To be published in *Behavioral and Brain Sciences* (in press)

Moral Heuristics

Cass R. Sunstein
University of Chicago, Law School and Department of Political Science
1111 E. 60th Street
Chicago, Illinois 60637
Email: csunstei@uchicago.edu
Home page: 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 peculiarly unlikely to be unreliable. 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.

Keywords: acts and omissions, biases, cognition, heuristics, morality, punishment.

I. 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 and Kahneman 1974, p. 1124). Intense controversy has developed over the virtues and vices of the heuristics, most of them "fast and frugal," that play a role in many areas (see Gilovich et al. 2002; Gigerenzer et al. 1999). But the relevant literature has only started to investigate the possibility that in the moral and political domain, people also rely on simple rules of thumb that often work well but that sometimes misfire (see Baron 1994; Baron 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 essay 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 1994), 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 want to urge 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 where the error is both highly intuitive and on reflection uncontroversial – where 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 at least potentially erroneous. In the search for reflective equilibrium, understood as coherence among our judgments at all levels of generality (Rawls 1971; Daniels 1993), it is important to see that some of our deeply held moral beliefs might be a product 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; 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.

II. Ordinary Heuristics and An Insistent Homunculus

A. 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 and 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 people 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 and 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 observes 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 1999 et al.; Gigerenzer 2000). On 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 and Funder, forthcoming; compare Kahneman and 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; compare Kahneman and 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 and Gigerenzer 1999), 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.

B. Attribute Substitution and Prototypical Cases

What is a heuristic? Kahneman and Shane 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 (see Myers, 2003), and these intuitions sometimes are biased, in the sense that they produce errors in a predictable direction. Consider the question 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 perception (see Kuran and Sunstein 1998). For the size of cities, the recognition heuristic, which is 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 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, 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 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 below). 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; compare 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, Bjorkland, and Murphy 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.

III. 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 on <u>any</u> view about what morality requires; and second, by showing that such heuristics are at work on a minimally contentious view about what morality requires. I will identify a number of such heuristics in Part IV.

A. 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, 1971, pp. 28-29; Sidgwick, 1981, pp. 199-216; originally published 1907; 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, 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

[1] On 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 and Shavell 2003).

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 that 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 people 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 Anderson 1993; Nussbaum 1984; Sen 1980-81).

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. On 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, where the facts or simple logic provide a good test 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 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.

B. Neutral Benchmarks and Weak Consequentialism

My goal here is therefore not to show, with Sidgwick and Mill, that common sense morality is a series of heuristics for the correct general theory, 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 Part V, 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; Sen 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, at 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.

C. 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; Sober and Wilson 1999; Katz 2000). 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 to 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 (see below), and evolutionary pressures need not have any role at all.

Second, and relatedly, 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 complex 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 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.

IV. The Asian Disease Problem and Moral Framing

In a finding closely related to their work on heuristics, Kahneman and Tversky themselves find "moral framing" in the context of what has become known as "the Asian disease problem" (Kahneman & Tversky1984). 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 instead "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 and 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 (Revesz 1999; Morrison 1998). 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 (1994) suggest that people are indifferent between saving one life today and saving 45 lives in 100 years. 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

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

people's moral judgments about obligations to future generations are very much a product of framing effects (for a similar result, see Baron 2000).[3]

The same point holds for the question 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 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.

V. Moral Heuristics: A Catalogue

My principal interest here is the relationship between moral heuristics and questions of law and policy. I separate the relevant heuristics into four categories: those that involve morality and risk regulation; those that involve punishment; those that involve "playing God," particularly in the domains of reproduction and sex; and those that involve the act-omission distinction. The catalogue is meant to be illustrative rather than exhaustive.

A. Morality and Risk Regulation

1. Cost-Benefit Analysis. An automobile company is deciding whether to take certain safety precautions for its cars. In deciding whether 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 lives 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

[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).

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 Viscusi 2000; Tetlock 2000). The oddity here is that under tort law, it is unclear that a company should not 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 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 that they will cause a death, and do so anyway, are regarded as cold-hearted monsters.[4] On 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 and 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 1 in 1 million risk of death. Its product is used by ten million people. The cost of eliminating the risk would have been $100 million.

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

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 the two.

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 perspective. 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 instead a product of System I, and a crude but quite tenacious moral heuristic.

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

[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).

> [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 so 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 so 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 violate the law or are unconnected with desirable activities, 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.

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 1 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 pure moral judgments; they are choosing what they like best, and it just turns out that a moral judgment, involving antipathy to betrayals, 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 and trust in nature and antipathy to "playing God,' as discussed below.) 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.

B. Morality and Punishment

1. Pointless punishment? In the context of punishment, moral intuitions are sometimes disconnected with the consequences of punishment, suggesting that a moral heuristic may well be at work (see Darley et al. 2000; Darley and Carlsmith forthcoming). 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 1993, 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

1993). 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.

2. Probability of Detection. Now turn to 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 (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 large question whether people accept this account, and if not, why not. (For the moment, let us put to one side the question 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, Schkade, and Kahneman 2000). In the first, subjects were given cases of wrongdoing, arguably calling for punitive damages, and also 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 to reduce penalties when the probability of detection was high, and to increase penalties when the probability of detection was low. Subjects were questioned 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 be consider the possibility that System I has gotten the better of them.

C. 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 (pp. 17-19, 1998) 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 whether human cloning is ethically defensible. But I want to suggest that moral heuristics, and System I, are responsible for Kass seeks to celebrate as "we intuit and feel, immediately and without argument." In fact Kass' catalogue of alleged errors seems to me 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 (for example, "mass production of human beings, with large clones of look-alikes, compromised in their individuality"). The representativeness heuristic plays a similar role (for example, "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 2002, pp. 31, 38).

Studies show that people overestimate the carcinogenic risk from pesticides and underestimate the risks of natural carcinogens (Rozin 2002). People also believe that nature implies safety, so much that they will prefer natural water to processed water even if the two are chemically identical (Rozin 2002).

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 deliberative, 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 2001). 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 harms 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 return to the incest taboo. We can easily imagine incestuous relationships, say between first cousins or second cousins, that ought not to give rise to social opprobrium, but that 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.

D. 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 v. 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 <u>always</u> 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 above; many people show an omission bias, favoring inaction over statistically preferable action (Baron and 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 1994). 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 here, 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 and 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 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 that 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 reflect 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 the authors' 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 1998), in a way that suggests a possible physical location for some of the operations of System I. So too, perhaps, in the context of morality, politics, and law (Greene & Haidt 2002).[6]

[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 Dezhbakhsh, Rubin, and Shephard, 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 the 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

VI. 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 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 out reasons, exotic cases might be positively harmful (compare Flanagan 1993).

---

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?

In short, I believe that some philosophical and philosophical analysis, based on exotic moral dilemmas, is inadvertently and even comically replicating the early work of Kahneman and Tversky: uncovering situations in which intuitions, normally quite sensible, turn out to misfire. The irony is that while 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 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 (compare 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 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 – better than the alternatives even if productive of error in imaginable cases. I have noted that moral heuristics might show a kind of "ecological rationality," working well in most real-world contexts (Gigerenzer 2000); recall 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 discussions of wide and narrow reflective equilibrium in Daniels 1993; 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 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.

References

Ackerman, F. and Heinzerling, L. (2004) *Priceless: On Knowing the Price of Everything and the Value of Nothing*. The New Press.

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

(1994) Nonconsequentialist decisions. *Behavioral and Brain Sciences* 17:1-10.

(1998) *Judgment misguided: Intuition and error in public decision making.* Oxford University Press.

(2000) Can we use human judgments to determine the discount rate? *Risk Analysis* 20:861-68.

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

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

Benarzi, S. and Thaler, R.H. (2000). Myopic Loss Aversion and the Equity Premium Puzzle. In *Choices, Values and Frames,* Kahneman and Tversky eds. Cambridge University Press.

Camerer, C. (2000). Prospect Theory in the Wild: Evidence from the Field. In *Choices, Values and Frames*, Kahneman and Tversky eds. Cambridge University Press.

Churchland, P. (1995). *The Engine of Reason, the Seat of the Soul*. The MIT Press.

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 & Uncertainty* 8:243-65.

Darley, J., Carlsmith, K., and Robinson, P. 2000. Incapacitation and Just Deserts as Motives for Punishment. Law and Human Behavior **24**: 659-683.

Darley, J. and Carlsmith, K. Incapacitation and Just Deserts as Motives for Punishment *Law and Human Behavior*, forthcoming.

De Waal, F. (1996). *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Harvard University Press.

Dezhbakhsh, H, Rubin, P., and 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

Flanagan, O. (1993) *Varieties of Moral Personality: Ethics and Psychological Realism.* Harvard University Press.

Frederick, S. (2003) Measuring intergenerational time preference: Are future lives valued less? *Journal of Risk and Uncertainty* 26(1):39-53.

Gigerenzer, G. (1996) Narrow Norms and Vague Heuristics: A Reply to Kahneman and Tversky. Psychological Review, 103, 592-596.

Gigerenzer, G. (2000) *Adaptive Thinking: Rationality in the Real World.* Oxford University Press.

Gigerenzer, G. & Todd, P. (1999) *Simple heuristics that make us smart.* Oxford University Press.

Gilovich, T., Griffin, D., & Kahneman, D., eds. (2002) *Heuristics and biases: The psychology of intuitive judgment.* Cambridge University Press.

Gould, S.J. (1991) *Bully for brontosaurus: Reflections in natural history*. W.W. Norton & Company.

Greene, J. & Haidt, J. (2002) How (and where) does moral judgment work? *Trends in Cognitive Sciences* 6:517-23.

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.

Haidt, J. (2001) The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108(4): 814-34.

Haidt, J. & Baron, J. (1996). Social roles and the moral judgement of acts and omissions. *European Journal of Social Psychology*, 26, 201-218.

Haidt, J., Bjorklund, F., & Murphy, S. (2004). Moral dumbfounding: When intuition finds no reason. (Unpublished manuscript, University of Virginia).

Hare, R.M. (1981). *Moral Thinking*. Oxford University Press.

Hooker, B. (2000). *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*. Oxford University Press.

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.

Kahneman, D., Schkade, D., & Sunstein, C.R. (1998) Shared outrage and erratic awards: The psychology of punitive damages. *Journal of Risk & Uncertainty* 16:49-86.

Kahneman, D. & Tversky, A. (1984). Choices, values, and frames. *American Psychologist* 39: 341-50.

Kahneman, D., Slovic, P., & Tversky, A., eds. (1982) *Judgment under uncertainty: Heuristics and biases.* Cambridge University Press.

Kahneman, D. & Tversky, A. (2000) Choices, values and frames. In: *Choices, values, and frames* ed. D. Kahneman & A. Tversky. Cambridge University Press.

Kahneman, D., & Tversky, A. (1996) On the reality of cognitive illusions: A reply to Gigerenzer's critique. *Psychological Review*, 103: 582-91.

Kahneman, D., Knetsch, J.L., and Thaler, R.H. 1986. Fairness As A Constraint on Profit-Seeking: Entitlements in the Market. *American Economic Review* 76: 728-41.

Kamm, F. (1998). Moral Intuitions, Cognitive Psychology, and the Harming-versus-Not-Aiding Distinction. *Ethics* 108: 463-488.

Kamm, F. (1993). *Morality, Mortality, vol 1: Death and Whom to Save from It*. Oxford University Press.

Kaplow, L. and Shavell, S. (2003) *Fairness Versus Welfare*. Harvard University Press.

Kass, L. (1998) The wisdom of repugnance. In: *The ethics of human cloning*, L. Kass & J.Q. Wilson. American Enterprise Institute.

Katz, L.D. (2000). *Evolutionary Origins of Morality: Cross-Disciplinary Perspectives*. Academic.

Kelman, S. (1981) *What Price Incentives? Economists and the Environment*. Auburn House.

Kuran, T. and Sunstein, C.R. (1999). Availability Cascades and Risk Regulation. *Stanford Law Review*. 683– 768

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.

Krueger, J. & Funder, D. Toward A Balanced Social Psychology: Causes, Consequences and Cures for the Problem-Solving Approach to Social Behavior and Cognition. *Behavioral and Brain Sciences* (forthcoming).

Ledoux, J. (1996) *The emotional brain: the mysterious underpinning of emotional life*. Touchstone.

McHughen, A. (2000) *Pandora's Picnic Basket*. Oxford University Press.

McKenzie, C. (2004) Framing Effects in Inference Tasks – And Why They're Normatively Defensible. *Memory and Cognition*, in press

Mellers, B., Hertwig, R., & Kahneman, D. (2001) Do frequency representations eliminate conjunction effects? *Psychological Science* 12(4):269-75.

Messick, D. (1993) Equality as a decision heuristic. In: *Psychological perspectives on justice*, ed. B. Mellers & J. Baron. Cambridge University Press.

Mill, J.S. (1971). *Utilitarianism*. Bobbs-Merrill Company.

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.

Nussbaum, M. (1984) Plato on Commensurability and Desire. *Proceedings of the Aristotelian Society*, suppl. Vol. 58: 55-80.

Nussbaum, M. (2002) *Upheavals of Thought*. Cambidge University Press.

Pizarro, D.A. & Bloom, P. (2003) The intelligence of the moral intuitions: Comment on Haidt. *Psychological Review* 110(1):193-98.

Polinsky, A.M. & Shavell, S. (1998) Punitive damages: An economic analysis. *Harvard Law Review* 111:869–76.

Rawls, J. (1971) *A theory of justice.* Harvard University Press.

Raz, J. (1994) The Relevance of Coherence. In J. Raz, *Ethics in the Public Domain*, 277-326. Oxford: Clarendon

Revesz, R. (1999) Environmental regulation, cost-benefit analysis, and the discounting of human lives. *Columbia Law Review* 99(4): 941-1017.

Ritov, I. & Baron, J. (2002) Reluctance to vaccinate: Omission bias and ambiguity. In: *Behavioral Law and Economics*, ed. C.R. Sunstein. Cambridge University Press.

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.

Rozin, P. & Nemeroff, C. (2002) Sympathetic magical thinking: The contagion and similarity "heuristics." In: *Heuristics and biases: The psychology of intuitive judgment*, ed. T. Gilovich, D. Griffin, & D. Kahneman. Cambridge University Press.

Sandel, M. (1997) *It's Immoral to Buy the Right to Pollute*, N.Y.TIMES, Dec. 15, 1997, at A23

Schlosser, E. (2002) *Fast Food Nation: The Dark Side of the All-American Meal*. HarperCollins.

Sen, A. (1980-81) Plural Utility. *Proceedings of the Aristotelian Society*, 81, 193-215

Sen, A. (1982) Rights and Agency. *Philosophy and Public Affairs* 11: 3-39

Sen, A. (1984) Well-Being, Agency, and Freedom: The Dewey Lectures. *Journal of Pholosophy* v. 82, 169-221.

Sen, A. (1996) Fertility and Coercion. *University of Chicago Law Review*, 63, 1035-1061.

Sigdwick, H. (1981) (originally published 1907). The Methods of Ethics. Hackett Publishing Company.

Slovic, P., Finuncane, M., Peters, E., and MacGregor, D.G. The Affect Heuristic. In Gilovich, T., Griffin, D., & Kahneman, D., eds. (2002) *Heuristics and biases: The psychology of intuitive judgment.* Cambridge University Press.

Smart, J.J.C. An Outline of A System of Utilitarian Ethics. In Smart, J.J.C. & Williams, B. (1973) *Utilitarianism: For and against.* Cambridge University Press.

Smart, J.J.C. & Williams, B. (1973) *Utilitarianism: For and against.* Cambridge University Press.

Sober, E. & Wilson, R. (1999) *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press.

Sorenson, R. (1992). *Thought Experiments*. Oxford University Press.

Stein, E. (2001) Ethics and Evolution. *The MIT Encyclopedia of the Cognitive Sciences*. Edited by Robert A. Wilson and Frank C. Weil. MIT.

Stein, E. (1996) *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science*. Oxford University Press.

Sunstein, C.R. (2004) Lives, life-years, and willingness to pay. *Columbia Law Review*. 104: 205-52.

Sunstein, C.R. (1999) *One case at a time: Judicial minimalism on the Supreme Court.* Harvard University Press.

Sunstein, C.R., Hastie, R., Payne, J.W., & Viscusi, W.K. (2002) *Punitive damages: How juries decide*. University of Chicago Press.

Sunstein, C.R., Schkade, D., & Kahneman, D. (2000) Do people want optimal deterrence? *Journal of Legal Studies* 29(1):237-54.

Sunstein, C.R. (2003) *Why Societies Need Dissent*. Harvard University Press.

Tetlock, P. (2000) Coping with tradeoffs. In: *Elements of reason: Cognition, choice, and the bounds of rationality*, ed. A. Lupia, S. Popkin, M.D. McCubbins. Cambridge University Press.

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.

Tversky, A. and Kahneman, D. (1991). Loss Aversion in Riskless Choice: A Reference-Dependent Model. Quarterly Journal of Economics 106:4, 1039-61.

Tversky, A. and Kahneman, D. (1984). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review*, 90, 293-315.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.

Viscusi, W.K. (2000) Corporate risk analysis: A reckless act? *Stanford Law Review* 52:547-97.

Washington v. Glucksberg (1997) *West's Supreme Court Reporter* 521:702-89.

Williams, B. (1973) A Critique of Utilitarianism, in *Utilitarianism: For and Against*. J.J.C. Smart and B. Williams. Cambridge University Press.

Acknowledgements: 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, Jonathan Haidt, Robert Goodin, Steven Pinker, Edward Stein, Peter Singer, and a number of anonymous reviewers.