sketched, although intriguing, does not have enough empirical support to dethrone the rival view that moral reasoning about risks balances multiple considerations.

Investigating the reliability of moral reasoning teaches us which moral judgments to trust. Its lessons require accurate identification of the reasoning that yields a judgment. Moral judgments

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

arise in various ways. A judgment may be immediate, rather than a product of a principle's application. Also, many moral principles, of differing reliability, may yield the same judgment. The judgment may be correct even if some principle generating it is unreliable. For example, a moral heuristic and reflection may both yield an intuition about an exotic case. The intuition may be trustworthy on account of support by reflection, despite its arising also from an unreliable heuristic. A strong argument against trusting a judgment requires showing that the judgment rests exclusively on an unreliable process. Moral heuristics are just a step toward the necessary, full account of moral reasoning.

Author's Response

On moral intuitions and moral heuristics: A response

Cass R. Sunstein

University of Chicago, Law School and Department of Political Science, Chicago, IL 60637. **csunstei@uchicago.edu**
http://www.law.uchicago.edu/faculty/sunstein/

Abstract: Moral heuristics are pervasive, and they produce moral errors. We can identify those errors as such even if we do not endorse any contentious moral view. To accept this point, it is also unnecessary to make controversial claims about moral truth. But the notion of moral heuristics can be understood in diverse ways, and a great deal of work remains to be done in understanding the nature of moral intuitions, especially those that operate automatically and nonreflectively, and in exploring the possibility of altering such intuitions through modest changes in context and narrative.

R1. Introduction

I am grateful to the commentators for their exceptionally illuminating discussions of moral intuitions and moral heuristics. A pervasive theme is the automatic and pre-reflective character of many moral judgments. The commentaries raise three questions: (1) Do moral heuristics really produce moral blunders? (2) Isn't there an important difference between heuristics and freestanding moral principles? (3) To make progress, don't we need to settle on some method for establishing moral truth, or at least on the origins of moral judgments by human beings?

Recall that I am understanding moral heuristics to be mental shortcuts, in the form of simple rules of thumb that generally work well, but that also misfire. On this understanding, moral heuristics are parallel to the heuristics that people use when assessing simple questions of fact. It follows that the answer to the first question is yes; moral blunders are often a product of moral heuristics. True, moral heuristics can sometimes lead us to correct moral outcomes on erroneous grounds, as opposed to leading us to morally erroneous outcomes; and both phenomena are important. Of course, there is a difference between heuristics and deeply held principles. What a utilitarian sees as a moral heuristic (never lie!) might be regarded as a freestanding moral principle by a deontologist. But whatever one's view of the foundations of morality, I believe that many apparently freestanding moral principles can be shown to be mere heuristics. For this reason, we can learn a lot about moral reasoning and moral error without getting philosophically ambitious. Nothing in my argument is meant, for example, to suggest that utilitarianism provides the right foundation for morality. It is clear that moral heuristics are pervasive; we need to learn much more about their sources and their nature. Of particular interest are moral judgments that operate automatically and without much reflection from those who make them.

R2. Moral heuristics and moral blunders

Many of the commentaries raise the question of whether moral heuristics produce moral blunders. It is best to begin with an example, which **Koehler & Gershoff** helpfully provide. In their view, betrayal aversion is quite reasonable. (Their excellent work on betrayal aversion helped to inspire my article; should I feel betrayed?) As they note, betrayal aversion makes people willing to *double* their death risk in order to eliminate a betrayal risk. But they contend that this seemingly irrational behavior makes perfect sense, because a "safety-product betrayal" will have all sorts of negative consequences. It may, for example, produce a deep societal mistrust of car manufacturers and government agencies as well as increase our feelings of vulnerability. Koehler & Gershoff conclude that in seeking to avoid the negative consequences associated with betrayals, people are not led astray by a heuristic; they are acting sensibly.

For two reasons, I am unconvinced. The first is that for the argument to work, people must be perfectly altruistic, increasing their own safety risks in order to prevent adverse effects for others. The second reason is more fundamental: **Koehler & Gershoff** have shown only that *betrayal aversion is reasonable given that people are averse to betrayals*. The unfortunate consequences they describe are artifacts of betrayal aversion; the aversion cannot, without circularity, be justified by reference to itself. To be sure, this problem might be reduced if we assume that those who show betrayal aversion are not only perfectly altruistic but also exceedingly sophisticated. Perhaps they know that by itself, betrayal aversion is a form of bounded rationality – but they also know that human beings are boundedly rational. They display betrayal aversion only because they want to protect their fellow citizens from the multiple problems that arise when betrayal risks come to fruition. Even if this is so, a kind of circularity remains: Sophisticated people are willing to display betrayal aversion only because they know that other, less sophisticated people are subject to betrayal aversion. In any case, I much doubt that the subjects in the experiments conducted by Koehler & Gershoff are thinking so elaborately. In short, the special aversion to betrayal risks, which leads people to make themselves less safe, seems to me to be a clear case in which moral heuristics produce errors.

Might some moral intuitions, based on moral heuristics, protect us from moral error? **Bartsch & Wright** think so. Moral maturity, they contend, may not be so different from expertise in skiing, music, or chess, in which extremely rapid judgments usually lead experts to make correct judgments. Bartsch & Wright are right to say that from the moral point of view, heuristic-driven intuitions may do far better than moral principles, even if the latter are based on

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

System II. We can easily imagine monstrous moral principles; a system of morality based on Nazism, for example, runs afoul of widespread and deeply felt moral intuitions. Such intuitions work, every day of every year, to save people from making terrible moral errors. I do not suggest that moral heuristics are more likely to produce error than more systematic moral thinking of any particular sort. And I agree with the suggestion, ventured by Bartsch & Wright, that the most mature moral thinkers respond intuitively and well to a wide range of relevant particulars. But their point should be taken as a celebration of moral maturity, not of moral heuristics as such. It remains true that in numerous domains in morality, politics, and law, such heuristics lead people to err.

Mikhail offers the most sustained challenge to my focus on moral mistakes. He thinks that to identify mistakes as such, we need an account of moral competence. Mikhail objects to my criticism of the use of exotic cases to uncover intuitions; he believes that such cases can teach us a great deal about how the mind works. Perhaps so; but what exactly do we learn? I am not at all sure that those who encounter exotic problems "spontaneously compute unconscious representations" of the complex kind depicted by Mikhail. (It would be very lucky indeed if spontaneous computations turn out so elaborate.) But even if so, Mikhail has said nothing to demonstrate that we should have confidence in the reliability of people's reactions to weird dilemmas never encountered in ordinary life. At the very least, there is a problem if people have different moral judgments about identical (but differently framed) moral problems – and also if people's moral judgments can be shown to be wrong by reference to a fairly uncontroversial moral benchmark. My discussion of moral framing effects and of weak consequentialism is meant to establish that this problem is all too real.

Sharply disagreeing with **Mikhail**, **Pizarro & Uhlmann** believe that the psychology of moral judgment can be studied without any moral benchmark – that it is illuminating and for many purposes adequate to proceed without claiming that people make moral mistakes. With respect to error, Pizarro & Uhlmann suggest that we should focus on whether people's moral judgments are produced by factors that *they themselves believe to be irrational grounds for judgment*. For example, many people believe that their punishment judgments are meant to deter crime, but the evidence suggests that for most people, deterrence plays a modest role, at best, in their punishment judgments. Hence, people might be embarrassed by their own moral intuitions. That is an important possibility, one that those interested in moral heuristics should exploit. But I think that we should be more ambitious. Sometimes people can't easily be embarrassed, but it is very much worth considering the possibility that they have been led astray by a heuristic. We wouldn't want to say that people's immunity from embarrassment is proof that they are on the right track. If we did, that would mean that the louder and more tenacious the homunculus, the less likely System II is to correct the System I error, and the less likely it is that a moral heuristic will be seen to be at work – not the most productive way to approach moral judgments.

Which brings us to **Haidt**'s entertaining essay, contending that policy makers (and others) are likely to resist the claim that they are being led astray by their homunculus. In a way he is obviously right (and he is supported by his own evidence on moral dumbfounding, see Haidt et al. 2004). But I think he is too pessimistic. In environmental law, consider tradable emissions permits, by which companies are given a license to pollute, one that can be bought and sold on the open market. For many years, such permits were rejected on the ground that they violated deeply held moral intuitions. But they are clearly the wave of the future in environmental law, having received bipartisan endorsement on many occasions in the last decade. Squawking all the way, the homunculus is on the run. System II is often a slave to System I, but in politics and law, it's not always clear who is the master.

In this regard, I much appreciate **Herzog**'s emphasis on the inconsistencies in people's thinking about animals. Herzog stresses that the Animal Welfare Act protects dogs but not rats, perhaps because of a heuristic to the effect that "Rats are pests; pests are bad." With respect to animals, many people have immediate, affective reactions that drive their moral judgments; perceived similarity to human beings evidently has a large impact. Indeed, we might speculate that a particular heuristic is operating here for human judgments: *Protect nonhuman animals to the extent that they resemble human beings*. To know whether this heuristic generally works well, we would need to settle on a moral judgment about the appropriate status of animals. But even without so settling, we might well agree, after a little reflection, that the heuristic is likely to misfire in identifiable cases – as, for example, in cases in which a failure to protect nonhuman animals creates preventable suffering. A great deal of work remains to be done on this topic.

R3. Moral heuristics and freestanding moral principles

Some moral principles might be freestanding; they may or may not be moral heuristics. Consider the prohibition on torture; is this a mere heuristic? The answer depends on what morality requires; if torture is always forbidden from the moral point of view, then the prohibition is a freestanding principle, one that does not misfire. Many of the commentators suggest that I have misdescribed ordinary moral reasoning by arguing that it is driven by heuristics when a freestanding principle, or a set of freestanding principles, might be at work instead.

Weirich offers the most ambitious defense of ordinary moral reasoning. He thinks that in the risk regulation cases I discuss, people are not using a simple heuristic; in fact, their reasoning is often quite complex. In the context of automobile safety, for example, he thinks that people object whenever companies profit from disregarding life. In the context of emissions trading, he contends that people see pollution as non-cooperative behavior, and they want to ban that behavior, rather than to allow pollution markets. But Weirich provides no evidence to support his views about how ordinary people reason, and some of his claims seem to me implausible. (Do people really want to ban pollution – and hence to ban both cars and major sources of electricity? Do people really think that hard questions about how to trade off prices and risks can be answered by proclaiming, homunculus-style, that companies shouldn't profit from disregarding life?) Even if his account is right, Weirich has supported my general argument, simply because he has offered an alternative set of heuristics to explain my risk

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

regulation examples. Weirich does not defend his suggestion that people considering regulatory issues "recognize the complexity of the issue" and "weigh pros and cons." On the contrary, his own reconstruction of their reasoning suggests that they fail to weigh pros and cons and instead rely on simple (and crude) heuristics.

Ritov rightly insists on the distinction between heuristics and deontological rules. She understands the former to operate automatically and effortlessly, and also affectively, and therefore to be distinct from the latter, which operate consciously and reflectively. A minor qualification: Some heuristics, such as availability, need not be accompanied by affect. A larger qualification: I believe that many apparent deontological rules are productively seen as heuristics, in the form of simple rules of thumb that generally work well but that also lead to major blunders. The literature on protected values, emphasized by Ritov, strongly supports this understanding; it shows that protected values become less absolute once people's attention is drawn to tradeoffs (Baron & Leshner 2000). People's automatic moral intuitions, based on heuristics, can often be shown to be too crude, even to the satisfaction of the very people who strongly hold those intuitions. One of my major goals, in fact, is to raise the question of whether numerous deontological rules are best seen in this light.

Adler accepts the claim that moral heuristics lead to moral mistakes, but he would like me to endorse consequentialism. He thinks that without some kind of contentious moral position, my examples may not be able to get off the ground. Before reading Adler's discussion, I would have thought that moral framing was the simplest response to this concern. We need not adopt a contentious moral position to think that different moral judgments ought not to be produced by different framing of identical problems. Ingeniously, however, Adler defends framing effects with the suggestion that moral loss aversion might be an effort to economize on the costs of deliberation while also reducing the aggregate level of moral error. Going much further, Adler suggests, correctly, that if weak consequentialism is really very weak, then many of my examples will not work. If we are thoroughgoing retributivists, or believe that harmful acts are much worse than harmful omissions, or don't want to tamper with nature, then my examples do not involve blunders at all.

In the end, **Adler** might be right. In the moral domain, clever readers might well be able to generate plausible, principled, heuristic-free explanations for why people think as they do. But I wonder how many of these explanations can be made convincing. A rule-utilitarian defense of moral framing may not be entirely implausible, but in deciding what morality requires, people can surely do far better than to rely on the easily manipulated distinction between gains and losses. True, some of my examples won't work unless consequences have a nontrivial weight. Those who believe that consequences matter exceedingly little might insist that freestanding moral principles, rather than moral heuristics, are responsible for their belief in, for example, pointless punishment. But my hunch is that unless subjects are extremely stubborn, they can be brought, on reflection, to agree that their moral intuitions have misfired, by their own lights, in many of the relevant cases.

Baron would also like me to adopt a contentious theory – indeed he would like me to endorse utilitarianism (a more controversial idea than consequentialism, which need not accept the idea that all consequences must be described in terms of their effects on utility). Baron is certainly right to say that the fact of disagreement is not a decisive objection to ambitious normative theories. But disagreement is nonetheless a fact; and it is at least useful to show, if it can be shown, that moral heuristics can be described as such, and can be shown to produce error, even when we disagree about many normative questions. There is something to be gained by learning about the uses and abuses of moral heuristics even when people cannot converge on a moral theory. Baron is certainly correct to say that it is productive to demonstrate that if people are concerned about consequences, some moral judgments will make them better, and others will make them worse.

Fried is also interested in the possibility that some moral intuitions are freestanding moral principles rather than mere heuristics. She contends that my own approach favors moral coherence and that it is not, in fact, neutral among different moral principles. In her view, an approach that favors coherence is biased in favor of welfarism, simply because welfarists have an easier time in achieving coherence among their principles (as compared, for example, to deontologists). To know whether a heuristic is at work, she argues that we should shed the idea of coherence in favor of a purely procedural test, to the effect that a freestanding moral principle counts as such only if, after scrutiny, one judges it as an end in itself.

As **Fried** suggests, her test might support many of my examples. But I am not at all sure that she is right to say that my approach is biased in the direction of welfarism; deontological principles can certainly cohere, and deontologists work extremely hard to achieve coherence. Rawls, after all, is a deontologist – indeed, the most powerful critic of utilitarianism – and an understanding of the search for reflective equilibrium (more on this later) is one of his primary contributions to philosophy (Rawls 1971). What Fried misses, I think, is that to identify moral heuristics as such, it is not necessary to make especially ambitious claims about coherence. Far less stringent tests seem to me sufficient in many cases. I have suggested that some heuristic-driven intuitions cannot be defended by reference to any moral theory; others seem wrong on a minimally contentious moral theory, one that does not force people to choose between utilitarianism and deontological approaches.

Sounding much like **Adler**, **Weber & Ancker** object that my approach depends on a consensus on morally correct answers. They contend that the heuristics-and-biases approach can be applied in the context of morality only if utilitarianism is accepted. But this is not so. As Adler demonstrates, we can find heuristics *within* theories; for those who accept any particular theory, heuristics can be shown to be at work in the real world. (Deontologists can find heuristics for what deontology requires; so too for utilitarians; so too for Aristotelians.) And in many cases, I believe that moral heuristics can be identified as such regardless of one's theory. Contrary to what Weber & Ancker suggest, I do not mean to accept utilitarianism. Weber & Ancker are right to say that weak consequentialism is not entirely uncontroversial, but I believe that in many of the cases I discuss, most people would be willing, on reflection, to agree that their intuitive judgment is difficult to defend.

Weber & Ancker also suggest that it is helpful to identify different "modes" of decision-making, including consequential reasoning, decision by affect, story-based decision,

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

and recognition-and-rule-based decisions. The existing research on this count is indeed illuminating. What I would add is that several of their "modes" are strong candidates for analysis as moral heuristics. If, for example, people make decisions by consulting their affective reactions, they are probably using those reactions as a heuristic for something – in the moral domain no less so than in the personal one. (Of course, affective reactions to moral problems can be better, from the moral point of view, than moral principles, for reasons suggested by **Bartsch & Wright**.) And if people assess moral dilemmas by recognizing them as part of some category that triggers a rule, then a heuristic, in the sense of a moral shortcut, is undoubtedly at work.

Hahn, **Frost & Maio** (**Hahn et al.**) rightly emphasize that the term "heuristic" can be understood in many different ways. They point out that, most of the time, I understand moral heuristics to be general propositions that usually work well but that also produce errors in identifiable cases. They worry that this understanding renders my argument essentially pointless. General propositions, as used in morality or even law, typically give rise to undesirable outcomes if they are applied rigidly. Along with **Ritov**, Hahn et al. want to restrict the term "moral heuristics" to processes in which subjects are not aware of the motivations for their judgments. In this way, they seem to think that the study of moral heuristics should emphasize the difference between the system of reasoning and the intuitive system, where processing is unintentional.

I agree that this difference deserves much more investigation, but **Hahn et al.** seem to me to underrate the importance of studying moral rules of thumb that generally work well but that also misfire. In the abstract, it is no news to say that general principles have exceptions. But it is far from pointless to identify a number of cases where intuitive moral objections are extremely strong, and where the moral homunculus is squawking even though no moral wrong is being done. As in cognition, so to in morality: Mental shortcuts can be exceedingly helpful, but they can get us in a lot of trouble, too.

R4. Method: Reflective equilibrium, evolution, and beyond

What is the relationship between moral heuristics and the search for reflective equilibrium? My answer here is cautious. I do not suggest that reflective equilibrium is required to expose heuristics as such; more modest reflection is often enough. Nor do I suggest that an understanding of moral heuristics is incompatible with the search for reflective equilibrium. If we know that moral heuristics are pervasive, we will be aware that some of our most deeply held moral intuitions might be unreliable, and we will scrutinize them in light of the possibility that they are overgeneralizations from ordinarily sound principles. Many of the commentators argue for closer attention to reflective equilibrium and to the foundations of moral judgments.

Tetlock believes in the search for reflective equilibrium, contending that it is the best available tool for producing thoughtful responses to moral problems. In his view, those who attempt to reach equilibrium are ultimately able to enlist System II in favor of more demanding efforts to describe the boundaries of principles and to create what he calls "higher-order composites." For example, Tetlock's own empirical work suggests, importantly, that people's initial agreement with arguments against biotechnology tend to dissipate when they are made alert to tradeoffs (see Tetlock 2000). Other work strongly supports his general point, demonstrating that once people are alerted to tradeoffs and complexities, their strong moral intuitions tend to shift (Baron & Leshner 2000). So far, we are in complete accord. But Tetlock is also concerned that in some cases, what I call moral heuristics are not merely an effort to simplify decision tasks; perhaps people are instead defending sacred values by refusing to consider taboo tradeoffs. I wonder about the sharpness of the distinction between moral heuristics and sacred values, and I think that Tetlock's own work throws that distinction into question (cf. Tetlock 2000). Some sacred values, and some refusals to consider tradeoffs, are best seen as moral heuristics, simplifying decision tasks. Some religious taboos, involving practices of eating and cleaning, can be understood in just this way. I also believe that the moral opprobrium directed against cost-benefit analysis, certain sexual practices, and "tampering with nature" are best seen in the same terms. As Tetlock suggests, the search for reflective equilibrium can be a helpful corrective here.

E. Anderson does not refer to reflective equilibrium as such, but her discussion of John Dewey is not inconsistent with Tetlock's position. She argues that when people are deliberating successfully, they do not abandon heuristics; they use them. In her view, I adopt a "standard reason–emotion dichotomy." Following Dewey, Anderson argues that in contexts that are well-suited to deliberation, our heuristics are not really supplanted; they are incorporated into moral deliberation. I agree with her. I do not mean to accept a reason–emotion dichotomy, in part because emotional reactions are usually based on reasons. (If jurors are outraged by corporate misconduct, it is for reasons.) I hope that nothing I have said is incompatible with Anderson's account of the proper place of heuristics.

In the most philosophically ambitious commentary, **Stein** also wants to emphasize the search for reflective equilibrium, so long as it is wide rather than narrow. By wide reflective equilibrium, Stein means a situation in which our moral judgments are in accord not only with our other moral judgments, including our general moral commitments, but also with our theories of personal identity, metaphysics, and more. Stein believes that it is necessary to focus on the origins of moral norms, and thus on wide reflective equilibrium, to rebut what would otherwise be a strong conceptual attack on the very idea of moral heuristics, to the effect that whatever moral norms exist are indexed to human moral competence. If moral norms are so indexed, moral errors are not possible, and the project of identifying moral heuristics cannot get off the ground. Stein invokes the analogy of humor, where standards of what is funny are indexed to human faculties. He thinks that the search for wide reflective equilibrium can show that moral heuristics exist, because those who seek that equilibrium are perfectly willing to revise their moral views and to concede, on reflection, that they are moral errors.

To identify many such errors, however, I am not sure that it is necessary to seek wide reflective equilibrium. (To paraphrase an unpublished remark by Rawls: "No deep thinking here. Things are bad enough already.") In the case of

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

moral framing (and notwithstanding Adler), mistakes are easy to show without much worrying about reflective equilibrium. And in many of the cases I discuss, no particularly contentious theory is necessary to demonstrate mistakes. Stein's example of humor actually supports my strategy. Even if funniness is indexed to human capacities, there are heuristics for humor; these usually work well, but they also misfire. It's often funny if someone slips on a banana peel, but not if the person who slips has Down's syndrome. In fact, bad television comedies typically fail because they use heuristics for humor. But Stein may be right to say that in order to be certain that moral error has occurred, we will have to become ambitious about the nature and origins of moral norms.

Also concerned with the sources of moral intuitions, **C. Anderson** focuses in particular on the act–omission distinction. In his view, omission bias stems from a distortion in human perception, a distortion that leads people to focus more on the consequences of action than on the consequences of inaction. One reason for omission bias may be an acute awareness of the losses produced by action, alongside relative inattentiveness to the losses produced by inaction. C. Anderson's account might be right. Compare the endowment effect, by which people tend to demand more to give up goods that they own than they are willing to pay for the identical goods when in the hands of others (Thaler 1993). The endowment effect might well be explained by "opportunity cost neglect": People are relatively unconcerned with the opportunity costs of failing to trade a good that they already have. I believe that opportunity cost neglect is a pervasive cognitive phenomenon, and it may help to explain omission bias. In any case, I much appreciate C. Anderson's suggestion that we might be able to isolate the sources of moral error without making contentious philosophical claims.

I anticipated that some readers would use an understanding of moral heuristics as a basis for a large-scale challenge both to deontology and to the search for reflective equilibrium. **Singer** has done exactly that. Where **Pizarro & Uhlmann** and **Fried** object that I have not been neutral on normative questions, and make suggestions for achieving greater neutrality, Singer contends, approvingly, and along the lines of **Baron**, that my article supports utilitarians and undermines deontologists. The reason is that deontologists often argue against utilitarians by demonstrating that, in particular cases, utilitarianism leads to unacceptable outcomes. With Baron, Singer says, rightly, that the problem may lie with our intuitions, not with utilitarianism. Indeed, he goes much further. He thinks that we ought not to test moral judgments by seeking reflective equilibrium, because the intuitions that go into the search for equilibrium may themselves be misconceived. But as I have said, my argument can easily be incorporated into the search for reflective equilibrium; it need not be taken as an attack on that search. Nonetheless, Singer is right to say that utilitarianism cannot be defeated by showing that it leads to results that seem strongly counterintuitive. I agree with Singer that many challenges to utilitarianism, based on our reactions to exotic cases, are much weaker than they seem.

Several commentaries go beyond the search for reflective equilibrium to suggest other foundations for moral beliefs. Evolution looms large here. **Hinde** suggests that an evolutionary approach can help to identify the sources of moral norms. He emphasizes norms of reciprocity, which he sees as pan-cultural. In his view, such norms help to explain some of the phenomena I describe, such as the preference for pointless punishment. Certainly, moral heuristics may have evolutionary foundations. Note, however, that in the cognitive domain, heuristics (such as availability) are of interest whether or not we can explain them in evolutionary terms. And in the domain of morality, many heuristics have social rather than evolutionary origins. To Hinde, then, my response is that it is certainly valuable to explore the evolutionary sources of moral norms, but that we can learn a great deal without taking a stand on that question.

Hauser wants to understand moral reasoning by analogy to the language faculty. He thinks that there is some "circuitry" for deciding what is permitted and forbidden; he believes that we can make a great deal of progress in understanding moral thinking by studying that circuitry. Intriguingly, he suggests that diverse people agree on many judgments involving permissible harm even if they do not have access to the underlying principles (see Kahneman et al. 1998 for a similar finding). Hauser speculates that some aspects of our moral judgments may well be universal. The speculation is both interesting and plausible, but I think that Hauser should be more careful about the analogy to language. In the most interesting cases, people dispute what morality requires. These disputes are crucial to ethics, politics, and law. What I am trying to show is that heuristics often lead us in the wrong directions in those domains.

Casebeer's ambitious commentary asks, among other things, for a neurobiological foundation for heuristics and for an appreciation of neo-Aristotelian virtue theory. I am agnostic on both. Some heuristics may have an identifiable neurobiological source, but others probably do not. The availability heuristic is a heuristic regardless of what neurobiology says. Neo-Aristotelian virtue theory has many defenders, but its claims are controversial, and neo-Aristotelians disagree among themselves. My hope is that neo-Aristotelians will find at least some of my examples plausible.

Gorman is also concerned with the sources of moral intuitions. He thinks that it is important for people to use their moral imagination so as to identify a range of possible solutions. In his view, we should attempt to track the process of moral reasoning as it occurs, in part to learn about how people will react to emerging technologies, including cloning and the extension of human capabilities. What I would add to Gorman's sensible account is that for issues at the frontiers of scientific understanding, people are especially likely to use unreliable heuristics. Often they rely on a version of the view that we should not "tamper with nature." As **Tetlock** suggests, an effort to achieve reflective equilibrium, through encounters with specific practices as well as general principles, can provide an important corrective here.

Gerrig makes the intriguing suggestion that within narratives, different identifications, and in a sense different heuristics, take hold. If we identify with the hero of a movie, we're likely to want him to succeed, even if he is somewhat evil, and even at the expense of other characters who are both worthy and honest. Gerrig is right and his point seems to me quite important. Suppose that your best friend is in the midst of an unpleasant divorce. His wife is seeking to claim well over half of the couple's assets. His income is sig-

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

nificantly lower than hers, and he believes that he is entitled to at least half. You might well end up taking his side. Because his moral claim is embedded in a compelling narrative, and because he is, in a sense, the hero of that narrative, you might even feel moral outrage on his behalf.

But your moral judgment might well be reversed if the narrative context, and the emotional identification, is different. Suppose that it is the wife, in our little narrative, who is your best friend. Suppose that she has worked extremely hard to earn a decent living. Suppose that she is claiming a share of the couple's assets that is over 50% but that is still significantly below her economic contribution. Suppose finally, that in her view, her ex-husband-to-be is quite lazy, and his relatively weaker economic position is a direct product of his laziness. If so, you might feel moral outrage on her behalf rather than his.

There is a larger point here. Each one of us sees ourselves as a principal player in the narrative of our own lives, and each one of us identifies closely with that particular player. Moral judgments will inevitably be skewed as a result. I am not exactly sure what all this has to do with heuristics. But if **Gerrig** is right to say that we follow a heuristic to the effect that *The hero should succeed*, and if it is easy to identify the hero of your own life (you!), then moral blunders are inevitable.

In a way that is closely linked to **Gerrig**'s emphasis on narratives, **Fernandez-Berrocal & Extremera** are concerned with the role of emotions in producing moral decisions. They offer two interesting findings. First, those who understand their own emotions are more likely to think that a wife and mother (Meryl Streep, as it happens) should stick with her husband (instead of running away with Clint Eastwood). Second, those who can regulate their feelings are less likely to choose a definite outcome in the Asian disease Problem; such people are more likely to gamble. These findings raise at least two more general puzzles: When emotions are intensified, how, exactly, are people's moral judgments affected? And when emotions are intensified, do people's moral judgments get better or worse? The first question is an empirical one on which it would be good to know much more. The second question is only partly empirical (should Meryl Streep have run away with Clint Eastwood?), but we could undoubtedly make much progress on it as well. My discussion of moral heuristics does not specify or explore the role of emotion, nor does it explore the contested distinction between reason and emotion in the moral domain. There is much more to do on these questions.

This, in fact, seems to me to be the largest lesson of the diverse commentaries presented here. The idea of moral heuristics can be understood in many different ways, and there is a great deal to learn about the sources and nature of moral intuitions – the processes that give rise to them, their occasionally automatic character, and their substance in diverse contexts. Let's go to work.

ACKNOWLEDGMENTS

I am grateful to Elizabeth Emens and Adrian Vermeule for helpful comments on this response.

References

Letters "a" and "r" appearing before authors' initials refer to target article and response respectively.

Ackerman, F. & Heinzerling, L. (2004) *Priceless: On knowing the price of everything and the value of nothing*. The New Press. [aCRS]

Adams, H. (1876) *Essays on Anglo-Saxon law*. Little, Brown. [RAH]

Adler, M. D. & Posner, E. A. (1999) Rethinking cost-benefit analysis. *Yale Law Journal* 109:165. [MDA]

Ames, D. R., Flynn, F. J. & Weber, E. U. (2004) It's the thought that counts: On perceiving how helpers decide to lend a hand. *Personality and Social Psychology Bulletin* 30:461–74. [EUW]

Anderson, C. J. (2003) The psychology of doing nothing: Forms of decision avoidance result from reason and emotion. *Psychological Bulletin* 129:139–67. [CJA]

Arnhart, L. (1998) *Darwinian natural right: The biological ethics of human nature*. State University of New York Press. [WDC]

Bankowski, Z., White, I. & Hahn, U., eds. (1995) *Informatics and the foundations of legal reasoning*. Kluwer. [UH]

Baron, J. (1985) *Rationality and intelligence*. Cambridge University Press. [JB]

(1993a) Heuristics and biases in equity judgments: A utilitarian approach. In: *Psychological perspectives on justice*, ed. B. Mellers & J. Baron. Cambridge University Press. [UH, aCRS]

(1993b) *Morality and rational choice*. Kluwer. [EUW]

(1994a) Nonconsequentialist decisions. (Target article with commentary and response). *Behavioral and Brain Sciences* 17:1–42. [JB, UH, aCRS]

(1994b) *Thinking and deciding*, 1st edition. Cambridge University Press. [EUW]

(1998) *Judgment misguided: Intuition and error in public decision making*. Oxford University Press. http://www.sas.upenn.edu/~baron/vbook.htm [JB, aCRS]

(2000a) Can we use human judgments to determine the discount rate? *Risk Analysis* 20:861–68. [aCRS]

(2000b) *Thinking and deciding*, 3rd edition. Cambridge University Press. [JB]

(2004) Normative models of judgment and decision making. In: *Blackwell handbook of judgment and decision making*, ed. D. J. Koehler & N. Harvey, pp. 19–36. Blackwell. [JB]

Baron, J. & Frisch, D. (1994) Ambiguous probabilities and the paradoxes of expected utility. In: *Subjective probability*, ed. G. Wright & P. Ayton. Wiley. [JB]

Baron, J. & Hershey, J. C. (1988) Outcome bias in decision evaluation. *Journal of Personality and Social Psychology* 54:569–79. [DAP]

Baron, J. & Leshner, S. (2000) How serious are expressions of protected values. *Journal of Organizational Behavior and Human Decision Processes* 89:1100–18. [rCRS]

Baron, J. & Ritov, I. (1993) Intuitions about penalties and compensation in the context of tort law. *Journal of Risk and Uncertainty* 7:17–33. [aCRS]

(2004) Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes* 94:74–85. [CJA]

Baron, J. & Spranca, M. (1997) Protected values. *Organizational Behavior and Human Decision Processes* 70:1–16. [IR]

Baron, J., Gowda, R. & Kunreuther, H. (1993) Attitudes toward managing hazardous waste. *Risk Analysis* 13(2):183–92. [aCRS]

Benarzi, S. & Thaler, R. H. (2000) Myopic loss aversion and the equity premium puzzle. In: *Choices, values and frames*, ed. D. Kahneman & A. Tversky. Cambridge University Press. [aCRS]

Boyd, R. & Richerson, P. J. (1991) Culture and cooperation. In: *Cooperation and prosocial behaviour*, ed. R. A. Hinde & J. Groebel, pp. 27–48. Cambridge University Press. [RAH]

(1992) Punishment allows cooperation (or anything else) in sizeable groups. *Ethology and Sociobiology* 13:171–95. [RAH]

Broome, J. (1991) *Weighing goods: Equality, uncertainty and time*. Blackwell. [JB]

Burghardt, G. M. & Herzog, H. A., Jr. (1980) Beyond conspecifics: Is Brer Rabbit our brother? *Bioscience* 30:763–68. [HAH]

(1989) Animals, evolution, and ethics. In: *Perceptions of animals in American culture*, ed. R. J. Hoage, pp. 129–51. Smithsonian Institution. [HAH]

Camerer, C. (2000) Prospect theory in the wild: Evidence from the field. In: *Choices, values and frames*, ed. D. Kahneman & A. Tversky. Cambridge University Press. [aCRS]

Carlsmith, K. M., Darley, J. M. & Robinson, P. H. (2002) Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology* 83:284–99. [DAP]

Carroll, N. (1990) *The philosophy of horror*. Routledge. [RJG]

Casebeer, W. D. (2003a) Moral cognition and its neural constituents. *Nature Reviews Neuroscience* 4:841–47. [WDC]

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

(2003b) *Natural ethical facts: Evolution, connectionism, and moral cognition*. MIT Press. [WDC]
Casebeer, W. D. & Churchland, P. S. (2003) The neural mechanisms of moral cognition: A multiple-aspect approach to moral judgment and decision-making. *Biology and Philosophy* 18:169–94. [WDC]
Chomsky, N. (1957) *Syntactic structures.* Mouton. [JM]
(1964) *Current issues in linguistic theory.* Mouton. [JM]
Churchland, P. (1996) *The engine of reason, the seat of the soul*. MIT Press. [aCRS]
Churchland, P. M. (1998) Towards a cognitive neurobiology of the moral virtues. *Topoi* 17:83–96. [WDC]
Cohen, L. J. (1981) Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences* 4:317–31. [ES]
Coleridge, S. T. (1817/1907) *Biographia literaria*. Oxford University Press. (Originally published 1817). [RJG]
Cropper, M. L., Aydede, S. K. & Portney, P. R. (1994) Preferences for life-saving programs: How the public discounts time and age. *Journal of Risk and Uncertainty* 8:243–65. [aCRS]
Damasio, A. (1994) *Descartes' error: Emotion, reason, and the human brain*. Putnam. [JH]
Daniels, N. (1979) Wide reflective equilibrium and theory acceptance in ethics. *Journal of Philosophy* 76:256–82. [ES]
(1996) *Justice and justification: Reflective equilibrium in theory and practice.* Cambridge University Press. [PS]
Darley, J., Carlsmith, K. & Robinson, P. (2000) Incapacitation and just deserts as motives for punishment. *Law and Human Behavior* 24:659–83. [aCRS]
Dawes, R. (1998) Behavioral decision making and judgment. In: *The handbook of social psychology*, ed. D. T. Gilbert, S. T. Fiske & G. Lindzey. Oxford University Press. [PF-B]
de Waal, F. (1996) *Good natured: The origins of right and wrong in humans and other animals*. Harvard University Press. [aCRS]
Denby, D. (1991) A guy, a girl, and their guns. *Premiere* 5:32–33. [RJG]
Dewey, J. (1915) The logic of judgments of practice. In: *Collected works of John Dewey. Middle works, vol. 8*, ed. J. A. Boydston. Southern Illinois University Press. [EA]
(1922) *Human nature and conduct*. Henry Holt. [EA]
Dezhbakhsh, H, Rubin, P. & Shephard J. (2004) Does capital punishment have a deterrent effect: New evidence from post-moratorium panel data. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=259538 [aCRS]
Dostoyevsky, F. (1879/1994) *The Karamazov Brothers* (Part 2, Book 5, Ch. 4), trans. I. Avsey. Oxford University Press. [PS]
Dresser, R. (1989) Developing standards in animal research review. *Journal of the American Veterinary Medical Association* 194:1184–91. [HAH]
Dreyfus, H. & Dreyfus, S. (1986) *Mind over machine: The power of human intuition and expertise in the era of the computer.* The Free Press. [KB]
(1991) Towards a phenomenology of moral expertise. *Human Studies* 14:229–50. [KB]
Dwyer, S. (1999) Moral competence. In: *Philosophy and linguistics*, ed. K. Murasugi & R. Stainton, pp. 169–90. Westview. [MDH, JM]
(2004) How good is the linguistic analogy. Available at: www.umbc.edu/philosophy/dwyer. [MDH]
Epstein, S. & Pacini, R. (1999) Some basic issues regarding dual-process theories from the perspective of cognitive-experiential self-theory. In: *Dual-process theories in social psychology*, ed. S. Chiaken & Y. Trope. Guilford. [IR]
Ericsson, K. A. & Simon, H. A. (1984) *Protocol analysis: Verbal reports as data*. MIT Press. [MEG]
Fernandez-Berrocal, P. & Extremera, N. (in preparation) Emotional intelligence and moral dilemmas. [PF-B]
Fernandez-Berrocal, P., Extremera, N. & Ramos, N. (2004) Validity and reliability of the Spanish modified version of the Trait Meta-Mood Scale. *Psychological Reports* 94:751–55. [PF-B]
Fernandez-Berrocal, P., Salovey, P., Vera, A., Extremera, N. & Ramos, N. (2005) Cultural influences on the relation between perceived emotional intelligence and depression. *International Review of Social Psychology* 18:91–107. [PF-B]
Flanagan, O. (1993) *Varieties of moral personality: Ethics and psychological realism*. Harvard University Press. [aCRS]
Frederick, S. (2003) Measuring intergenerational time preference: Are future lives valued less? *Journal of Risk and Uncertainty* 26(1):39–53. [aCRS]
Frisch, D. (1993) Reasons for framing effects. *Organizational Behavior and Human Decision Processes* 54:399–429. [JB]
Frost, J., Maio, G. R. & Hahn, U. (in preparation) Effects of value instantiation on behaviour. [UH]
Gerrig, R. J. (1993) *Experiencing narrative worlds*. Yale University Press. [RJG]
Gibbard, A. (1986) Risk and value. In: *Values at risk*, ed. D. MacLean, pp. 94–112. Rowman and Allenheld. [BHF]
Gigerenzer, G. (1996) Narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review* 103:592–96. [aCRS]
(2000) *Adaptive thinking: Rationality in the real world*. Oxford University Press. [WDC, aCRS, PW]
Gigerenzer, G., Todd, P. & the ABC Research Group (1999) *Simple heuristics that make us smart.* Oxford University Press. [aCRS]
Gilovich, T., Griffin, D. & Kahneman, D., eds. (2002) *Heuristics and biases: The psychology of intuitive judgment.* Cambridge University Press. [PF-B, aCRS]
Goldman, A. (1970) *A theory of human action.* Princeton University Press. [JM]
Goldstein, D. G. & Gigerenzer, G. (2002) Models of ecological rationality: The recognition heuristic. *Psychological Review* 109:75–90. [UH, aCRS]
Goldstein, W. M. & Weber, E. U. (1995) Content and discontent: Indications and implications of domain specificity in preferential decision-making. In: *Decision making from a cognitive perspective: The psychology of learning and motivation, vol. 32*, ed. J. R. Busemeyer, R. Hastie & D. L. Medin, pp. 83–136. Academic Press. [Reprinted in: W. M. Goldstein & R. M. Hogarth, eds. (1997) *Research on judgment and decision making,* pp. 566–617. Cambridge University Press.] [EUW]
Goody, J. (1997) *Representations and contradictions.* Blackwell. [RAH]
Gopnik, A. & Sobel, D. M. (2000) Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development* 71:1205–22. [JM]
Gorman, M. E. (1992) *Simulating science: Heuristics, mental models and technoscientific thinking*. Indiana University Press. [MEG]
Gorman, M. E. & Mehalik, M. M. (2002) Turning good into gold: A comparative study of two environmental invention networks. *Science, Technology and Human Values* 27(4):499–529. [MEG]
Gorman, M. E., Mehalik, M. M. & Werhane, P. H. (2000) *Ethical and environmental challenges to engineering*. Prentice-Hall. [MEG]
Gould, S. J. (1991) *Bully for brontosaurus: Reflections in natural history*. W. W. Norton. [aCRS]
Greene, J. & Baron, J. (2001) Intuitions about declining marginal utility. *Journal of Behavioral Decision Making* 14:243–55. [JB]
Greene, J. & Haidt, J. (2002) How (and where) does moral judgment work? *Trends in Cognitive Sciences* 6:517–23. [PF-B, WDC, UH, aCRS]
Greene, J., Somerville, R. B., Nystrom, L .E., Darley, J. M. & Cohen, J. D. (2001) An fMRI investigation of emotional engagement in moral judgment. *Science* 293:2105–08. [JH, JM, aCRS, PS]
Grossman, D. (1996) *On killing: The psychological cost of learning to kill in war and society*. Back Bay Books. [CJA]
Haidt, J. (2001) The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108(4):814–34. [CJA, JH, UH, DAP, aCRS]
Haidt, J. & Baron, J. (1996) Social roles and the moral judgement of acts and omissions. *European Journal of Social Psychology* 26:201–18. [aCRS]
Haidt, J., Bjorklund, F. & Murphy, S. (2004) Moral dumbfounding: When intuition finds no reason. Unpublished manuscript, University of Virginia. [arCRS]
Haidt, J. & Hersh, M. (2001) Sexual morality: The cultures and emotions of conservatives and liberals. *Journal of Applied Social Psychology* 31:191–221. [DAP]
Haidt, J., Koller, S. & Dias, M. (1993) Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology* 65:613–28. [DAP]
Haidt, J., Rosenberg, E. & Hom, H. (2003) Differentiating diversities: Moral diversity is not like other kinds. *Journal of Applied Social Psychology* 33:1–36. [JH, DAP]
Hammond, K. R. (1996) *Human judgment and social policy.* Oxford University Press. [EUW]
Hare, R. M. (1981) *Moral thinking: Its levels, method and point*. Oxford University Press/Clarendon Press. [JB, aCRS]
Harman, G. (1999) Moral philosophy and linguistics. In: *Proceedings of the 20th World Congress of Philosophy, vol. 1: Ethics*, ed. K. Brinkmann, pp. 107–15. Philosophy Documentation Center, Bowling Green, KY. [MDH]
Hauser, M. D. (in press) *Moral minds: The unconscious voice of right and wrong.* Harper Collins. [MDH]
Hauser, M. D., Cushman, F. & Young, L. (in press) Reviving Rawls' linguistic analogy: Operative principles and the causal-intentional aspects of moral actions. In: *Moral psychology and biology*, ed. W. Sinnott-Armstrong. Oxford University Press. [MDH]
Hauser, M. D., Cushman, F., Young, L., Jin, R. K.-X. & Mikhail, J. (under review) A dissociation between moral judgments and justifications. *Proceedings of the National Academy of Sciences.* [JM]
Herzog, H. A. (1993) Human morality and animal research: Confessions and quandaries. *The American Scholar* 62:337–49. [HAH]
Herzog H. A. & Galvin, S. (1997) Anthropomorphism, common sense, and animal awareness. In: *Anthropomorphism, anecdotes, and animals*, ed. R. W. Mitchell & N. S. Thompson, pp. 237–53. State University of New York Press. [HAH]
Hinde, R. A. (1997) *Relationships: A dialectical perspective.* Psychology Press. [RAH]
(2002) *Why good is good*. Routledge. [RAH]

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

Hooker, B. (2000) *Ideal code, real world: A rule-consequentialist theory of morality*. Oxford University Press. [aCRS]
Hooker, B. & Little, M. O. (2001) Moral particularism. Oxford University Press. [WDC]
Jackendoff, R. (2004) *Language, culture, consciousness: Essays on mental structure*. MIT Press. [MDH]
Johnson, M. (1993) *Moral imagination*. University of Chicago Press. [MEG]
Johnson-Laird, P. N. (1983) *Mental models*. Harvard University Press. [MEG]
Kagan, J. (2000) Human morality is distinctive. *Journal of Consciousness Studies* 7:46–48. [RAH]
Kagan, S. (1998) *Normative ethics*. Westview Press. [MDA]
Kahneman, D. (2002) Maps of bounded rationality: A perspective on intuitive judgment and choice. Nobel Prize Lecture. In: *Les Prix Nobel. The Nobel Prizes 2002*, ed. Tore Frängsmyr, Stockholm, 2003, Nobel Foundation. Available at: http://nobelprize.org/economics/laureates/2002/kahneman-lecture.html [UH]
Kahneman, D. & Frederick, S. (2002) Representativeness revisited: Attribute substitution in intuitive judgment. In: *Heuristics and biases: The psychology of intuitive judgment*, ed. T. Gilovich, D. Griffin & D. Kahneman. Cambridge University Press. [PF-B, UH, IR, aCRS, EUW]
Kahneman, D. & Tversky, A. (1984) Choices, values, and frames. *American Psychologist* 39:341–50. [aCRS]
(1996) On the reality of cognitive illusions: A reply to Gigerenzer's critique. *Psychological Review* 103:582–91. [UH, aCRS]
Kahneman, D., Knetsch, J. L. & Thaler, R. H. (1986) Fairness as a constraint on profit-seeking: Entitlements in the market. *American Economic Review* 76:728–41. [aCRS]
Kahneman, D., Schkade, D. & Sunstein, C. R. (1998) Shared outrage and erratic awards: The psychology of punitive damages. *Journal of Risk and Uncertainty* 16:49–86. [arCRS]
Kahneman, D., Slovic, P. & Tversky, A., eds. (1982) *Judgment under uncertainty: Heuristics and biases.* Cambridge University Press. [DAP]
Kamm, F. (1993) *Morality, mortality, vol 1: Death and whom to save from it*. Oxford University Press. [aCRS]
(1998) Moral intuitions, cognitive psychology, and the harming-versus-not-aiding distinction. *Ethics* 108:463–88. [aCRS]
Kant, I. (1948/1964) *Groundwork of the metaphysics of morals*. Harper & Row. [KB]
Kaplow, L. & Shavell, S. (2002) *Fairness versus welfare*. Harvard University Press. [JB, JM, aCRS]
Kass, L. (1998) The wisdom of repugnance. In: *The ethics of human cloning*, ed. L. Kass & J. Q. Wilson. American Enterprise Institute. [aCRS]
Katz, L. D. (2000) *Evolutionary origins of morality: Cross-disciplinary perspectives*. Academic. [aCRS]
Kelley, H. H. & Thibaut, J. W. (1978) *Interpersonal relations*. Wiley. [RAH]
Kelman, S. (1981) *What price incentives? Economists and the environment*. Auburn House. [aCRS]
Koehler, J. J. & Gershoff, A. D. (2003) Betrayal aversion: When agents of protection become agents of harm. *Organizational Behavior and Human Decision Processes* 90(2):244–61. [JJK, aCRS]
Kohlberg, L. (1969) Stage and sequence: The cognitive-developmental approach to socialization. In: *Handbook of socialization theory and research*, ed. D. A. Goslin, pp. 347–480. Rand McNally. [DAP]
Krueger, J. & Funder, D. (2004) Towards a balanced social psychology: Causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences* 27(3):313–27. [aCRS]
Kruglanski, A. W. (1989) The psychology of being "right": The problem of accuracy in social perception and cognition. *Psychological Bulletin* 106:395–409. [DAP]
Kruglanski, A. W., Fishbach, A., Erb, H.-P., Pierro, A. & Mannetti, L. (2004) The parametric unimodel as a theory of persuasion. In: *Contemporary perspectives on the psychology of attitudes*, ed. G. Haddock & G. R. Maio, pp. 399–422. Psychology Press. [UH]
Kuran, T. & Sunstein, C. R. (1999) Availability cascades and risk regulation. *Stanford Law Review* 51:683–768. [aCRS]
Ledoux, J. (1996) *The emotional brain: The mysterious underpinning of emotional life*. Touchstone. [aCRS]
Lerner, J. S. & Keltner, D. (2001) Fear, anger, and risk. *Journal of Personality and Social Psychology* 81:146–59. [PF-B]
Loewenstein, G., Small, D. & Strnad, J. (2005) Statistical, identifiable and iconic victims. In: *Behavioral public finance: Toward a new agenda*, ed. E. McCaffrey & J. Slemrod. Russell Sage. [BHF]
Macnamara, J. (1986) *A border dispute: The place of logic in psychology.* MIT Press. [ES]
Maio, G. R. & Haddock, G. (in press) Attitude change. In: *Social psychology: Handbook of basic principles, vol. 2*, ed. A. W. Kruglanski & E. T. Higgins. Guilford Press. [UH]
McCloskey, H. J. (1957) An examination of restricted utilitarianism. *Philosophical Review* 66:466–85. [PS]
McDowell, J. (1998) *Mind, value, and reality*. Harvard University Press. [KB]
McGraw, P. & Tetlock, P. E. (2005) Taboo trade-offs, relational framing, and the acceptability of exchanges. *Journal of Consumer Psychology* 15:2–15. [PET]
McGraw, P., Tetlock, P. E. & Kristel, O. (2003) The limits of fungibility: Relational schemata and the value of things. *Journal of Consumer Research* 30:219–29. [PET]
McHughen, A. (2000) *Pandora's picnic basket*. Oxford University Press. [aCRS]
McKenzie, C.R. (2004) Framing effects in inference tasks – and why they are normatively defensible. *Memory and Cognition* 32(6):874–85. [aCRS]
Mellers, B., Hertwig, R. & Kahneman, D. (2001) Do frequency representations eliminate conjunction effects? *Psychological Science* 12(4):269–75. [aCRS]
Messick, D. (1993) Equality as a decision heuristic. In: *Psychological perspectives on justice*, ed. B. Mellers & J. Baron. Cambridge University Press. [aCRS]
Mikhail, J. (2002) Law, science, and morality: A review of Richard Posner's *The Problematics of Moral and Legal Theory*. *Stanford Law Review* 54:1057–27. [JM]
(in press) *Rawls' linguistic analogy.* Cambridge University Press. [JM]
Mikhail, J., Sorrentino, C. & Spelke, E. (1998) Toward a universal moral grammar. In: *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society,* ed. M. A. Gernsbacher & S. J. Derry, p. 1250. Erlbaum. [JM]
Mikhail, J. M. (2000) Rawls' linguistic analogy: A study of the "generative grammar" model of moral theory described by John Rawls in *A theory of justice*. Unpublished Doctoral dissertation, Cornell University. [JM, MDH]
Mikhail, J. M., Sorrentino, C. & Spelke, E. (2002) Aspects of the theory of moral cognition: Investigating intuitive knowledge of the prohibition of intentional battery, the rescue principle, the first principle of practical reason, and the principle of double effect. Unpublished manuscript, Stanford University. [MDH]
Mill, J. S. (1861/1957). *Utilitarianism*. Macmillan. [KB]
(1861/1971) *Utilitarianism*. Bobbs-Merrill. [aCRS]
Miller, A. (2003) *An introduction to contemporary metaethics*. Polity Press. [MDA]
Morrison, E. R. (1998) Comment: Judicial review of discount rates used in regulatory cost-benefit analysis. *University of Chicago Law Review* 65(4):1333–70. [aCRS]
Narvaez, D. (1999) Using discourse processing methods to study moral thinking. *Educational Psychology Review* 11:377–93. [EUW]
Newell, A. & Simon, H. A. (1972) *Human problem solving*. Prentice-Hall. [UH]
Nussbaum, M. (1984) Plato on commensurability and desire. *Proceedings of the Aristotelian Society* (Suppl.) 58:55–80. [aCRS]
(2002) *Upheavals of thought*. Cambridge University Press. [aCRS]
Parfit, D. (1984) *Reasons and persons*. Oxford University Press. [ES]
Patt, A. & Zeckhauser, R. J. (2000) Action bias and environmental decisions. *Journal of Risk and Uncertainty* 21(1):45–72. [CJA]
Pizarro, D. A. (2000) Nothing more than feelings? The role of emotions in moral judgment. *Journal for the Theory of Social Behaviour* 30:355–75. [DAP]
Pizarro, D. A. & Bloom, P. (2003) The intelligence of the moral intuitions: Comment on Haidt. *Psychological Review* 110(1):193–98. [CJA, aCRS]
Pizarro, D. A., Uhlmann, E. & Bloom, P. (2003) Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology* 39:653–60. [DAP]
Plous, S. & Herzog, H. (2000) Poll shows researchers favor lab animal protection. *Science* 290:711. [HAH]
(2001) Reliability of protocol reviews for animal research. *Science* 293:608–09. [HAH]
Polinsky, A. M. & Shavell, S. (1998) Punitive damages: An economic analysis. *Harvard Law Review* 111:869–76. [aCRS]
Prentice, D. A. & Gerrig, R. J. (1999) Exploring the boundary between fiction and reality. In: *Dual-process theories in social psychology,* ed. S. Chaiken & Y. Trope, pp. 529–46. Guilford Press. [RJG]
Profitt, D. R. & Kaiser, M. K. (2002) Intuitive physics. In: *Encyclopedia of cognitive science*, ed. L. Nadel. MacMillan. [UH]
Railton, P. (2000) Normative force and normative freedom: Hume and Kant, but not Hume versus Kant. In: *Normativity,* ed. J. Dancy, pp. 1–33. Blackwell. [KB]
Rashdall, H. (1907) *The theory of good and evil, vol. 1*. Clarendon Press. [PS]
Rawls, J. (1951) Outline of a decision procedure for ethics. *Philosophical Review* 60:177–97. [PS]
(1971) *A theory of justice.* Harvard University Press/Belknap Press. [MDH, JM, arCRS, PS]
(1974–1975) The independence of moral theory. *Proceedings and Addresses of the American Philosophical Association* 48:5–22. [ES]
Raz, J. (1994) The relevance of coherence. In: *Ethics in the public domain*, ed. J. Raz, pp. 277–326. Clarendon Press. [aCRS]
Rest, J. R. & Narvaez, D. (1994) *Moral development in the professions*. Erlbaum. [MEG]
Rest, J. R., Narvaez, D., Bebeau, M. J. & Thoma, S. J. (1999) *Postconventional moral thinking: A neo-Kohlbergian approach*. Erlbaum. [EUW]

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press

Revesz, R. (1999) Environmental regulation, cost-benefit analysis, and the discounting of human lives. *Columbia Law Review* 99(4):941–1017. [aCRS]
Rheingold, H. & Hay, D. (1980) Prosocial behavior of the very young. In: *Morality as a biological phenomenon*, ed. G. S. Stent, pp. 93–108. University of California Press. [RAH]
Ritov, I. & Baron, J. (1990) Reluctance to vaccinate: Omission bias and ambiguity. *Journal of Behavioral Decision Making* 3:263–77. [CJA]
(1999) Protected values and omission bias. *Organizational Behavior and Human Decision Processes* 97:79–94. [IR]
(2002) Reluctance to vaccinate: Omission bias and ambiguity. In: *Behavioral law and economics*, ed. C. R. Sunstein. Cambridge University Press. [aCRS]
Roco, M. C. & Bainbridge, W. S., eds. (2002) *Converging technologies for improving human performance: Nanotechnology, biotechnology, information technology and cognitive science.* NSF/DOC. [MEG]
Rozin, P. (2001) Technological stigma: Some perspectives from the study of contagion. In: *Risk, media, and stigma: Understanding public challenges to modern science and technology*, ed. J. Flynn, P. Slovic & H. Kunreuther. Earthscan Publications. [aCRS]
Salovey, P., Mayer, J. D., Goldman, S. L., Turvey, C. & Palfai, T. P. (1995) Emotional attention, clarity, and repair: Exploring emotional intelligence using the Trait Meta-Mood Scale. In: *Emotion, disclosure & health*, ed. J. W. Pennebaker, pp. 125–54. American Psychological Association. [PF-B]
Salovey, P., Stroud, L. R., Woolery, A. & Epel, E. S. (2002) Perceived emotional intelligence, stress reactivity, and symptom reports: Further explorations using the Trait Meta-Mood Scale. *Psychology and Health* 17:611–27. [PF-B]
Sandel, M. (1997) It's immoral to buy the right to pollute. *New York Times*, December 15, 1997, p. A23. [aCRS]
Schlosser, E. (2002) *Fast food nation: The dark side of the all-American meal*. HarperCollins. [aCRS]
Schwarz, N. & Vaughn, L. A. (2002) The availability heuristic revisited: Ease of recall and content of recall as distinct sources. In: *Heuristics and biases: The psychology of intuitive judgment*, ed. T. Gilovich, D. Griffin & D. Kahneman. Cambridge University Press. [PF-B]
Searle, J. (1983) *Intentionality.* Cambridge University Press. [DAP]
Sen, A. (1980–1981) Plural utility. *Proceedings of the Aristotelian Society* 81:193–215. [aCRS]
(1982) Rights and agency. *Philosophy and Public Affairs* 11:3–39. [aCRS]
(1985) Well-being, agency, and freedom: The Dewey lectures. *Journal of Philosophy* 82:169–221. [aCRS]
(1996) Fertility and coercion. *University of Chicago Law Review* 63:1035–61. [aCRS]
Serpell, J. A. (1989) Humans, animals, and the limits of friendship. In: *The dialectics of friendship*, ed. R. Porter & S. Tomaselli, pp. 111–29. Routledge. [HAH]
Shrager, J. (2004) Diary of an insane cell mechanic. In: *Scientific and technological thinking*, ed. M. E. Gorman, R. D. Tweney, D. C. Gooding & A. Kincannon. Erlbaum. [MEG]
Sigdwick, H. (1907/1981) *The methods of ethics.* Hackett. [aCRS]
Singer, P. (1974) Sidgwick and reflective equilibrium. *The Monist* 58:490–517. [PS]
Slovic, P., Finuncane, M., Peters, E. & MacGregor, D. G. (2002) The affect heuristic. In: *Heuristics and biases: The psychology of intuitive judgment*, ed. T. Gilovich, D. Griffin & D. Kahneman. Cambridge University Press. [aCRS]
Smart, J. J. C. (1973) An outline of a system of utilitarian ethics. In: *Utilitarianism: For and against*, ed. J. J. C. Smart & B. Williams. Cambridge University Press. [aCRS]
Smart, J. J. C. & Williams, B. (1973) *Utilitarianism: For and against.* Cambridge University Press. [aCRS]
Smith, A. (1759/1976) *The theory of the moral sentiments*. Clarendon. [MDH]
Smith, S. M. & Levin, I. P. (1996) Need for cognition and choice framing effects. *Journal of Behavioral Decision Making* 9:283–90. [PF-B]
Sober, E. & Wilson, R. (1999) *Unto others: The evolution and psychology of unselfish behavior*. Harvard University Press. [aCRS]
Sorenson, R. (1992) *Thought experiments*. Oxford University Press. [aCRS]
Spelke, E. S., Breinlinger, K. & Jacobson, K. (1992) Origins of knowledge. *Psychological Review* 99:605–32. [JM]
Stanovich, K. E. & West, R. F. (1998) Individual differences in framing and conjunction effects. *Thinking and Reasoning* 4:289–317. [PF-B]
(2000) Individual differences in reasoning: Implications for the rationality debate? *Behavior and Brain Sciences* 23:645. [EUW]
Stein, E. (1996) *Without good reason: The rationality debate in philosophy and cognitive science*. Oxford University Press. [aCRS, ES]
(2001) Ethics and evolution. *The MIT encyclopedia of the cognitive sciences*, ed. R. A. Wilson & F. C. Weil. MIT Press. [aCRS]
Stich, S. (1990) *The fragmentation of reason: Preface to a pragmatic theory of cognitive evaluation.* MIT Press. [ES]
Sunstein, C. R. (1999) *One case at a time: Judicial minimalism on the Supreme Court.* Harvard University Press. [aCRS]
(2002) *Risk and Reason: Safety, law, and the environment*. Cambridge University Press. [JB, aCRS]
(2003) *Why societies need dissent*. Harvard University Press. [aCRS]
(2004) Lives, life-years, and willingness to pay. *Columbia Law Review* 104:205–52. [aCRS]
Sunstein, C. R., Hastie, R., Payne, J. W. & Viscusi, W. K. (2002) *Punitive damages: How juries decide*. University of Chicago Press.
Sunstein, C. R. & Nussbaum, M. C., eds. (2004) *Animal rights: Current debates and new directions*. Oxford University Press. [HAH]
Sunstein, C. R., Schkade, D. & Kahneman, D. (2000) Do people want optimal deterrence? *Journal of Legal Studies* 29(1):237–54. [IR, aCRS]
Sunstein, C. R. & Thaler, R. H. (2003) Libertarian paternalism. *American Economic Review* 93:175–80. [JB]
Tetlock, P. (2000) Coping with tradeoffs: Psychological constraints and political implications. In: *Elements of reason: Cognition, choice, and the bounds of rationality*, ed. A. Lupia, M. D. McCubbins & S. Popkin. Cambridge University Press. [arCRS]
Tetlock, P. E. (1986) A value pluralism model of ideological reasoning. *Journal of Personality and Social Psychology: Personality Processes and Individual Differences* 50:819–27. [PET]
(2002) Social-functionalist metaphors for judgment and choice: The intuitive politician, theologian, and prosecutor. *Psychological Review* 109:451–72. [PET]
(2003) Thinking about the unthinkable: Coping with secular encroachments on sacred values. *Trends in Cognitive Science* 7:320–24. [PET]
Tetlock, P. E., Kristel, O., Elson, B., Green, M. & Lerner, J. (2000) The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology* 78:853–70. [PET]
Tetlock, P. E., Visser, P., Singh, R., Polifroni, M., Scott, A., Elson, B., Mazzocco, P. & Rescober, P. (under review) People as intuitive prosecutors: The impact of social-control goals on punitiveness and attributions of responsibility. [DAP]
Thaler, R. (1993) *Quasi-rational economics*. Russell Sage. [rCRS]
Thomson, J. J. (1986) The trolley problem. In: *Rights, restitution and risk: Essays in moral theory*, ed. J. J. Thomson & W. Parent. Harvard University Press. [aCRS]
Tversky, A. & Kahneman, D. (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185:1124–31. [RJG, UH, JJK, aCRS, EUW]
(1981) The framing of decisions and the psychology of choice. *Science* 211:453–58. [PF-B, JB]
(1984) Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review* 90:293–315. [IR, aCRS]
(1991) Loss aversion in riskless choice: A reference-dependent model. *Quarterly Journal of Economics* 106(4):1039–61. [aCRS]
Uhlmann, E., Pizarro, D. A. & Brescoll, V. (in preparation) The role of reason in moral judgment. [DAP]
Vacco vs. Quill (1997) *West's Supreme Court Reporter* 521:793–807. [JM]
Viscusi, W. K. (2000) Corporate risk analysis: A reckless act? *Stanford Law Review* 52:547–97. [aCRS]
Washington vs. Glucksberg (1997) *United States Reports* 521:702–89. [aCRS]
Weber, E. U., Ames, D. & Blais, A.-R. (2004) "How do I choose thee? Let me count the ways": A textual analysis of similarities and differences in modes of decision making in China and the United States. *Management and Organization Review* 1:87–118. [EUW]
Weber, E. U. & Hsee, C. K. (2000) Culture and individual judgment and decision making. *Applied Psychology: An International Review* 49:32–61. [EUW]
Weber, E. U., Böckenholt, O., Milton, D. J. & Wallace, B. (2000) Confidence judgments as expressions of experienced decision conflict. *Risk Decision and Policy* 5:1–32. [EUW]
Weirich, P. (2001) *Decision space: Multidimensional utility analysis*. Cambridge University Press. [PW]
Werhane, P. H. (1999) *Moral imagination and management decision making.* Oxford University Press. [MEG]
Wiggins, D. (1987/2002) *Needs, values, truth*. Blackwell/Clarendon Press. [KB, JH]
Williams, B. (1973) A critique of utilitarianism. In: *Utilitarianism: For and against*, ed. J. J. C. Smart & B. Williams. Cambridge University Press. [PS, aCRS]
Wilson, T. D., Centerbar, D. B. & Brekke, N. (2002) Mental contamination and the debiasing problem. In: *Heuristics and biases: The psychology of intuitive judgment*, ed. T. Gilovich, D. Griffin & D. Kahneman, pp. 185–200. Cambridge University Press. [CJA]
Yates, J. F. & Lee, J. W. (1996) Chinese decision making. In: *Handbook of Chinese Psychology*, ed. M. H. Bond. Oxford University Press. [EUW]
Zahavi, A. (2000) Altruism: The unrecognised selfish traits. *Journal of Consciousness Studies* 7:253–56. [RAH]

https://doi.org/10.1017/S0140525X0529009X Published online by Cambridge University Press