Chapter 3. The Case Against Intuition

> "I am afraid," replied Elinor, "that the pleasantness of an employment does not always evince its propriety."
>
> "On the contrary, nothing can be a stronger proof of it, Elinor; for if there had been any real impropriety in what I did, I should have been sensible of it at the time, for we always know when we are acting wrong, and with such a conviction I could have had no pleasure."
>
> Jane Austen, *Sense and Sensibility*

THE EVIDENCE OF SELF-EVIDENCE

Astronomers have stars; geologists have rocks. But what do moral theorists have to work with? For centuries, they typically claimed to proceed from truths that were self-evident—or, more modestly, from our moral intuitions. Moral theories would explain, and be constrained by, these intuitions—that's the way with theories—and so moral theorists had to have a story about what kind of intuitions counted. Thomas Reid, founder of the so-called Common Sense school (a sort of ordinary-language philosophy *avant la lettre)*, urged that we "take for granted, as first principles, things wherein we find an universal agreement, among the learned and the unlearned, in the different nations and ages of the world. A consent of ages and nations of the learned and vulgar, ought, at least, to have great authority, unless we can show some prejudice as universal as that consent is, which might be the cause of it. Truth is one, but error is infinite." At the same time, he held that "attentive reflection" was itself "a kind of intuition."[1] It had great authority, too, and would help chasten intuitions of the first kind.

His successors were often more fastidious when it came to listening to the unlettered. William Whewell, in his 1846 *Lectures on Systematic Morality*, insisted that he was proceeding from self-evident principles: but then grappled with the question of to whom they were self-evident. We could have no coherence in our views, he argued, if we submitted them to "a *promiscuous Jury* from the mass of mankind." Rather, self-evidence would be tested by "a Jury of men as men," thoughtful and wholesome specimens: "They are men as well as the first twelve you might take at Temple Bar, or at Timbuctoo," by which he meant that they are "true men; or at least men who have laboured and toiled, under favourable circumstances, to be true to their humanity." Whewell believed men such as these "pronounce on my side;—that they decide such fundamental Principles as mine to be true and universal Principles of Morality." In his view, "*Vox populi feri* is not *vox veritatis*. It would be more suitable to say *Vox humani generis vox veritatis*."[2]

Nor did Henry Sidgwick, half a century later, stray far when he called for philosophers to study, "with reverent and patient care," what he dubbed "the Morality of Common Sense," by which he did *not* mean the moral judgments of the common run of man (it would be "wasted labor" to systematize the morality of the worldly, debased as it was by "their sordid interests and vulgar ambitions"). The Morality of Common Sense referred, rather, "to the moral judgments—and especially the spontaneous unreflected judgments on particular cases, which are sometimes called moral intuitions—of those persons, to be found in all walks and stations of life, whose earnest and predominant aim is to do their duty." The moral philosopher was to be "aided and controlled by

---

NOTES TO CHAPTER 3

[1] Reid, *Essays on the Intellectual Powers of Man*, *op. cit.*, Vol. 1, 363, 361.

[2] William Whewell, *Lectures on Systematic Morality* (London: J.W. Parker, 1846), 34, 35, 121. (Translating the Latin, we get: "The voice of savage people is not the voice of truth. It would be more suitable to say that the voice of human kind is the voice of truth.")

them in his theoretical construction of the Science of Right."[3]

The basic picture here—moral theory as the perfection of elevated intuition—is anything but foreign to modern moral philosophy. In the past century, though, the endeavor has taken on an increasingly scientific tinge. Sir David Ross, in *The Right and the God* (1930), proposed that "the moral convictions of thoughtful and well-educated people are the data of ethics just as sense-perceptions are the data of a natural science." Not that such data must always be taken at face value: "Just as some of the latter have to be rejected as illusory, so have some of the former," he allowed, but "only when they are in conflict with convictions which stand better the test of reflection."[4] In the nineteen-fifties, John Rawls formalized the method of his predecessors in what became (after some further elaboration) his celebrated notion of "reflective equilibrium"—directing us to adjust our principles to our intuitions, and our intuitions to our principles, until, through mutual calibration, consonance was achieved—and in doing so, he too made explicit reference to such scientific theorizing.[5] And the model of moral philosophy as a sort of intuition refinery has maintained its allure. The philosopher Frank Jackson, writing a century after Sidgwick, takes the task of "moral functionalism" to be to construct "a coherent theory out of folk morality, respecting as much as possible those parts that we find most appealing, to form mature folk morality"—the latter being what folk morality ends up when exposed to "debate and critical reflection." Folk morality may be a complicated and untidy thing, but, he maintains (and Reid, Whewell, Sidgwick, Ross, and Rawls would have concurred) "we must start from somewhere in current folk morality, otherwise we start from somewhere *unintuitive*, and that can hardly be a good place to start from."[6]

Not surprisingly, our moral theories have clashing ambitions: if their plausibility comes from their ability to accommodate our intuitions, their power comes from their ability to challenge still other intuitions. A theory may be applauded for being revisionary, for showing us the error of our quondam judgments—or rejected precisely because it defies common sense. A theory may be commended for the ease with which it accommodates our existing intuitions—or castigated for its flaccid indulgence of our wayward prejudices. In one direction, we complain of normative systems that seem impossibly unmoored from human judgment, bicycles built for octopods. "To ask us to give up at the bidding of a theory our actual apprehension of what is right and what is wrong," Sir

---

[3] Henry Sidgwick, "My Station and Its Duties," *International Journal of Ethics* 4, 1 (October 1893), 9, 10.

[4] W.D. Ross, *The Right and the Good* (Oxford: Oxford University Press, 1930), 40.

[5] "The procedure is somewhat analogous to evidencing a proposition or theory in the real sciences, except that in oral discussions we try to validate or invalidate decisions and the action consequent thereto, given the circumstances and the interest in conflict (not acts of believing given a proposition or theory and its evidence) and the criteria we use are the principles of justice (and not the rules of inductive logic)," Rawls wrote in his "Outline of a Decision Procedure for Ethics," *The Philosophical Review* 60, 2 (April 1951): 195-6. Intuitions are the "data," in the consonant words of Nicholas Rescher, that "the theoretician must weave into a smooth fabric" in a process that is "closely analogous with the systematization of the 'data' of various levels in natural science." "Reply to Hare," in Ernest Sosa (ed.), *The Philosophy of Nicholas Rescher: Discussion and Replies* (Dordrecht: D. Reidel Publishing Co., 1979), 153–155. Nelson Goodman, writing in 1954, advanced the procedure as a general one for theory building. A given deductive argument, he wrote, is justified by being seen to accord with the rules of deductive inference. "Yet, of course, the rules themselves must eventually be justified ... Principles of deductive inference are justified by their conformity with accepted deductive practice. Their validity depends upon accordance with the particular deductive inferences we actually make and sanction. If a rule yields unacceptable inferences, we drop it as invalid.... The point is that rules and particular inferences alike are justified by being brought into agreement with each other. A rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend." Nelson Goodman*, Fact, Fiction, and Forecast, 4th ed.* (Cambridge: Harvard University Press, 1983), 63-4.

[6] Frank Jackson, *From Metaphysics to Ethics* (Oxford: Clarendon Press, 1998), 134, 133, 135.

David Ross wrote, "seems like asking people to repudiate their actual experience of beauty, at the bidding of a theory which says, 'only that which satisfies such and such conditions can be beautiful'"—a request that, he says, "is nothing less than absurd."[7] In the other direction, though, we get the bugbear of moral conservatism, propping up the disreputable old theories that our intuitions enrobe. It is the charge James Mill (quoting Alexander Pope) made against Edmund Burke: that in his philosophy "whatever is, is right." Such accusations, on either side of the ledger, often have more than a whiff of the ad hoc. On what basis do we settle them? Here we confront what I'll call the "intuition problem."

A quick pair of examples. For many, it is a point in Jeremy Bentham's favor that, in contemplating the principle of utility, he got the big issues right that his contemporaries got wrong: he was able to challenge the prevailing moral intuitions of his day about slavery, the subjection of women, homosexuality, and so forth. We can laud the triumph of sterling principle over the debased currency of moral common sense. Other moral thinkers, skeptical of the radical impartialism that utilitarianism seems to demand, have, by way of rebuttal, held up to scorn a notorious passage by William Godwin, Bentham's rough contemporary and fellow utilitarian. It's a passage in which Godwin imagines having to choose whom to rescue from a fire: either Archbishop Fénelon, who had much to contribute to the general welfare, or someone without the archbishop's moral distinction—perhaps Godwin's brother, his benefactor, his father. According to Godwin, justice requires that we rescue the archbishop, and let father perish in the flames.[8] Ethicists, in his day as in ours, have shaken their heads in disbelief: who would want to sign onto a doctrine that had this deeply unappealing result? Here, as so often, it is common sense, not principle, that is taken as dispositive.[9] What guides us here? In the realm of meta-ethics, it would seem, we have clashing intuitions about intuitions.

Reflective equilibrium remains the usual way that philosophers think about the vexed status of intuition in normative ethics. What has made this notion so durable is not that it has solved the difficulties it means to address but that the difficulties themselves have proved so durable—one of philosophy's many *Jarndyce v. Jarndyce*-style impasses. Indeed, one could be forgiven for thinking that reflective equilibrium is really another name for the problem, rather than a solution to it. It's worth recalling that in its earliest, 1951 formulation, Rawls's "decision procedure" was for moral judgments in particular cases. This modest, mid-century notion may have had certain advantages, but in the seventies and after, the notion was revised and expanded into a strategy not just for arriving at judgments in particular cases but for moral theory more generally; the list of things to be brought into equilibrium was broadened to include background theories and beliefs. As you might fear, the

---

[7] W. D. Ross, *The Right and the Good* (Oxford: Oxford University Press, 1930), 40.

[8] I've done so recently myself in *The Ethics of Identity* (Princeton: Princeton University Press, 2005). Ironically, a remark attributed to Fénelon suggests an impartialist perspective that Godwin might have commended. Desmoulins cites him in a talk to the Club des Jacobins as saying, "I love my family more than myself, my fatherland more than my family, and the universe more than my fatherland" ("J'aime mieux ma famille que moi, ma patrie que ma famille, et l'univers que ma patrie.") (Quoted by Camille Desmoulins "Sur la situation politique de la nation à l'ouverture de la seconde session de l'Assemblée nationale" Club des Jacobins, 21 octobre 1791 cited at http://www.royet.org/nea1789-1794/archives/discours/desmoulins_situation_politique_nation_21_10_91.htm)

[9] At the same time, these moral radicals enlist intuition, too. If one of two people must die, and one is an insignificant wretch and the other a great humanitarian, wouldn't most people feel relieved to learn the great humanitarian survived? Or consider a revisionist line of argument that was first advanced by Bentham and has reached its greatest force and influence in the works of Peter Singer. If you agree that it's wrong to mistreat an infant, whose sentience surely compares unfavorably to that of a mature horse or dog, doesn't it stand to reason that the capacity to suffer, not the capacity to think, is what entitles a creature to considerate treatment? Such arguments seek to overturn one element of common sense by appealing to another.

procedure for reaching equilibrium is less than determinate.

Suppose we have a theory, T_1, from which we can derive all our moral intuitions (so far) except intuition INT. (This is the simplest case. In general, of course, there'll be a class of intuitions that don't fit with our current best ethical theory.) Suppose we construct a different theory, T_2 that differs from our existing theory only as much as is necessary to accommodate INT. We cannot now reject INT because it isn't derivable from a theory: it can be derived from T_2. To reject INT on the basis that it can't be derived from T_1, we should need to have a reason for preferring T_1 to T_2 in the first place. There might be pragmatic reasons: T_1 might be simpler. (And, given that ethics is practical, simplicity might be a genuine theoretical virtue here.) But theoretical simplicity is both hard to measure and also, often, a function of what terms you're willing to take as basic, so it may be hard to apply this criterion. A more external kind of practical virtue in a theory might be that others in our own society endorse it (which would further incline us toward conservatism about ethics). At any rate, in the absence of a reason for preferring one theory to another, independently of the intuitions it supports, the contest is between INT and the other intuitions that we must abandon if we accept T_2. The point is that the standard reflective-equilibrium approach is only going to help us deal with conflicting intuitions if we have some independent ideas about the shape of ethical theory as well. It looks as though, as with our non-normative beliefs construed in a realist way, saving the phenomena—our appraisals; our perceptual judgments—is not enough.[10]

Suppose T_1 were some version of utilitarianism. Rawls discussed the objection that it might sometimes be welfare-maximizing to punish someone known to be innocent, a practice he dubbed "telishment"; he supposed that our intuitive revulsion toward telishment—our strongly held conviction that it's wrong to punish the innocent—counted against utilitarianism. So we could abandon utilitarianism UT_1 for UT_2 (modified utilitarianism), which says that one should maximize utility except where it entails the judicial punishment of innocent people. The utilitarian may want us to bite the bullet and reject the intuition in the name of coherence. But why, given that we can take the path of modified utilitarianism instead? In seeking reflective equilibrium between theory and intuition here, how to chose between rejecting INT and rejecting utilitarianism? Down the road, of course, further intuitions may lead us to abandon this new theory, too. But before we go any further, it is hard to apply reflective equilibrium as a procedure without already knowing whether to regard UT_2 as inferior to utilitarianism. (Because, say, it looks unacceptably ad hoc, as many people would agree in this case; the more familiar repair is to subject UT_1 to the constraint that nobody's rights are violated.) If what matters is fitting together all the bits—as Rawls's coherentist picture suggests—then we need a better theory than we've been offered for what a coherent theory should look like.[11]

---

[10] It won't help, of course, to say that we should aim at truth here, because, in each case, it looks like we only have access to the truth by way of the phenomena. This is the thought that leads to anti-realism. See my *For Truth in Semantics* (Oxford: Blackwell, 1986).

[11] Here's one counterproposal that has been bruited: Give up the demand that our intuitions and our theory never clash. On this view—engagingly advanced by the philosopher Ben Eggleston—what matters is not that the theory always gives us the outcome that accords with our intuition; what matters is whether the theory can endorse our having the intuition in question. Act utilitarianism may recommend telishment while also endorsing our intuitive revulsion at it, since such an intuition might conduce to acts that maximize utility. See Ben Egglestone "Practical Equilibrium: A New Approach to Moral Theory Selection" http://www.ku.edu/~utile/unpub/pe.pdf. The fact that this approach is more conserving of our actual intuitions is put forward as a recommendation. Whether you take it as one will depend, of course, on your intuitions about the reliability of people's intuitions. I'll return to the point that one might distinguish—as not all the literature does—between intuitions about cases and intuitions about principles. We have intuitions when confronted with general propositions—cruelty is bad; punishing the innocent is wrong—but further inquiry would be necessary to determine whether our intuitions in response to stated principles conform with our intuitions in response to the actual phenomena. Yet there's another ambiguity that bedevils the subject. We sometimes conflate "intuitions" as specific responses to specific cases with "intuition" as a general disposition to respond

Other philosophers, meanwhile, complain about what they regard as the method's conservatism. Indeed, while Rawls's own "theory of justice" is usually criticized by pointing to its counter-intuitive implications, his method of reflective equilibrium is often denounced as overly deferential to our intuitions. By subjecting our intuitions to an internal test of coherence, critics say, the method may simply give our old prejudices a haircut and a shave. Where would we be if philosophy had been compelled to respect what used to pass as common sense—that morality required religion, say, or that slavery was part of the natural order?[12] They notice that Rawls's own work is stippled with appeals to shared intuitions—to what "there is a broad measure of agreement" about, what "it seems reasonable to suppose," what "we are confident" about—and they ask: What do you mean "we," Kemo Sabe?

There's certainly reason to wonder whether "obvious" is a relation masquerading as a predicate (so that we ought to ask always, "Obvious to whom?") A great Cambridge mathematician, so the story goes, once stood before his class and filled the chalkboard with a vast and intricate equation. Underlining the result, he declared, "As you can see, it's obvious." Suddenly, though, he was seized by doubt, and, with furrowed brow, crept from the classroom. When he returned five minutes later, he was in fine spirits, his worries banished. "Why, yes, indeed," he assured his students, "it *is* obvious."[13]

---

to certain generalizable features across a range of cases. The first looks like the sort of raw data we can collect; the second—in which we distill a pattern from these specific responses—necessarily requires various background assumptions, inferences, and theories. In the account of "practical equilibrium" that Eggleston advocates, the intuitions to be endorsed cannot be the particular judgment in a particular case, but some larger disposition that's taken to issue in that judgment; and identifying that larger disposition is a far from trivial task, as we'll see. This problem should, of course, remind us of the discussion of what it is to identify the maxim of an action in Kant.

[12] See, for example, Richard Brandt, *A Theory of the Good and the Right* (Oxford: Clarendon Press, 1979), 21-2; or, more recently David Papineau, "The Tyranny of Common Sense," *The Philosophers' Magazine* http://www.philosophersnet.com/magazine/article.php?id=1005. "In ethics, as in mathematics, the appeal to intuition is an epistemology of desperation," Philip Kitcher writes in "Biology and Ethics," in David Copp, ed., *The Oxford Handbook of Ethical Theory* (New York: Oxford University Press, 2005) 176. R. M. Hare, in "Rawls' Theory of Justice—I," *Philosophical Quarterly* 23, 91 (April 1973), 146, started a list of Rawls's appeals to common sense and accused Rawls of "monistic intuitionism." Because these arguments, in various ways, reprise John Stuart Mill's critique of Whewell's moral theory—and because Mill's critique is better known than its object—it should be noted that Whewell's rationalist intuitionism is closer to the constructivism that Rawls espoused than to the moral-sense varieties of intuitionism in which moral truths are directly apprehended; for Whewell, elevated intuition was a starting point, it did not define the right answer (our conscience was as fallible as our reason); and he believed in progress in the moral realm as much as in the scientific realm. "Conscience," he readily acknowledged, is "a subordinate and fallible Rule," for man's "moral and intellectual progress is still incomplete; and this incompleteness is no justification for what is done under its influence." Whewell, *The Elements of Morality* (New York: Harper and Brothers, 1845), 246. And cf. J. S. Mill, "Whewell on Moral Philosophy" (1852) in J. M. Robson ed. *The Collected Works of John Stuart Mill,* Vol. X (Toronto: University of Toronto Press, 1969), 167-201. It will be obvious, I hope, that the intuition problem is hardly confined to those approaches to morality that are labeled "intuitionist."

[13] This story, which I learned in college, is presumably about G. H. Hardy, who wrote about a proposition in his *A Course of Pure Mathematics*, "This is almost obvious" … but then added a footnote in which he said:

> There is a certain ambiguity in this phrase which the reader will do well to notice. When one says "such and such a theorem is almost obvious" one may mean one or other of two things. One may mean "it is difficult to doubt the truth of the theorem," "the theorem is such as

The intuition problem, I should say, isn't peculiar to moral philosophy, but it's hardest to avoid there, as you find when you consider the alternative credos offered by the anti-intuition camp. David Papineau, for instance, proposes that, rather than giving quarter to common sense, "All claims should be assessed on their merits, against the tribunals of observation and reason": but when the claims are moral ones, we can't be optimistic that our observations and reasons could be cleansed of our intuitions. Richard Brandt, for his part, exhorts us to "step outside of our own tradition somehow, see it from outside": and that little word "somehow" betrays an understandable worry about the ease with which we can do so.[14]

The suspicion that our common sense may be littered with perishable and parochial prejudice is, of course, an ancient and enduring one, stretching from Herodotus to the pioneers of modern cultural anthropology. Utilitarian reformers who sought to combat entrenched moral intuitions found encouragement in Helvetius's *De l'Esprit*, and its argument that self-interest, including self-interest in the form of class interest, was the hidden wellspring behind talk of virtue and duty (presaging what Engels called "false consciousness"). Over the past decade or two, there has been a renewed assault on the status of moral intuitions, and from another direction. Even as the scientific paradigm has urged moral theorists to treat intuitions as data, a wave of empirically based research into human decision-making has depicted many of those intuitions to be—in ways that appear to be universal, and perhaps incorrigible—unreliable and incoherent.

THE PROSPECTS FOR COMMON SENSE

Consider the following experiment, which will be familiar to many of you. We divide our subjects into two groups. The task, we tell them, is to choose between two policy options as we prepare for an impending outbreak of the Asian Flu. If we do nothing, 600 or so people will die. Group 1 gets to choose between policies A and B, which we describe as follows. Policy A will save the lives of 200 people who would otherwise have died. Policy B has a 1/3 chance of saving 600 people and a 2/3 chance of saving nobody. So Group 1 has these options:

A	B
Save 200 people	1/3 chance of saving 600 & 2/3 chance of saving nobody

---

common sense intuitively accepts," as it accepts, for example, the truth of the propositions "2 + 2 = 4" or "the base angles of isosceles triangles are equal". That a theorem is "obvious" in this sense does not prove that it is true, since the most confident of the intuitive judgments of common sense are often found to be mistaken; and even if the theorem is true, the fact that it is also "obvious" is no reason for not proving it, if a proof can be found. The object of mathematics is to prove that certain premises imply certain conclusions; and the fact that the conclusions may be as "obvious" as the premises never detracts from the necessity, and often not even from the interest of the proof.

But sometimes (as for the example here) we mean by "this is almost obvious" something quite different from this. We mean "a moment's reflection should not only convince the reader of the truth of what is stated, but should also suggest to him the general lines of a rigorous proof". And often, when a statement is "obvious" in this sense, one may well omit the proof, not because the proof is unnecessary, but because it is a waste of time to state in detail what the reader can easily supply for himself.

(Cited in Andrew Lenard, "What can be learned from n<n!?" *Mathematics Magazine*, Feb 1998, http://www.findarticles.com/p/articles/mi_qa3789/is_199802/ai_n8788117.)

[14] Brandt, *A Theory of the Good and the Right, loc. cit*. Some philosophers have argued that it is *impossible* to distance oneself from one's own moral perspective. I should be clear that this is a claim (putting aside the sense in which it is trivially true, taking whatever perspective one adopts to be one's perspective) that I do not endorse. See my discussion of irony in Richard Rorty's *Contingency, Irony and Solidarity* by Richard Rorty in "Metaphys Ed." *The Village Voice* (September 19 1989), 55.

Group 2 also gets a couple of options, policy C and policy D, described as follows:

C	D
400 people die	1/3 chance nobody dies & 2/3 chance that 600 will die

You will notice at once that, granted the background assumption that 600 people will die if we do nothing, A and C are the same policy. It takes only a moment longer to see that, since, in this context, "600 people are saved" and "nobody dies of flu" are the same and so are "saving nobody" and "600 people die of flu," B and D are the same policy, too. These, then, are the same options, differently described.

What happened, though, when the choice was actually put to people? The answer, as the psychologists Daniel Kahneman and Amos Tversky showed in a justly famous experiment, was rather surprising.[15] Those who had to choose between A and B generally favored A; people who had to choose between C and D were inclined to favor D. In fact, roughly three-quarters of each group had these opposed preferences.

This is one of a multitude of examples of what are called "framing effects." People's choices often depend on exactly how the options are framed, even when the descriptions are, rationally speaking, equivalent. Now Kahneman and Tversky, being respectable scientists, didn't just do an experiment, they offered an explanation of why the experiment comes out this way. They claimed that they could say what it was about the framing that made the difference. Indeed, as is normal, they did the experiment to test the explanation, which they called *prospect theory*. It holds that:

1) People are risk-averse when thinking about gains over what they think of as the background *status quo* but
2) they're willing to take risks when they're faced with possible losses with respect to that *status quo*.
3) They're also more concerned to avoid losses than they are attracted by equivalent gains.

Prospect theory explains these results because when we say we are *saving* 200 people, we take the death of 600 people as the baseline. So we think of the 200 people saved as a gain, even though 400 people are still dying. Since we're risk-averse about gains, we prefer A to B, because B risks that gain of 200. On the other hand, if we think of 400 people *dying*, we're taking the baseline to be nobody dying. Now the 400 people dying are a loss. And since we're willing to take risks to avoid losses, we'll prefer D to C; because now we see C as offering the possibility that no one will die. If you get people to focus on the worst case—that 600 people still die, despite our best efforts—they'll pick the guarantee that we save at least 200. If you get them to focus on the best case—nobody's dying—they'll aim for it, even though it's the less likely outcome.

Now I don't know what you would have expected people to say here or what you would have said yourself. But, as we saw, a large slice of moral philosophy has consisted in thinking about such cases, forming an intuition about the right answer, and then trying to discover principles that will explain why it's the right answer. (People can, of course, have intuitions about principles as well as about cases, but, except where I say so explicitly, I'm going to use the word "intuition"—as Sidgwick suggested—to refer to responses to cases, or, more precisely, classes of cases.)

One challenge to our methods in moral philosophy posed here is simply this. It looks as though our intuitive judgments about what it's right to do are determined in part by things that cannot possibly be relevant. Given that the policy choices are exactly the same whichever way you describe them, people are responding here to something that cannot possibly matter. What's nice

[15] Amos Tversky and Daniel Kahneman, "The Framing of Decisions and the Psychology of Choice," *Science,* 221 (1981), 453-458. In the original experiment 72% preferred A to B and 78% preferred D to C. *loc. cit.*, 453.

about the Asian Flu experiment, in other words, is that it makes its point whichever policy you think is right; it doesn't require moral premises to get its conclusion. When our intuitions are guided by irrelevant factors, they *can't* be reliable guides.

As it happens, many philosophers, like traditional economists, think we should pick options on the basis of their expected costs and benefits, and so would rank A, B, C and D equally.[16] (You calculate the expected value of an option by multiplying the value of each of the mutually exclusive possible outcomes by its probability and adding those products all up.) For A and C, the expected loss of life is (1 x 200) people, since to say that a policy will save 200 people, is to say, in effect, that it will save 200 people with probability 1; for B and D the expected loss of life is (1/3 x 600) + (2/3 x 0) which is 200 again. So from that point of view preferring any of these options to any other is *irrational* because the policy of assessing options by their expected values is the policy with the greatest probability of maximizing your benefits over the long haul. Here the judgment about what's reasonable is based, then, not on an intuition but on an argument.[17] But an argument of that sort, too, directs us to revisit the status we accord to intuitions about scenarios like this one.

And, once you start to look, you can start to find the influence of baseline considerations everywhere. Take a little experiment that Thomas Schelling once conducted with his students, to try to draw out their intuitions of justice. He asked them what they thought about a particular tax policy that gave a bigger child deduction—that paid a larger bonus—to rich people who had kids than it

---

[16] Of course, even economists and philosophers know that some irrational people will be lucky and some rational people will be unlucky. Lotteries are designed to make money for the people that run them. As a result, usually when someone wins the lottery, they've made out on an investment that had a negative financial expected value; and, conversely, the person who takes the drug with a small risk of negative side effects and a large prospect of medical benefits and then becomes one of the rare people who suffer the side effects was nevertheless doing the rational thing. Given this fact, people sometimes respond to the maximizing argument by saying, "How can it be the best policy if rational people can suffer so badly and irrational people can profit so much?" The question presupposes that the best policy ought to guarantee rewards for those who stick to it. As I say, I'll generally be sticking to intuitions about cases; but you could call that presupposition an intuition, too, if you like. And it's a mistaken intuition, because it ignores the reality that we live in a world of chance, where the best you can do is to try to maximize the probability of getting what you want. There are, generally speaking, no guarantees. But this mistaken thought is not the sort of intuition that I have in mind, because it's a thought, as I say, about a principle. And the intuitions I have in mind are intuitions about cases, scenarios like the Asian Flu scenario: they are what happens when we read or hear a scenario described and just find we have a view about what is the right thing to do.

[17] Of course, a judgment in a social-policy scenario like this one is, in an important sense, normative; and you can imagine a morally fraught argument between those who favor different choices. Imagine an Option D partisan arguing with an Option C partisan: "What kind of monster are you that you would countenance the certain death of four hundred people!" Imagine an Option A partisan arguing with an Option B partisan: "What kind of monster are you that you could risk the deaths of all six hundred?" In fact, there might be relevant differences between the probabilistic outcomes and the determinate ones, such that I should prefer one to the other, depending on other background assumptions—especially those to do with further consequences that have been cropped out by the scenario. Perhaps my community of six hundred would be so devastated by grief and despondency were four hundred to die that it would be utility-maximizing to pursue the 1/3 chance of no deaths. Perhaps the survival of the germ line is all important, so that people would prefer the guaranteed survival of a few to the possible extinction of all. And of course the actual policies that would achieve these results would have to be different, even if the outcomes were not. The Kahneman and Tversky scenario brackets a question that preoccupies deontologists, namely, how these results are arrived at. Suppose the choice were between (a) murdering four hundred people in order to extract enough gamma globulin from their serum to save two hundred and (b) administering everyone a vaccine that would save only a third of them. (In each case, let the alternative be that all 600 people die.)

gave to their poor counterparts. As you'd guess, students didn't think that policy was fair. Then he redescribed the situation. Suppose the tax code has, as a default, a couple with children, and imposes a surcharge, a penalty, for the childless. Should that surcharge be larger for the poor than for the rich? The students all reversed their position, even though the two scenarios specified the very same distribution.[18]

As with the Asian Flu case, our intuitions hang on what we counted as a loss and what as a gain. There are others ways of manipulating our moral choices that don't involve anything like the gain/loss distinction, and seem even less respectable. In a recent experiment conducted by Thalia Wheatley and Jonathan Haidt, at the University of Virginia, people were taught under hypnosis to feel disgust—a brief sense of "a sickening in your stomach"—when they came across the emotionally neutral words "take" or "often." Then they were presented with different scenarios. One was about a person acting in morally troubling ways (a hypocritical bribe-taking congressman); another was about someone whose actions were quite morally untroubling (a student council representative who was in charge of scheduling discussions about academic issues, and who tried to choose topics that would appeal to both professors and students). The scenarios came in two versions that were almost identical, except that one version contained the cue word. The researchers found that when their subjects were responding to the versions that had the cue word, they judged moral infractions much more harshly: the congressman who will "take" bribes from the tobacco lobby was judged a greater villain than the congressman who is "bribed by" the tobacco lobby. As for Dan, the student council rep, nobody disapproved of him when he sought to stimulate discussion by trying to pick topics of interest both to students and to professors. What's not to like? But when he "often" picked such topics, a significant number of people couldn't help but disapprove. Asked to explain, they'd write down such comments as: "It just seems like he's up to something." One subject fingered him for being a "popularity-seeking snob"; another wrote, "It just seems so weird and disgusting." Even more mordant was the comment: "I don't know why it's wrong, it just is."[19]

It isn't comforting to learn that people's moral judgments can be shaped by a little hypnotic priming, given that the everyday world fills us with all sort of irrelevant associations that may play a similar role. And though the situationist studies I discussed in the previous chapter chiefly involved conduct—and, in the main, what look like acts of supererogation—rather than expressions of moral judgments or justifications, they, too, surely bolster the worries about intuitions we've touched on.

Kahneman and Tversky themselves suggested that prospect theory provides a better explanation of most people's intuitions about a wide range of cases than do traditional moral philosopher's distinctions. Of course, philosophers who think that a choice is right according to some principle do not have to hold that the principle explains our actual responses, not least because it is open to them to say that many of our intuitive responses are wrong. The philosopher is bound to regard the pattern of responses in the Asian Flu case, for example, as irrational; so she will not provide a theory that justifies this pattern of responses at all. Indeed, where many people regularly make the wrong choice, a psychological explanation of why they do so supports the philosopher's claim that they are mistaken, because it relieves us of the worry that they are being guided by some rational principle that we have failed to discover.

But this sort of research will, at least at first, provide less comfort than dismay. Since we moral philosophers invoke intuitions about cases all the time, the effect of thinking about Schelling's tax-code case or the Asian Flu case is a bit like the effect that thinking about visual illusions and hallucinations has on our confidence in our visual judgments. Descartes, you'll recall, pointed out that we often make mistakes about what's really there, as when we're dreaming. And his question was: If you know you make mistakes sometimes, why aren't you worried that you might be making them all the time? Similarly, those studies in the psychology of decision-making leave us nervously aware that, at least sometimes, our intuitions are deeply unreliable. What's even more nervous-

---

[18] Thomas C. Schelling, "Economic Reasoning and the Ethics of Policy," *Public Interest* (63), 37-61.

[19] Thalia Wheatley and Jonathan Haidt, "Hypnotic Disgust Makes Moral Judgments More Severe," *Psychological Science* 16, 10 (October 2005), 780-784.

making is that psychologists have offered what we might call *undermining* explanations for judgments that many moral philosophers have taken to be *correct*, at least in part because they comport with those philosophers' intuitions.

TROLLEYOLOGY

To approach the problem, consider another line of recent experimental work that starts by observing a troubling conflict between the moral intuitions on which philosophers rely; a conflict that some psychologists suggest should lead us to question the intuitions some important moral theories are intended to justify and support. The argument starts from a long-established discussion in philosophy about a range of so-called "trolley problems," which derive originally from scenarios dreamed up by Philippa Foot and elaborated by Judith Jarvis Thompson, among (eventually) many others.[20] Consider a pair of them. In one scenario, there's a runaway trolley hurtling down the tracks, and it's on course to kill five people who are, unavoidably, in its path. You can save those five people, but only by hitting a switch that will put the trolley onto a side track, where it will kill one person. Should you do it? Philosophers generally say yes. And research conducted in recent years shows that the intuition is widely shared. When polled, most people—eighty or ninety percent of us—will say yes.

In a second scenario, the trolley, once again, is hurtling toward the five people. This time you're on a footbridge over the tracks, next to a 300-pound man. The only way you can save those five people is by pushing the 300-pound man off the bridge and onto the tracks. His body mass will stop the runaway trolley, but he'll be killed in the process. Should you do it? Most of us, theorists and civilians alike, say no.[21]

Those responses are, in themselves, neither surprising nor troubling. Though the body count is the same whether you hit the switch or sacrifice the 300-pound man, there are plausibly relevant differences between the cases. When Philippa Foot introduced the trolley scenario she was exploring a traditional moral idea—the doctrine of double effect—according to which there's a significant difference between doing something that has a bad outcome as a foreseen but unintended consequence and intentionally doing that bad thing, even if the overall outcome is the same.[22] This is

---

[20] Philippa Foot, "The Problem of Abortion and the Doctrine of the Double Effect," *Oxford Review* 5, 8-9 (1967), reprinted in Philippa Foot, *Virtues and Vices and Other Essays in Moral Philosophy* (Berkeley: University of California Press, 1978), 19, 23. Judith Jarvis Thomson, "Killing, Letting Die, and the Trolley Problem" *The Monist* 59 (1976), 204-217, reprinted in Thomson, *Rights, Restitution, and Risk*: William Parent ed. (Cambridge: Harvard University Press, 1986); Judith Jarvis Thomson, "The Trolley Problem," Yale Law Journal 94, (1985), 1395-1415, reprinted in Thomson, *Rights, Restitution, and Risk*, *op. cit.,* 94. Thomson's earlier arguments concerning the trolley-car cases centered on claims to rights and "origin of harm"; in *The Realm of Rights* (Cambridge: Harvard University Press, 1990), she has turned to the notion of "hypothetical consent." In the bystander-at-the-switch case, a person, even if he knew that he would later be one of the people on the tracks, would rationally consent (here, the notion of "rationally" is normatively fraught); cf. William Godwin: "It would have been just in the valet to have preferred the archbishop to himself. To have done otherwise would have been a breach of justice."

[21] Researchers have collected data from around the world and the results are quite robust. Some deeply held convictions, though, turn out to be surprisingly culturally variable, including our repugnance toward "telishment." Kaiping Peng et al. asked students to respond to a "magistrate and the mob" scenario: if authorities don't falsely convict and punish an innocent man, murderous ethnic rioting will break out, resulting in many deaths and injuries. Chinese students were much more likely to consider telishment in this scenario to be justified than were American students. (The work, by Peng, Doris, Nichols, and Stich, is described in John M. Doris and Alexandra Plakias "How to Argue About Disagreement: Evaluative Diversity and Moral Realism" in Walter Sinnott-Armstrong ed., *The Biology and Psychology of Morality* (Oxford: Oxford University Press, forthcoming).)

[22] Discussion of this question usually assumes that we can decide what someone intended

relevant to traditional Catholic views about abortion, for example, according to which if a fetus dies in the course of saving a mother's life (because, say, you have to remove her cancerous uterus) the doctor is blameless for the death. Similarly, in wartime, civilian casualties, where unavoidable, are usually seen as a permissible side-effect of military activities, but the deliberate killing of civilians in order, say, to put pressure on an enemy government, is a war-crime—even if the net effect of that pressure is to force an end to the war and reduce overall fatalities.

The doctrine of double effect seems relevant here, at least at first glance, since in the footbridge example you are stopping the trolley *by* killing someone: the death of the large stranger isn't an unintended by-product of your stopping the trolley, it's the *means* by which you do it. In the original trolley case, by contrast, the man on the track dies as a foreseen but unintended consequence of your diverting the lethal machine.[23] That's a possible reason, a justification, for our differing intuitions in the two cases. Some philosophers, especially those of deontological inclinations, see an illustration here of the distinction between bringing-about and allowing-to-happen, or between aiding and not harming.[24] Still other philosophers … well, uncountable trolley-related arguments and

---

independently of a moral evaluation of their action. Not so, according to experiments I'll discuss later in this chapter. See Joshua Knobe, "The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology," *Philosophical Studies* (forthcoming). As found at http://www.unc.edu/~knobe/PhilStudies.pdf.

[23] These scenarios are, of course, extraordinarily artificial; but there are real situations where we have to contemplate killing the innocent in order to save lives. When the attacks on the World Trade Center and the Pentagon happened on September the 11th, fighter planes were scrambled, whose job, if they had been successful, would have been to kill the many innocent passengers on those planes, in order to save other people's lives. Now, of course, the people on the planes were going to be dead in either case. But suppose the President had been asked for permission to use fighter jets, or some sort of avionic intervention, to get the plane to crash not in Manhattan but in White Plains, where the death toll on the ground would be a good deal lower. This choice is structurally exactly the choice in the trolley problem. There are people you can save by stopping a dangerous vehicle in one place, thus killing fewer people than if you had let the vehicle travel on. The deaths in White Plains are foreseen but unintended side effects of saving Manhattan. Here most people would respond (and expect the President to respond) as in the trolley problem: authorizing someone to cause the death of some people in order to save the lives of many more.

[24] Nor do trolley ethicists restrict themselves to the rails. To exemplify a situation that produces the similar response as the footbridge scenario, philosophers often mention the fact that every transplant surgeon knows that he could extend the lives of at least five people by killing a random healthy adult, who is the potential source, after all, of two kidneys, a liver, a heart and a pair of lungs. There are people waiting for just such organs right now in many hospitals in the world. Without them, they'll soon die. Yet not only do most people think it's completely obvious that it would be wrong to take the life of a healthy person for this purpose, we don't even allow people to take these organs from *dead* people without their consent or that of their relatives. So all the time, faced with the option of killing someone to save the lives of several others, doctors abstain … exactly as most people say they would do in the footbridge scenario. We can call it the organ-harvesting problem. This generalizes a scenario—the "Transplant Case"— that was introduced by Philippa Foot, in "The Problem of Abortion and the Doctrine of Double Effect," *Oxford Review,* 5 (1967), 28-41 and elaborated by Judith Jarvis Thomson, "Killing, Letting Die, and the Trolley Problem," *The Monist* 59 (1976), 204-217, and that stipulated a choice between killing one particular healthy person and letting five patients die. Though the scenarios are importantly different, it's plausible to suppose that our response to the Transplant case is influenced by our response to the organ-harvesting nightmare, which is a low-level generalization of it. That's why it's perfectly reasonable to forbid the surgeon from killing his healthy visitor even if you favor toppling large men on footbridges to stop runaway trolleys. Who'd want to live in that fear-stricken society where any board-certified surgeon had a license to kill? Indeed, a policy of allowing any surgeon who needs two kidneys, a heart and a liver, to

elaborations have been published in the past quarter century; by now, the philosophical commentary on these cases makes the Talmud look like *Cliffs Notes,* and is surely massive enough to stop any runaway trolley in its tracks.

A solution to the problem of why it sometimes is and sometimes isn't permissible to end one life to save more than one would start by classifying the trolley problem and the footbridge dilemma as *different* in some important way. To justify the different responses, that important difference would have to be one that *warrants* the different responses; that makes sense of treating the cases differently. One challenge to this approach comes then from the psychologists who are able to classify the two cases differently and thus explain our differential response, but do so in a way that does not seem to warrant the differential treatment. That was what Kahneman and Tversky did for the Asian Flu case. Prospect theory shows why the different frames lead to different responses, but it wasn't supposed to show that different frames *justify* different responses; indeed, it helped undermine that claim.

As you will already have guessed, some philosophers have detected the shadow of prospect theory over those trolley tracks: Robert Nozick, for one, wondered whether "the relation of the bringing about/allowing to happen" didn't "seem suspiciously like that of the gain/loss distinction." After all, in the real world, we can never be sure what the consequences will be down the line. As a result, however the scenario is described, we're likely to treat any action as taking a risk to secure some outcome. Perhaps, then, when we're deciding what to do with the man on the footbridge, his being alive is the background status quo, the baseline, which we seek to preserve because we're risk-averse, even when the net gain is four lives. And perhaps in the original trolley case, by contrast, the five people alive right now are the baseline, so we risk diverting the trolley to avoid the loss; we'll take risks to avoid losses. Of course, if I can get you to see what will happen if you don't intervene as the baseline, then intervening to reduce the death toll is taking a risk to produce a gain. Since you're risk averse, you should be inclined to leave well alone. And some people do respond to the original trolley problem in that way.[25]

Consider, too, the possible role of order effects. The order in which moral options are presented to us, research has shown, can affect which option we choose. In the bystander-at-the-switch version of trolley problem, we are always told first about the five people whose lives are imperiled; we learn about the one person on the side track as a subsequent complication. Having resolved to do whatever you can to stop the train from its destructive course—and having learned about the switch, and the side track—you have already formed a strong resolve: you look to be confirmed in your decision (as confirmation bias might suggest) and view countervailing considerations with a grain of skepticism.

These thoughts are disruptive, of course, because the question is what it's right to do, and it's worrying that we can be shifted one way or another on this by redescribing the same situation.

A SCANNER DARKLY

An even more disruptive proposal, however, has been made by the experimental moral psychologist Joshua Greene and his colleagues, who conclude that the our different responses to the

---

grab and kill the first healthy person who comes along would mean that a lot of other people would be sick and dying because they had decided that it was risky to be around doctors.

[25]Robert Nozick, *The Nature of Rationality* (Princeton: Princeton University Press, 1993), 60. And cf. Tamara Horowitz, "Philosophical Intuitions and Psychological Theory," in M. DePaul and W. Ramsey, ed., *Rethinking Intuition: The Psychology of Intuition and its Role in Philosophical Inquiry* (Lanham, MD: Rowman and Littlefield, 1998), 153. Nozick's suspicion isn't accepted by everyone, and Frances Kamm—in a lucid discussion of the general issues of method raised here, "Moral Intuitions, Cognitive Psychology, and the Harming-versus-Not Aiding Distinction" *Ethics* 108 (April 1998), 463-488—has marshaled the pertinent objections to conflating the loss-versus-no-gain distinction with the aiding-versus-harming distinction. If our baseline is what would happen without our intervention—a thesis she ascribes to the psychologist Jonathan Baron—then, as she notes (skeptically), we would always worry less about not preventing deaths than about causing them.

trolley scenario and the footbridge scenario has to do with the emotional traction of the latter: "The thought of pushing someone to his death is, we propose, more emotionally salient than the thought of hitting a switch that will cause a trolley to produce similar consequences, and it is this emotional response that accounts for people's tendency to treat these cases differently."[26]

How did they confirm this hypothesis? By using functional magnetic resonance imaging (fMRI), which allows you to see which parts of a person's brain are most active at any time. Independent coders divided up various scenarios into moral and non-moral ones, and then into personal or impersonal ones. The switch-throwing scenario was classified as a "moral-impersonal" scenario; the footbridge example was classified as a "moral-personal" scenario, being, as Greene says, "up close and personal." Then Greene and his colleagues showed that when people were faced with the footbridge dilemma the parts of the brain that "lit up"—the medial frontal gyrus, the posterior cingulate gyrus, and the angular gyrus—were regions associated with emotion. On the other hand, they went on, "areas associated with working memory have been found to become less active during emotional processing as compared to periods of cognitive processing," and, indeed, the right middle frontal gyrus and the bilateral parietal lobe—both of which are associated with working memory—were "significantly less active in the moral-personal condition than" when subjects were thinking about "moral-impersonal" choices, like the trolley problem, or about choices that were not moral at all.[27]

Even those people who *were* in fact willing to kill the stranger on the footbridge confirmed Greene's fundamental claim. Because it looks as though they *overrode* their emotional instinctual response. People who decided that they should push the stranger off the bridge took significantly longer to respond than those who decided not to. They had the emotion-laden response, it seemed, and then they reasoned their way out of it.

Now, so far all we have is a psychological explanation. It may or may not impress you, depending on your confidence in our current grasp of functional neuroanatomy, but let's take the analysis at face value.[28] Confronted with the footbridge dilemma, what *is* the right thing to do? It is true—we didn't need scanners to tell us this—that, given the way our brains work, most people will find it distressing to push a stranger over a bridge, even for a very good reason. But is that a good enough reason not to do it? We have a conflict here between what our intuition tells us—don't kill an innocent stranger—and a powerful reflective thought: Isn't it better if only one person dies? Being told that our "intuition" involves the engagement of a different part of our brain from the part that has the reflective idea might just make it easier to side with that reflective judgment (which, let's grant, is sponsored by another set of intuitions). After all, as I said, philosophers have been engaged for a long time in seeking to find reasons that justify responding one way to the trolley case and

---

[26] Joshua D. Greene, R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, Jonathan D. Cohen, "An fMRI Investigation of Emotional Engagement in Moral Judgment," *Science*, Vol. 293, no. 5537, (14 September 2001), 2105-2108.

[27] Greene, et al. *loc. cit.* In a more recent experiment, two researchers at Northwestern, Piercarlo Valdesolo and David DeSteno, found a link between mood and moral judgment. Before presenting their subjects with the two trolley-car scenarios, they had half of their subjects watch five minutes of "Saturday Night Live," and the other half watch an unengaging documentary about a Spanish village. Those buoyed by the comedy show were more likely to say they would topple the large man. The researchers concluded that the negative emotions inspired by contemplating the up-close-and-personal homicide were offset by the cheerful spirits inspired by the TV show. See Piercarlo Valdesolo and David DeSteno, "Manipulations of Emotional Context Shape Moral Judgment," *Psychological Science* Vol. 17 Issue 6 (June 2006), 476. (So there are times when cheerful people may be less pleasant to be around!)

[28] There are plenty of skeptics around with other explanations of these data (or who don't believe they're reliably reproducible). As I say, my interest here isn't so much in whether these claims are true as in what would follow if they were. After all, we need to know *that* in order to decide what to think if they *are* true.

another to the footbridge case. It is fair to say that they have not come up with an answer that satisfies most people who have thought about the matter. I mentioned just now that the distinction between using someone to stop the trolley in the footbridge case and killing him as an unintended side effect might be helpful. But now consider another oft-discussed modification of the trolley case: the *loop scenario*. This time, when you throw the switch, the trolley goes onto a loop, which has the heavy stranger on it, but then comes back onto the track with the five people on it. What saves their lives now is the fact that the stranger weighs enough to stop the trolley. Once more, it seems, you're saving them *by* killing him, as in the footbridge case. His death is the means by which you save them, not an unintended side effect. But many people think that switch-throwing in this case is just as permissible as the switch-throwing in the first trolley case I described.[29] Being deluged with trolley problems is one of the professional hazards of modern moral philosophy.

Greene's psychological explanation raises the possibility that (as he thinks) there's just no good moral reason why we make this distinction. Our brains are so constituted that most of us will go one way in one scenario and another in the other: so we've been told *what makes us* respond as we do. We have (at least) two ways of deciding what to do. Sometimes one kicks in, sometimes the other. Which one kicks in is decided by whether the killing in question is "up close and personal" or not: and while that makes a difference to how we will *feel*, it doesn't have any real moral significance. In the absence of a compelling rational explanation, it's possible to conclude that that's all there is to it. As Greene has written in a discussion of other conflicting intuitions:

> Consider that our ancestors did not evolve in an environment in which total strangers on opposite sides of the world could save each other's lives by making relatively modest material sacrifices. Consider also that our ancestors did evolve in an environment in which individuals standing face-to-face could save each others' lives, sometimes only through considerable personal sacrifice. Given all of this, it makes sense that we would have evolved altruistic instincts that direct us to help others in dire need, but mostly when the ones in need are presented in an "up-close-and-personal" way.
>
> … Maybe there is "some good reason" for this pair of attitudes, but the evolutionary account given above suggests otherwise.[30]

How should a moral philosopher respond?

MORAL EMERGENCIES

We might begin by noticing something special about the trolley problem and the footbridge problem. They are both what I will call *moral emergencies*. They are cases, that is, where

1) You, the agent in the story, have to decide what to do in a very short period of time;
2) there is a clear and simple set of options,
3) something of great moral significance—in this case, the death of five innocent people—is at stake; and
4) no one else is as well placed as you are to intervene.

Those four features—the need for a quick decision, clear options, high stakes, the fact that you're the best placed person—together make these situations very different from most of the decisions that face us.

---

[29] See Frances Kamm, "Toward the Essence of Nonconsequentialism," in Alex Byrne, Robert Stalnaker, and Ralph Wedgwood, ed. *Fact and Value: Essays on Ethics and Metaphysics for Judith Jarvis Thomson* (Cambridge, Mass.: The MIT Press, 2001), 155-182.

[30] Joshua Greene "From neural 'is' to moral 'ought': what are the moral implications of neuroscientific moral psychology?" *Nature Reviews Neuroscience* 4, 846-850 (01 Oct 2003) *Perspective*.

First of all, it's in part the high stakes that allow us to narrow the range of options to consider. An unstated but obvious presupposition of these scenarios is that anything else we could be doing has much lower stakes, which is why we can ignore it. To make that clear, just add to each scenario, for example, the further stipulation that you see the trolley while rushing to defuse a nuclear bomb that only you can defuse, a bomb that will kill hundreds of thousands if you stop to deal with the trolley. Now, plainly, you must neither throw the switch nor push the 300-pound man. You must keep on your way.

Second, the fact that you, the agent in the story, have to decide fast means that you don't have time to get more information. Suppose the five people were old and suicidal; they'd gathered on the tracks in order to end their lives—they were expecting the trolley—and would be displeased to find their plans foiled. Suppose the other person—the man on the line, the stranger on the bridge—is about to find a cure for Alzheimer's. It's a convention of such scenarios that you've been told all you need to know. (Or perhaps it's just an application of the Gricean maxim of quantity: make your contribution as informative as is required for the purpose of the discussion, but not more so.)

Third, the fact that you are the person on the spot means that you bear a responsibility that you wouldn't have if there were others closer or better equipped to act. So it looks as if the choice is yours. You can't just say, "It's none of my business."

The fact that the footbridge and the trolley problem are moral emergencies suggests that they are exactly the kinds of situations in which we will be guided, as Greene might put it, by the medial frontal, posterior cingulate and angular gyri rather than the parietal lobe. The person at the footbridge, the bystander at the switch, has little time for reflection, information gathering and processing, deliberation. Adrenaline will be pulsing through her veins. It will be hard, in Hamlet's happy phrasing, for one's "native hue of resolution" to be "sicklied o'er with the pale cast of thought." In such circumstances, there is much to be said for an instinctive reluctance to push strangers over bridges, even if it might be, all things considered, in these horrific circumstances, the best thing to do.

That response looks, in short, as though it meets the conditions I set out in the last chapter for being a good heuristic. Suppose that, in fact, one ought to push the heavy stranger off the footbridge. Even so, there might be reasons for wanting to be the kind of people who couldn't do it. Holding onto revulsion against killing people "up close and personal" is one way to make sure we won't easily be tempted to kill people when it suits our interests. So we may be glad that our gyri are jangled by the very idea of it. That heuristic will lead us to miss the chance to save those five strangers on the trolley-track. But how often does a situation like that arise? As fast and frugal policies go, this one is very much more likely to lead us in the right direction.

Evolutionary psychologists have offered similar explanations for the prospect theory that was offered in explanation of the results in the Asian Flu case. For most of human history we lived in conditions of subsistence, close to starvation, where the risk of loss was the risk, in effect, of dying. While gains above subsistence are nice, they're not worth taking that risk for. So creatures that do what prospect theory requires seem more likely to survive. (It would take us a long way afield to discuss how good an explanation this is.) If acting according to prospect theory usually leads us to do what is, so far as the normative philosopher judges, the right thing, then, despite the fact that people are acting for the wrong reasons, they are often going to do what is right. She will treat it as a rule of thumb, a heuristic, and, so long as she's a consequentialist (like Greene) and not, say, a virtue theorist, that heuristic might suit her just fine. She may, of course, want to persuade people that there is a better way of getting to their conclusions, a route that will never lead them astray. But that project will be less urgent if the implicit principle that actually motivates them is, in most normal circumstances, a good enough guide. Notice that appeal to the idea of a heuristic here is not open to the objection I raised in the last chapter. In this instance, the heuristic is one that guides our action according to a presumed standard: it is about what we do, not what we are. We can therefore use the standards of normal means-end rationality to evaluate whether the heuristic does well as a substitute for applying the standard directly. And if there is no currently known way to change our intuitions—if they are a fixed feature of our psychologies—then learning when and how they mislead us will help

us to overrule them when we should.

Nor will evolutionary theorists be surprised to learn that we'll overcome our aversion to up-close killing under the right circumstances. Suppose—this is another chestnut of moral philosophy—that you and some fellow spelunkers have allowed a fat man to lead you out of the cave you've been exploring, and he gets stuck, trapping the rest of you inside. Worse still, the cave is flooding, and if you can't get out, you'll all drown, except the fat man. But one of you has a stick of dynamite, and can blast the man out of the mouth of the cave. Can you blow him up to save yourselves? In at least one recent survey, most people said yes.[31] Here, we might suppose, some element of self-defense enters into our judgment: in contrast to the trolley cases, the actor is securing his or her own survival, and, when the stakes are life and death, our moral common sense permits a special concern for oneself.

It is an interesting and unobvious assumption, which hasn't had the attention it deserves, that our responses to imaginary scenarios mirror our responses to real ones. Greene's account of these cases presupposes, in other words, that our intuition about what to do in the imaginary case is explained by the activation of the very mechanisms that would lead us to act in a real one. Without that assumption, his explanation of why we respond as we do fails. When I think about the footbridge scenario there is, in fact, no 300-pound man around and no one to save. There is no emergency. I'm making a guess about how I'd respond in those circumstances.[32] Greene would say, not implausibly, that the responses activated through the process of imagining yourself to be in that situation look like the ones we should expect to be activated if we really were in that situation. And, of course, there are moral accounts of long standing according to which the right thing to do is what a benevolent but fully-informed observer would advise you to do: in which case we should perhaps put more stock, morally, in the questionnaire-answerer's counsel than in the switch-thrower's conduct.

FOLK PSYCHOLOGY UNPLUGGED

Let me consider one final line of empirical investigation, which shines an equally harsh light on the folk psychology of moral appraisal. Here, the subject of inquiry is the relation between our moral judgments of an act and our ascriptions of responsibility or intent.

In one set of trials, the philosophers Shaun Nichols and Joshua Knobe set out to elicit the intuitions that non-philosophers had about an old philosophical debate: the relation between determinism and moral appraisal. If we live in a deterministic universe, where all events have causes—are necessitated by the laws of nature and the previous state of the cosmos—can people be held morally responsible for their actions? "Compatibilists," in the philosophical shorthand, think that determinism is, in some sense, compatible with free will (and thus the moral appraisal of actions); "incompatibilists" do not. Nicholas and Knobe were able to provoke both compatibilist and incompatibilist intuitions from people—depending on whether the scenarios they presented to them were "affect laden."

---

[31] The survey was conducted by the BBC, which collected the response of 12,000 participants, 76.1% of whom voted to ignite the stick of dynamite: http://xrl.us/fatcaver. The stakes for the actor count. Imagine a conductor of an out-of-control trolley who, approaching a fork, must choose one of two tracks, one with five pedestrians who will die if he hits them, the other with just one. Suppose, perversely, he chose the track with the five pedestrians. We'd condemn his decision. Yet even if the solitary pedestrian had the ability (via some remote-control device) to switch the train back to his own track, and save their lives at the cost of his own, almost nobody would consider that sacrifice to be obligatory. That sacrifice, if he made it, would be viewed as one of heroic supererogation, and venerated as such.

[32] Indeed, experimental research into "affective forecasting" raises some skepticism about our ability to predict how we will feel in some future situation, although (so far as I know) none of this research—pioneered by Jonathan Schooler, Timothy Wilson, and Daniel T. Gilbert—involves moral appraisal in any obvious way.

Here's how they did it. One group of participants was instructed to consider a universe that was described as being completely deterministic. (Everything that happens *had* to happen.) When asked whether people could be "fully morally responsible for their actions" in such a universe, 86% of subjects said no. When they were asked about a particular inhabitant of this universe, Mark, who, as he has done in the past, "arranges to cheat on his taxes," most people denied that he was fully responsible for his actions. So their intuitions about the abstract issue were preserved when they responded to this rather humdrum case. But their intuitions shifted when they were presented with a much more affect-laden scenario: "In Universe A"—a fully deterministic one—"a man named Bill has become attracted to his secretary, and decides that the only way to be with her is to kill his wife and 3 children. He knows that it is impossible to escape from his house in the event of a fire. Before he leaves on a business trip, he sets up a device in his basement that burns down the house and kills his family."[33] Most subjects given this scenario said they thought that Bill was indeed fully morally responsible.

Nichols and Knobe are inclined to think that people's intuitions about Bill reflect some sort of "performance error," in which case these intuitions shouldn't count in favor of compatibilism (whether or not compatibilism is correct). They wonder whether we're looking at a "moral illusion" analogous to the optical illusion that makes the two lines of the Müller-Lyer diagram look like they're of different length, even when we know they're the same. They're also struck by the fact that when you point out the seeming contradiction and ask people to reconcile their warring intuitions, half jump one way, half jump the other. The dissensus among lay people, they notice, seems awfully similar to the dissensus among philosophers.

Of course, results of this sort don't interpret themselves; they have to be interpreted. For one thing, it may be overreaching to ascribe the philosophical doctrine of compatibilism to the Bill blamers. Especially if a "performance error" is involved, they may not be asserting the compatibility of determinism and free will; they may be so outraged that they've put the stipulated background assumption *way* in the background. We can also fuss about details: for instance, the word "decided," in the Bill scenario, has a distinctly volitional hue. The vignette situates Bill (but not Mark) in a web of interpersonal relationships, the sort of context that, P. F. Strawson thought, was bound to elicit "participant reactive attitudes," like reproach, approbation, gratitude, and other responses that suggest an ascription of moral agency.[34] Indeed, to call the Bill scenario "affect-laden"—as if "affect" were some fungible substance, like blood or bile—is to under-describe it: the scenario elicits a particular kind of affect, moral repugnance. One can imagine other "affect-laden" scenarios that involve no moral judgment (e.g., a man who suffers a series of gruesome calamities, for which nobody is to blame) and that wouldn't trigger a suspension of the incompatibilist intuition.

Nor would the results have come as a surprise to Gilbert Ryle. "We discuss whether someone's action was voluntary or not only when the action seems to have been his fault," he ventured more than half a century ago. It was a *déformation professionelle* of philosophers, he thought, to

---

[33] Shaun Nichols and Joshua Knobe, "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions," *Nous*, forthcoming. (Available at http://www.unc.edu/~knobe/Nichols-Knobe.pdf. This quote is from page 12 of that version. The 86% figure comes from page 13.) Their study draws on and expands research published in Eddy A. Nahmias, Stephen Morris, Thomas Nadelhoffer, and Jason Turner, "Surveying Freedom: Folk Intuitions about Free Will and Moral Responsibility," *Philosophical Psychology*, Volume 18, Number 5, October 2005, 561-584. These researchers tend to describe the deterministic universe in scientific terms to their subjects (there may be talk of the laws of physics, of a supercomputer able to predict events with complete accuracy); but the basic concept has been familiar to people from other times and cultures under other terms: curses, oracles, predestination, and so on. A Greek from the fifth century BC could ponder whether Oedipus was fully morally responsible for his actions. See A. W. H. Adkins *Merit and Responsibility* (Oxford: Clarendon Press, 1960).

[34] P. F. Strawson, "Freedom and Resentment," *Proceedings of the British Academy* (1962), 187-211. Reprinted in *Freedom and Resentment and other Essays (*London*:* Methuen, l974).

describe as "voluntary" acts that were commendable or satisfactory. In his view, "The tangle of largely spurious problems, known as the problem of the Freedom of the Will, partly derives from this unconsciously stretched use of 'voluntary' and these consequential misapplications of different senses of 'could' and 'could have helped.' "[35]

We know, too, how easily an engaging story can defeat our allegiance to this or that dictum. Should cars parked in front of fire hydrants, such as one belonging to Bob the accountant, be towed? Sure. Now tell me a story about Joanna, a good woman who's having a bad day, with details about her hopes and dreams, her kindness to an ailing friend, her preoccupation with a troubled child. I don't want *her* car to be towed. In the abstract, we're opposed to theft. But any competent screenwriter could concoct a story about a cool, big-hearted car thief, shadowed by a Javert-like detective as he prepares to pull off one last heist—and get you to switch sides, at least until the credits roll. In the abstract, people often hold that it's inappropriate to respond to machines with reactive attitudes like reproach or approbation; but the "Terminator" movies handily sweep aside those theoretical commitments.[36] If there's a discordant element in the story, in short, I just might read past it.

Still, none of these considerations undermines the basic idea that strong disapproval may swamp our previous theoretical commitments about, say, causality. Indeed, in another study, Joshua Knobe showed how people's intuitions about whether an act was intentional seemed to depend on their appraisal of that act. This time, he devised a pair of scenarios that didn't require any science-fiction stipulations or, indeed, much imagination. In one of them, the chairman of a company is asked to approve a new program that will increase profits and also help the environment. "I don't care at all about helping the environment," the chairman replies. "I just want to make as much profit as I can. Let's start the new program." So the program is launched and the environment is helped. A second version is identical—except that the program will *hurt* the environment. Once again, the chairman is indifferent to the environment, and the program is launched in order to increase profits, with the expected results.

When Knobe presented these scenarios to subjects in a controlled experiment, he found that when the program helped the environment, only 23% percent agreed that the chairman had "helped the environment intentionally." When the program harmed the environment, though, 82% agreed that the chairman had "harmed the environment intentionally." And the pattern recurred when various other scenarios were tested. In the "harm" scenarios, people will assent to the statement that the chairman harmed the environment in order to increase profits; but, in the "help" scenarios, they generally won't agree that he helped the environment in order to increase profits. So while you might have supposed that whether you find an act blameworthy depends on whether you have concluded that the act was intended, Knobe's research suggests that something like the reverse is true.

In general, Knobe argues, our intuitions about intentional actions "tend to track the psychological features that are most relevant to praise and blame judgments," and different features become relevant "depending on whether the action itself is good or bad." His analysis here doesn't fault folk psychology for being distorted by moral considerations; the idea, rather, is that such

---

[35] Gilbert Ryle, *The Concept of Mind* (Chicago: University of Chicago Press, 1949, 2002), 71. Ryle is probably right about "most ordinary employments" of the terms *voluntary* and *involuntary* but it's easy to come up with not-very-exotic counterexamples—as, for instance, if you find out that someone has been kind to you because he's been coerced; or when we commend a soldier's brave act as especially heroic because he volunteered for a risky mission, rather than having been ordered to do it.

[36] At the same time, media scholars have shown that the way people apprehend narratives will be shaped by that they take to be common sense. For instance, researchers found that a group of Israeli Arabs watching "Dallas" wrongly assumed that Sue Ellen, having left her husband, had sought refuge with her father, because returning to your family in those circumstances is what made sense to them. Their effort to make the show culturally intelligible overrode the information that she'd actually taken refuge at the family home of her previous lover. See Tamar Liebes and Elihu Katz, *The Export of Meaning: Cross-Cultural Readings of Dallas* (London: Polity Press, 1994).

considerations "are playing a helpful role in people's underlying competence itself." (He has come to praise folk psychology, not to bury it.) Once we revisit the question what folk psychology is *for*, we might see that its tasks include the apportioning of praise and blame. Knobe's findings are, as he recognizes, congruent with Strawson's arguments about reactive attitudes—in particular, the hovering thought that such attitudes might be the basis of our judgments about responsibility, rather than vice-versa.[37]

Inspecting these help/harm studies, as with the compatibilism studies, philosophers will want to poke and prod, of course. Perhaps most people wouldn't describe an action's unsought side effect as either intentional or unintentional without the forced choice of a questionnaire. Perhaps there's something wonky about our use of that English adverb "intentionally," such that the study could not be reproduced in Korean. In Knobe's analysis, the word "behavior" is used to refer not to the chairman's decision to launch the program, but to the chairman's harming or hurting the environment. Perhaps that term is question-begging: philosophers might disagree about whether to designate downstream consequences of an agent's decision, even if foreseen, as an agent's "behavior." For the subjects of the experiment, the question has similar framing effects. That is, the question *Did the agent* Φ *intentionally?* foregrounds the proposition that *the agent* Φ*'d,* and (arguably) produces its own salience distortions.

Knobe has ventured that "the process leading up to people's intentional-action intuitions is a kind of heuristic," constructed so that it normally accords with our ascriptions of praise and blame.[38] But perhaps those ascriptions of praise and blame themselves flow from a heuristic of another kind. The intuitions elicited by those scenarios about the mercenary chairman look to be socially beneficial: nobody needs to be encouraged to promote their self-interest, whereas people do need to be discouraged from promoting their self-interest at the expense of others. In a manner congenial to consequentialist accounts of moral responsibility, this practical asymmetry corresponds to the asymmetry in our intentional-action intuitions. (For similar reasons, perhaps, we penalize firms for negative externalities but don't reward them for positive ones.) If these considerations help explain our judgments, we would expect praise for the sought-after positive effects of courageous, self-sacrificing acts, even if success was the result of sheer fortuity (and, *mutatis mutandis*, blame for sought-after but fortuitous negative effects), which is just what Knobe has found in other studies.[39]

There are other reasons to suspect that the basic results will prove pretty robust. For one

---

[37] Joshua Knobe, "The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology," *op. cit.* "The existence of the general framework of attitudes itself is something we are given with the fact of human society," Strawson argues. "As a whole, it neither calls for, nor permits, an external 'rational' justification." The optimist is prone to an "incomplete empiricism, a one-eyed utilitarianism. He seeks to find an adequate basis for certain social practices in calculated consequences, and loses sight (perhaps wishes to lose sight) of the human attitudes of which these practices are in part the expression… It is a pity that talk of the moral sentiments has fallen out of favor. The phrase would be quite a good name for that network of human attitudes in acknowledging the character and place of which we find, I suggest, the only possibility of reconciling these disputants to each other and the facts." (Strawson *op. cit.*)

[38] Joshua Knobe, "Folk Psychology and Folk Morality: Response to Critics," *Journal of Theoretical and Philosophical Psychology.* Available at http://www.unc.edu/~knobe/ResponseCritics.pdf. "People do not normally praise or blame agents for reluctant side-effects, and the process is therefore constructed in such a way that it classifies all reluctant side-effects as unintentional," he elaborates. "In certain cases, however, reluctant side-effects may still elicit feelings of blame. In those cases, people's feelings of blame end up diverging in tell-tale ways from their intentional action intuitions." When a drunk driver runs over a child, the fact of intoxication has two faces; it tends to exempt him from direct responsibility in the act while inculpating him in (let's say) an act of criminal negligence.

[39] Joshua Knobe, "The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology," *op. cit.*

thing, the intuitions Knobe describes are intaglioed on common law, not to mention our civil and criminal codes. We regularly hold people culpable for the unsought consequences of their behavior. We have a rich legal vocabulary—"depraved indifference," "reckless disregard," "murder in the second degree," "involuntary manslaughter"—for such violations. They establish agent ownership, in effect, of harms that are not even foreseen but merely foreseeable, and they accord with a broad moral consensus that holds people responsible for such side effects of what they do. All these features of the law reflect intuitions about moral responsibility that can be elicited with questionnaires and scenarios.

When you take a closer look at folk theories of moral responsibility, in fact, they start to resemble a patchwork quilt: that is, there *are* patterns … many different ones. Here's a criterion that researchers have identified, on the basis of various scenario-based experiments: although—as the vignettes about the mercenary chairman showed—we'll hold someone responsible for the foreseen-but-unsought side effects of an action when those effects are bad, we'll credit the person with good consequences only when the person was aiming for them. Here's another criterion, operative in other situations: you bear greater responsibility for an accident that happens in the course of your doing something discreditable than you do for an accident that happens in the course of your doing something commendable. And yet another: you bear greater responsibility for an accident that has severe consequences than you do for an accident with mild ones. Reviewing the research literature, Knobe and Doris identify an assumption, which they call "invariance," that underlies previous work on moral responsibility. It's the idea that "people should apply the *same* criteria in *all* their responsibility judgments." The evidence, they argue, shows that people in fact apply different criteria in different sorts of cases. Then they throw down the gauntlet: "This discovery in empirical psychology leaves us with a stark choice in moral philosophy. One option would be to hold on to the goal of fitting with people's ordinary judgments and thereby abandon the assumption of invariance. The other would be to hold on to the assumption of invariance and thereby abandon the goal of fitting with people's ordinary judgments. But it seems that one cannot have it both ways."[40]

As long as the patterns identified are countable and compassable, we might suppose that one invariance could be replaced with, say, five invariances. We shouldn't be surprised, though, to find many situations that aren't tightly cabined by one or another ideal type but lend themselves to multiple descriptions. Such variegation would also be predicted by an account of moral epistemology that arises from work that cognitive psychologists have done on neural networks and the nature of categorization. Ever since Eleanor Rosch's work in the early 1970s, it has been a common thought that people categorize objects, situations, and concepts not on the basis of abstract definitions, but by reference to prototypes. We immediately categorize an apple as a fruit; it takes us longer to decide that an olive is one, too. For Paul Churchland, then, moral disagreements should be seen not as disputes about "rules" so much as disputes about which prototypes best accord with a given situation. (Is a fetus more like a tiny person or more like a growth?)[41] On a prototype model, we

---

[40] Josh Knobe and John Doris, "Strawsonian Variations: Folk Morality and the Search for a Unified Theory," forthcoming in John Doris et al. *The Handbook of Moral Psychology* (Oxford: Oxford University Press). Available at http://www.unc.edu/~knobe/Knobe-Doris.pdf. Notice that all these patterns have criminal-justice correlatives. Accidentally causing a death in the course of committing a felony, for instance, is usually an offense in itself, punished more severely than the accidental death would be otherwise. A driver who runs over a child who has suddenly darted into the road may be held blameless—but not if he was also drunk, or had run a light, or was racing from a bank he'd robbed, even when there's no specific causal link between the two circumstances. As Bernard Williams has famously observed, in our intuitions about responsibility, we have more in common with the authors of the Greek tragedies that we often suppose.

[41] Paul Churchland, "The Neural Representations of the Social World," in Larry May, Marily Friedman, and Andy Clark, ed., *Minds and Morals* (Cambridge: The MIT Press, 1996), 103. "People with unusually penetrating moral insight will be those who can see a problematic moral situation in more than one way, and who can evaluate the relative accuracy and relevance of those competing

shouldn't expect invariance, because our perceptions, in a range of situations, will be subject to Necker-cube-style ambiguity, typicality effects, and fuzzy boundaries, and thus a competition among (or an overlay of) several plausible prototypes. You don't need that apparatus to explain why people may have a variety of moral responses to events, though. The numerous allies of moral or value pluralism often suppose that the judgments people make will depend on what's perspicuous in a particular situation, that we have no fixed algorithm for deciding among our normative concerns. Either way, if it's simply a fact about us that our folk moral psychology is deeply heterogeneous—not to mention riddled with moral mirages—then the convergence and coherence promised by reflective equilibrium are bound to be all the more elusive.

SEEING REASON

Human beings are inclined to believe that a tethered ball that circles around a pole will, once cut loose, retain a curved trajectory. That's a deep feature of our "folk physics," and it's completely wrong. Knowing that we're prone to this misapprehension—the way we're prone to the Müller-Lyer illusion—can help us to overcome it. The mechanisms of the eye are usually reliable in guiding us to correct beliefs about the visible world. But there are, of course, illusions that are consistently produced in specifiable circumstances. Once we know this, we can learn to avoid reliance on our eyes in such circumstances, even though we cannot change the built-in mechanisms of the visual system. I don't slow down to avoid the illusory puddles of water that appear on the highway before me in the heat of summer. I can't stop "seeing" them. I *can* learn that they aren't there. Does research into the vagaries of our folk moral psychology enable us to do something similar?

Understanding where our intuitions come from can surely help us to think about which ones we should trust. Other psychological research will suggest that some of our intuitions will survive, even in circumstances where they have misled us. Here the analogy with the scientific study of our cognitive processes is, as I've said, quite natural. Just as an awareness of optical illusions helps us avoid being misled by them, cognitive psychologists have shown that all of us—including professional statisticians—are bad at taking account of probabilistic information. But this very discovery comes with the knowledge that there are ways around the problem: learning to do statistical calculations rather than following our hunches, or, perhaps, recasting the probabilities as frequencies, about which our intuitions seem more reliable. The proposal that—to put it very crudely—it's our feelings that guide us to the intuition about the footbridge case, while our reason guides us in the original trolley problem is the sort of thing we might want to consider in deciding whether that intuition is right. So is the fact that even if killing the stranger is, in these bizarre circumstances, the best thing to do, I have good reason—based in our current best understanding of human psychology—not to try to make myself the kind of person that would do it.

Again, we might, as various theorists have suggested, treat our intuitions in clusters of cases as the output of a heuristic. But that, of course, entails that some other, better guide, determines whether the heuristic has steered us right or wrong. To identify a heuristic, remember, we need a standard. And here's where the analogy with folk physics falters: professional physicists can happily build models of the universe that involve structures—Calabi-Yau manifolds and Gromov-Witten invariants—that are utterly remote from our capacity to imagine them, in any meaningful sense. By contrast, moral psychology, however reflective, can't be dissociated from our moral sentiments, because it's basic to how we make sense of one another and ourselves; in a deliberately awkward

---

interpretations," he argues. "Such people will be those with unusual moral *imagination*, and a critical capacity to match." In his view, "The morally successful person has acquired a complex set of subtle and enviable skills: perceptual, cognitive, and behavioral," and he draws a connection to Aristotle's insistence that virtue involved skills accumulated over a lifetime of experience, and consisted in "largely inarticulable skills, a matter of practical wisdom. Aristotle's perspective and the neutral network perspective here converge." (105, 106.) What this (otherwise rather plausible) account of moral expertise leaves out is the task of providing explicit moral justification, which can also be a practical affair.

formulation of Bernard Williams's, moral thought and experience "must primarily involve grasping the world in such a way that one can, as a particular human being, live in it."[42] Could it be that your moral intuitions themselves, fallible though they are, provide a standard for judgment? Moral agents often think so, which is why, in the footbridge cases, philosophers have generally spurned the "obvious" way out of the problem—which is to maximize lives saved by killing the stranger—and struggled to find some other way. Do we have a way to settle the matter? The analogy with color perception might be helpful here. Suppose someone were to propose, as a heuristic, this "blue rule":

> If something looks blue and you have no special knowledge about your eyes or the lighting, you should believe that it's blue.

I think it is natural to respond that this isn't just a heuristic. If this isn't the right way to tell whether something is blue, either there's no such thing as looking blue or there's no such thing as being blue. Yes, following this rule can lead you astray; but only if there is something odd about the circumstances that you didn't know about. If it turned out that red things sometimes looked blue when your eyes and the lighting were normal, you'd have reason to conclude that there was something incoherent about the way we understand color. Color is, as Bishop Berkeley used to say, one of the proper objects of vision. It's part of what vision is for; and, from the point of view of a perceiver, the blue rule is the only way of proceeding that makes any sense.

Now consider this rule:

> If something seems intuitively wrong (or right) and you have no special knowledge that suggests your moral intuition is distorted, you shouldn't (or should) do it.

This, someone might argue analogously, isn't a heuristic either; from the point of view of the agent, the rule is a constitutive element of the very idea of wrongness. Now, I could have couched the rule in terms of belief: if something seems wrong, you should believe you shouldn't do it. But our interest here is practical: we want to consider what qualifies as a reason to Φ, a reason not to believe but to do something. The analogy I want to explore is between certain kinds of reasons for perceptual belief and certain kinds of reasons for action, not between reasons for perceptual and reasons for normative beliefs.[43]

To see how far we can get with the comparison between the perceptual and the ethical, let's consider the general schema the two cases share. In each, there is a response—belief about a color in the first, action in the second—that you ought to make when a certain condition obtains, provided there is no further reason that you should not respond in that way. Philosophers say you have a "pro tanto" reason to Φ, when you are aware of a state of affairs, S, that is a reason for Φ-ing, even if there are other further states of affairs awareness of which would mean you ought not to Φ. If you have a pro tanto reason to Φ and you're aware of no other relevant considerations, you ought to Φ.[44]

---

[42] Bernard Williams, *Moral Luck, op. cit.,* 52

[117] There are many disanalogies, to be sure. Usually, for example, reasons for action are generated by way of our imagining a situation not by our observing one.

[44] Ross made a similar move when he suggested that we cast duties in terms of "tendencies"—to say that it tends to be our duty to pay our debts or to relieve distress and thus resolve the "theoretical problem of conflict of duties," by allowing that the two may conflict, even if we decide that, on the whole, it's our duty to do one rather than another. "The absoluteness of the law of justice and of the law of benevolence is preserved if they are stated as laws of tendency," he concluded. W.D. Ross, "The Basis of Objective Judgments in Ethics," *The International Journal of Ethics,* vol. 37, no. 2 (January 1927), 127. Though he became identified with the less felicitous term "prima facie duties" (which sound as if they may vanish upon inspection, rather than being trumped), he intended the tendency interpretation; these duties (including duties of fidelity, reparation, gratitude, justice, beneficence, self-

So we can say, in this convenient shorthand that a thing's looking blue is a pro tanto reason to believe that it is blue: you ought to believe something is blue, that is, if you are aware that it looks blue. And that's so, even if, were further evidence to come along—that a wicked scientist put blue contact lenses in your eyes overnight, say—the new evidence might rationally require you to believe that it wasn't blue.

Similarly, perhaps the fact that something seems wrong is a pro tanto reason not to do it: you ought not to push the heavy stranger off the bridge if it seems to you wrong to do so. And, once more, that would be true even though there are circumstances in which further reasons might require or permit you to do it.[45] The proposal is, then, that our moral intuitions give us pro tanto reasons to act (or not to act), just as our senses give us pro tanto reasons for belief.

The obvious problem with this proposal is that, as we've seen again and again, our moral intuitions may be unreliable. Remember that comment offered by the experimental subject who was asked to think about Dan, the student council rep: "I don't know why it's wrong; it just is." How are we to decide which among our intuitions provide valid pro tanto reasons? Does the mere fact that it strikes me that I don't feel like doing something—reaching out to Joe, who has fallen into a sewage tank, say—give me a pro tanto reason not to help him? Perhaps the fact that I'm disgusted is a pro tanto reason not to stick my hand out; one that is defeated by the fact that Joe needs help (which gives me a pro tanto reason to reach out to him, in the context, that is not defeated, let us suppose, by any other reasons). How can I distinguish between a pro tanto reason and the mere fact that I do or don't feel like doing something? In the case of perception, we can distinguish the visual properties of things and say that it is our visual awareness of these that gives us pro tanto reasons for (visual) perceptual judgment. And we can identify the visual properties by identifying our eyes as the organs that detect them. But we can't identify the reason-for-action-giving properties of situations by pointing to any organ: and, indeed, moral psychologists are inclined to doubt that there are any such organs. They don't think we have what the Enlightenment philosophers called a "moral sense."

What we do have are many, many, responses that give us pro tanto guidance as to how we should feel and act. Let's call these responses "evaluations." One sort of evaluation is an emotion: fear is a pro tanto reason for avoiding something. More precisely, when I fear that P, I have a pro tanto reason to take action to prevent it coming about that P. (Fear of objects is grounded in the belief that some state of affairs whose coming about we fear is likely in their presence.) Reasonable fears are therefore grounded in judgments that our interests will be set back if we don't try to prevent them, because the fact that something will set back our interests is a pro tanto reason for us to prevent it. But we have not only fears but also phobias. Phobias arguably don't give us reason to do anything ... except to try to get rid of them. And so, when we must distinguish the cases where we should and should not take notice of our feelings, we will want to take account of the empirical research at hand. By illuminating odd features of our evaluations, showing where judgments about cases of a certain form diverge from our other, considered judgments, that research can be a useful adjunct to our intuitions, especially if we consider "attentive reflection," as Thomas Reid urged, to be a form of intuition, too.

---

improvement, and nonmaleficence) were, as he also put it, "conditional duties," and one became actual when an agent decided that it was "more incumbent than any other." Obviously, some ostensible duties may be prima facie in the sense that they aren't really duties at all.

[45] I should make a technical aside here: the state of affairs which warrants not pushing him here is not it's being wrong (which is a state of affairs that some people don't believe in) but your having the intuition that it's wrong (which is a state of affairs whose existence is sometimes quite uncontroversial). But sometimes you ought to believe something not because you are aware of a state of yours but because you are aware of the world. So something's being blue is also a (boring) pro tanto reason to believe that it's blue. If you're aware that it's blue you have a reason to believe that it is. So, too, if something is wrong you have a pro tanto reason not to do it.

All this sounds soothingly hygienic and prudent. But we are bound to be left with a lingering sense of unease. In a sense, all this talk of *mistakes*—of loss/gain anomalies, help/harm asymmetries, priming effects, performance errors, and so forth—has been a distraction from a larger threat. Learning that our intuitions are imperfect has, at least, one comforting implication: to be told that we sometimes get things wrong implies that we can, in principle, get them right. The lingering unease I mentioned has a different source: the dissonance between viewing moral sentiments as, in some measure, constitutive of a moral judgment and viewing them as a device that nature has bequeathed us for social regulation.

Four decades ago, Strawson noted that human attitudes had become objects of study in the disciplines of history and anthropology but also—and, he thought, of greater significance—in psychology. The humanists could tell us what aspects of our intuitions were specific to our time and place; but the psychologists, more corrosively, could make us distrust habits of mind that were universal and enduring. Strawson had, in part, a sociological point to make: "the prestige of these theoretical studies," he feared, "is apt to make us forget that in philosophy, though it also is a theoretical study, we have to take account of the facts in all their bearings; we are not to suppose that we are required, or permitted, as philosophers, to regard ourselves, as human beings, as detached from the attitudes which, as scientists, we study with detachment."[46] But the more fundamental concern has to do with the task of moral theory. We cannot content ourselves with the claim that a given bundle of attitudes solves a social coordination problem—that it is, from some objective point of view, *adaptive*. That's a possible standard for a heuristic, but (at least without further elaboration) it's not moral in nature. The claim that someone, on prudential grounds, should be punished does not entail that he or she is blame*worthy*. As Hume put it, when a man reproaches another morally, he "must choose a point of view, common to him with others; he must move some universal principle of the human frame, and touch a string to which all mankind have an accord and symphony."[47] Moral thought aspires to a register that is universal without being impersonal. It can't be just an esoteric guidebook for a supreme legislator; it has to be intelligible to us ordinary persons. That's why an explanation cannot replace a reason, why a causal account cannot supplant a moral justification. In ordinary life, you'll notice, we invoke psychological explanations only when we're seeking exemptions from moral agency. ("I'm sorry I said that—I haven't been sleeping well lately.") To introduce them into a conversation about justification must seem, in effect, to change the subject.

The vast majority of our decisions, to be sure, are made unreflectively, almost on autopilot; an outfielder can catch a ball without the help of Euler's Method for solving differential equations, and most of the time you can make expert moral decisions without engaging in conscious axiology. Indeed, you'd better, or you'll never get anything done. We needn't worry that, in actual appraisals, we generally start with a verdict, and work backward from there. Tellingly, while most people agree, in the footbridge scenario, that it would be wrong to topple the 300-pound man, they do not agree about why it is wrong: in experiments, the justifications that subjects offer are all over the map. Given that morality is practical, we should expect it to involve (in Ryle's venerable distinction) "knowing how," not just "knowing that"—tacit skills in contrast to the possession of propositional content, or the apprehension of truths. Yet the outfielder's intuitions about where the ball will fall do track with the actual movements of the ball. Many philosophers (especially those inclined toward moral realism, which may hold that moral predicates have causal power) will look for *some* link between the explanation and the justification for a decision. And even philosophers who are satisfied

---

[46] Perhaps we ought to modify our attitudes in the light of such studies, Strawson allowed; but we should not expect our moral sentiments—our reactive attitudes—to disappear. "What *is* wrong is to forget that these practices, and their reception, the reactions to them really *are* expression of our moral attitudes and not merely devices we calculatingly employ for regulative purposes. Our practices do not merely exploit our natures, they express them." Strawson, *op. cit.,* 210.

[47] David Hume, *An Enquiry into the Principles of Morals (1777),* Section IX, Conclusion, Part I http://etext.library.adelaide.edu.au/h/hume/david/h92pm/chapter9.html.

with good-enough-for-government-work heuristics, as a default procedure, will insist that people be able to provide reasons when reasons are called for.

"To say that a judgment is due to causes is to imply that it is not based on reasons," Ross wrote, "and so far as this is the case we have no ground for believing it to be true; it will be a mere accident if it is true."[48] This was meant to be a perspectival point, so to speak, not a logical one. As self-conscious moral agents, we can sometimes reflect on ourselves as natural creatures, part of a causal system, as well as reason-responsive ones. Our moral universe is both caused and created, and its breezes carry the voices of both explanations and reasons.

In moving from the psychologist's concern with naturalistic explanation to the philosopher's concern with reasons, we move from the issue of what sort of dispositions it might be good for us to have to the issue of what sorts of sentiments might count as moral justifications. These aren't just clashing accounts; they are—in ways I'll explore further in the next chapter—two perspectives. One is that of a sort of cosmic engineer, crafting our natures; the other is that of the moral agent thus crafted. And only a misguided monism would force both perspectives into one. In ethics as in optics, we need stereoscopy to see the world in all its dimensions. The claim that it's good, all things considered, to have certain sentiments, dispositions, and attitudes can be separated from the claim that these sentiments, dispositions, and attitudes are (at least in part) internally constitutive of normativity—of the content of moral propositions. But we needn't choose between them. The first claim defends a disposition by, say, its efficacy in securing some good or ensemble of goods. The second claim holds that some good derives, in a specific sense, from the disposition. The claims are different, but not rivalrous: they do not have to compete in the same explanatory space.

We can, after all, hive off the evolutionary question of why we have some of the evaluations that we have from the question of what, given that we do have these evaluations, we ought to do. Appeals to fitness might explain the evolution of our reactive attitudes; by itself, that fact takes nothing away from the authority of those reactive attitudes—nor, indeed, does it undermine our higher-level moral appraisals of our first-order moral appraisals. And so we can begin to dispel the unease I mentioned a little earlier. Nature taught our ancestors to walk; we can teach ourselves to dance.

This story can only be made compelling by saying more about how our picture of ourselves as natural creatures, the subjects of psychology and the social sciences, can fit with a view of the world in which our intuitions are not just feelings, emotions bubbling up from our neurons and our glands, but also responses to normative demands. We need to explore the relationship between the perspective of the cosmic engineer and the perspective of the agent she engineers; and—a project I'll discuss further in the next chapter—we need to be able to live with both perspectives.

[48] W. D. Ross, "The Basis of Objective Judgments in Ethics," *op. cit.,* 113.