The Affective Dog and Its Rational Tale

Peter Railton

For the Yale Legal Theory Workshop, 14 March 2013

Dear Colleagues, This is paper on the possible nature and standing of "intuitive" assessment in both normative and non-normative domains. It has the form, in part, of a commentary on some recent, influential work by psychologists and experimental philosophers on intuitive judgment, along with a coda on some well some well-known examples from moral philosophy. I hope that this paper will raise issues of interest to those who work in law and social policy, and who might wonder about the role of intuitive assessment in creation, evaluation, and application of law and policy, or in moral assessment more broadly. Looking forward to the opportunity to discuss these questions with you, PR

The Affective Dog and Its Rational Tale[1]

Peter Railton (Philosophy, University of Michigan)

Psychological research and experimental philosophy have recently given new life to the idea that *intuition* plays a central role in moral judgment—though perhaps not in a way that gives much comfort to the Intuitionist tradition in moral philosophy. In Jonathan Haidt's influential "social intuitionist model", for example, intuitive moral judgments are seen as fast, automatic, affective responses, rather than self-evident cognitions of a rational faculty (Haidt, 2001; compare Ross, 1930, 29-31). On this model, although reasoning can enter into the formation and revision of intuitions, the typical role of reasoning in moral judgment is *post hoc*, as we attempt to find—for ourselves or others—some rationale for our immediate emotional reactions. In the typical case, the "emotional dog" wags the "rational tail".

If this is true, does it undermine the credibility or normative authority of our moral intuitions? For those who conceive of emotion as non-cognitive or arational, it certainly could. But there is another possibility. Perhaps understanding the primary role of affect in spontaneous moral assessments helps us to see just what sort of information they might be carrying, and thus when they might be less or *more* credible. Although the affective system has been characterized in the recent literature on moral psychology as "ancient", "basic", "heuristic-based", "point-and-shoot", or "button-pushing" in its responses, increasing evidence suggests instead that the affective system is a flexible, sophisticated learning system whose chief function is to attune individuals' attitudes and behavior to the complex evaluative and social

[1] Under review at *Ethics*. Please feel free to cite, but do not circulate without permission. Thanks.

landscape they face—its prospects and perils. This system, it appears, is critical for learning in general, and plays a large role in our ability to respond aptly to a wide array of reasons.

The Affective Dog and Its Rational Tale

Peter Railton, University of Michigan

Glendower: I can call spirits from the *vasty deep*.

Hotspur: Why, so can I, or so can any man; But will they come when you do call for them?

—Shakespeare, *Henry IV*, Pt. 1

(1) Introduction. Many a philosopher has called intuitions from the depths of Plato's cave or the human psyche, but those that have come have been less impressive than one might have hoped. They have been fickle (notoriously sensitive to language and context), unreliable (18th-19th century intuitions that, necessarily, physical space is Euclidean and laws of nature deterministic), quarrelsome (Classical and Intuitionist mathematicians each find their own view of appropriate standards of proof self-evident), and provincial (a recent study of small-scale societies found that people's intuitions about fairness vary as a function of whether their predominant mode of hunting or farming is collective or individualist, see Gintis *et al.*, 2005).

Still, it is difficult to see how we can manage to avoid appealing to intuition—in thought generally and moral thought especially. Nearly everyone will agree that chronic, debilitating pain is, other things equal, a bad thing, to be avoided or alleviated. Yet what is this if not an intuition? Certainly it seems more obvious and compelling than any argument we might give for it. Similarly for the intuitive thought that, other things equal, actively causing a harm is morally worse than passively allowing the same harm to occur.

Can we, suitably chastened by the misadventures of philosophers who have called upon intuitions in the past, still make a case for attributing to intuition some *prima facie* rational or moral authority?

(2) Intuition and intuitions. The term 'intuition' has no proprietary sense, and while intuition was conceived by Intuitionists as a distinctive, quasi-perceptual rational faculty, the actual phenomenon of intuition is quite diverse and gives little evidence on its surface about its origins. What we observe is that people often find that they have (i) spontaneous notions of what is right or wrong, true or false, credible or implausible, promising or preposterous, reasonable or mad, which (ii) they find in some degree compelling or motivating and (iii) are reluctant to give up or ignore, even though they cannot articulate a satisfactory—or, indeed, often *any*—further explanation or justification.

While it is customary among philosophers to speak of intuitive *judgments*, intuition often, perhaps paradigmatically, initially appears in the form of a "sense", "feeling", or "hunch" prior to judgment. This "sense" can then shape judgment and belief, but sometimes it will run precisely contrary to what is explicitly judged or believed. A convinced functionalist in the philosophy of mind, for example, may retain an unaccountable sense that a machine perfectly replicating the functional patterns of the human brain would simply lack consciousness; a strict Kantian may be unable to avoid thinking that in some cases telling the truth, though a perfect duty, would simply be unconscionable.

(3) An example. Intuition, moreover, has the further feature that (iv) it appears capable of guiding extended, structured, spontaneous action. We can act as well as feel or think intuitively. Consider an experienced trial lawyer who has an intuition about how well her case

is going that is quite distinct from her best professional judgment. "Any rational person observing this trial would say that everything is going our way," she tells her assistant, "But I have this sinking feeling that we're about to lose. Somehow, I just can't shake it."

An intuition of this kind can emerge over time—unnoticed at first, then vaguely felt, then increasingly intense and pointed. Despite the lack of any evident justification, it can feel as if it makes a claim upon one. Our lawyer's "gut feeling" is more than a queasy stomach. It has, for example, an intentional object, "I can't say what it is, but there's something going on in that courtroom that just feels wrong." And it issues in further feelings and thoughts focused on that object—as the sense becomes more insistent, so, too, does her restless feeling that she *must* to do something, moving her to rack her brains rehearsing each day's events trying to figure out where the sinking feeling comes from.

Imagine now that the last day of the trial has come, and she still has no more definite idea of the source of her unease. With so little to go on, she resolves to stick to her tried-and-true formula in the summation: revisit the evidence step-by-step, making the logic of her conclusion seem inescapable. "Win their minds and you will win their hearts," she has always said. She thinks, "It'd be crazy to try something completely new with my client's future at stake. And besides, I have no idea what it would it be." But a third of the way into her summation, hearing her own voice echoing in the large courtroom and feeling a strange reluctance to look the jurors in the eyes as she marches through the evidence, finally, she loses heart altogether and can't go on. She feels at a complete loss. She tries restarting. "But before I continue with the evidence, let me remind you of the details of the charge brought against the defendant …". This feels wooden, preachy, hopeless. Her throat is going dry and she's beginning to flush. It feels as if everyone must know that she's struggling, lost. So once the charge repeated, she

breaks off again. With no new idea, she tries to pull herself together, buying time by straightening her tensed body, taking a long breath and walking slowly over to the jury box. As she does so, she forces herself to try to meet the jurors' eyes. Focusing her energy as best she can, she resumes speaking, though no longer in her courtroom voice, "… But I know that you know the facts of this case backwards and forwards. So what is there left to say?" An awkward pause, her mind racing. More words come, "What's left is talk about what this case is *really* about—where justice lies, and why it matters."

The jurors, too, are uneasy, evading her eyes. She is in unfamiliar territory, straining to keep her concentration and composure. Having taken up the thread, she follows it, sentence by sentence, not stopping to think what lies ahead. She moves her eyes from face to face as slowly as she dares—talking about the case as if talking to friend late at night, her taught face softening. At last she begins to feel she is making headway. One by one, the jurors stop staring past her or looking at their feet, and begin to fix upon her and what she's saying, as she explains what moves *her* about this case. The eyes of a juror in the front row are starting to tear up, and she's beginning to feel her own emotion rise in her throat. She is following where spontaneous thought and feeling lead her, now with a clear sense that she is building toward the conclusion. But what *is* the conclusion?—She has no ringing phrase at the ready. Somehow, the words come. Barely controlling her voice she says, "I have spoken to you from my heart. And I hope that I have reached your hearts. Because *that* is where you must search to find justice in this case. I know you will. Thank you. The defense rests." The courtroom is dead silent. Her legs are shaking beneath her as she returns to her seat, but she feels that her work is done.

Later she learns indirectly that her trademark meticulous method had from the outset rubbed certain influential jurors the wrong way. She had come across as indifferent and remote, while the prosecutor, legally outmaneuvered at every turn, nonetheless seemed to care strongly about the crime and its victim. Pressed afterwards by a colleague to say what tipped her off and led her to depart from her prepared summation and show her true feelings, she still draws a blank. "All I know is, at that moment I felt like I was *dying* out there—talking to myself with no one listening. And I just couldn't go on. I *had* to do something. So I tried to start over—but that was worse. I was clutching at straws, and everybody knew it. But somehow, when I slowed down and concentrated on their faces, trying to speak *to* them, it began to come. My brain was going full tilt, but I just tried to make sure I kept eye contact. Once I got into it, it began to feel like, 'Yes—keep going.' Don't ask me why."

(4) Intuitive guidance. Although she did not act on the basis of deliberation—indeed, she acted in violation her explicit deliberative resolution—our lawyer certainly wasn't sleep-walking, talking in a trance, or acting from habit. On the contrary, she had snapped *out* of habit, and her conscious and unconscious mind were working flat out, wholly focused and drawing upon all her resources, intellectual and emotional. Over the final days of the trial, it seems, she had implicitly noticed certain signs—facial expressions or body language, perhaps—that she was losing touch with jurors, so that when she gave free rein to "intuition", her response was targeted to overcoming just that alienation. Like a mature musician improvising, she was focused rather than diffuse, perceptually acute, sensitive to crucial features of the situation, inventive, and spontaneously expressive of her distinctive personality. One is never more alive, or less an automaton, than at such moments.

Nonetheless, the success of her last-minute improvisation should puzzle us somewhat. With no contingency plan to fall back on, what could account for the effective coordination of all her faculties in shaping her response: how she focused her gaze, what thoughts occurred to her, how her own feelings evolved, how she was moved to act, and what she said? Of course, it didn't happen *instantly*, as if by reflex. Her first attempt to start over failed miserably. But somehow, as she refocused on the jurors themselves, all of her faculties started to work together toward her dominant aim. And it began to feel like headway, encouraging her to continue. The groundwork for this spontaneous, coordinated response was no doubt laid, unawares, over the course of the trial, but also over years of lawyering and learning to get along with others—a sensitivity to what others might be feeling or thinking, an implicit knowledge that sharing one's own feelings or showing one's own vulnerability can enable one to reconnect with someone who has become disaffected, an acquired distaste for preachiness in herself and others. Like a skilled athlete whose years of experience, and implicit learning of her opponent's strengths and weaknesses over the course of a game, enable her to improvise just the right move in the final seconds, our lawyer was able to make an unpremeditated but unconsciously prepared and highly effective response in the clutch.

The undramatic intuitions of everyday life—what seems like a good idea, feels "off", appears likely, or looks dicey, even if we would be hard-pressed to say exactly why—similarly evolve over time and enable us to adapt with more or less intelligence and success to changing circumstances and possibilities, typically without requiring much self-aware, explicit deliberation. Given the volume of sensory information we take in at each moment, and the varied and sometimes conflicting goals, social expectations, and relationships that are in play from one moment to the next in daily life, it perhaps should not be surprising that the conscious,

deliberate mind does not manage everything on its own, but rather draws upon wider and deeper perceptual and mental resources.

(5) Normative force. Henry Sidgwick observed that intuitions tend to present themselves with a certain normative authority, and are seldom present without some "impulse or motive" to act accordingly (Sidgwick, 1907/1981, 34). If, as in the case of the lawyer or athlete, an "intuitive sense" provides this impulse and shapes action, prior to any judgment or distinguished normative attitude, what does it mean to say that it "presents itself" as making a claim to our attention or having some degree of action-guiding authority?

Well, what is it like when you are writing a condolence letter to a friend and finding all the usual phrases empty and repellant, when a certain phrase or image comes to mind and "feels" as if it belongs there? Or when the look on a friend's face when you are about to pass her on the sidewalk "tells you" to stop a bit and talk? Or when the look on your neighbor's face when you see him in a downtown club "tells you" not to? Though no explicitly normative judgments or concepts are in play, the felt authority and motivating force of such experiences is unmistakable—though I am quite certain I could not adequately convey the character of such experience to someone who had never felt anything like it.

(6) Aristotelian evaluative perception. In such experiences, the emotional meaning or evaluative significance of one's situation can seem integral to perception itself. The lawyer experiences the courtroom as "dead" to her words, or her first attempt to restart as "wooden" and "preachy", or your experience your friend's or neighbor's facial expression as "calling out" to you to stop or look away. In each case there is a distinctive phenomenology that involves more than a bare representation. We are used to thinking of the geometry of

three-dimensional objects in space as part of the "Gestalt" of visual experience, as part of the "given". In the lawyer's experience of the courtroom's "deadness", or your experience of your friend's or neighbor's face as "calling out" to you, a negative or positive "charge" inhabits the experience, a "felt quality" that makes some features salient and some thoughts and feelings immediately available.

Aristotle seems to have had a similar view of the evaluative character of perception in a well-prepared mind. A trained and experienced warrior "sees" a situation as calling for a certain response—and motivation and action "straightway" follow suit. How is this possible? He asks us to consider the working of courage. A courageous warrior is not someone who feels no fear (such a person is heedless to risk, a danger to himself and his fellow warriors), or who is like the coward (such a person feels excessive fear and will not stand his ground even when this is essential), but someone in whom the situation perceptually elicits an amount of fear proportional to the risk involved, which galvanizes attention, thought, bodily arousal, motivation, and action in risk-appropriate ways. According to Aristotle, "virtue is to do with feelings and action" (*NE* 1109b), so that training for virtue centrally involves the cultivation of *well-attuned* affect: "the courageous person is the one who endures and fears—and likewise is confident about—the right things, for the right reason, in the right way, and at the right time" (*NE* 1115b). Through training and experience, "the courageous person feels and acts in accordance with the merits of the case", so that, even though he acts *through* feeling—through a spontaneous, affectively-charged, evaluative perception of what the situation calls for—he acts "as reason requires" (*NE* 1115b), that is, in accord with the risks, goals, and values at stake. This notion of "accord", like the notion of proportionality in the Doctrine of the Mean in

general, suggests the idea that proper attunement in feelings tends to effect proper attunement in attitude and action.

(7) Attunement and warranted attitudes. In such a case, we speak of the courageous individual's fear as a *warranted attitude.* Warrant, here, involves a *double attunement*: in a properly-trained and experienced individual, the fear attitude (a) *is attuned in strength and content* to the evidence of risk, its source and magnitude, and (b) *has the effect of directly attuning her dispositions* to think, feel, and act in risk-appropriate.

The case of *belief*, an attitude often seen as warranted, helps us see this. Why is a degree of *belief* that *p* a warranted response to perceptual experience as of *p*, other things equal? Why not a *pretense* that *p* or *idle thought* that *p* or *denial* that *p*? Well, consider how belief differs from these other attitudes in the role it plays in our psychic economy. To believe that *p* is, among other things, to be disposed to think and act as if *p* actually obtains—to take *p* to be true, to expect that *p*, and to rely upon *p* in reasoning and action. These are thought- and action-guiding roles that perceptual experience as of *p* speaks in favor of, other things equal. The same perceptual experience does not speak in favor of pretense or idle thought or denial that *p* because sensory evidence that *p* does not bear favorably upon—afford a reason for—*their* respective thought- and action-guiding roles, and so they are not warranted responses to it.

To fear that *p*—as opposed, say, to *idly thinking* that *p*, or *liking* that *p*, or *being amused* that *p*—is to be disposed to act in the world as if *p* posed a risk of harm, according to *p* a negative expectation, heightening vigilance toward *p*, focusing attention and thought on trying to avoid or overcome *p*, and readying corresponding action. Evidence that *p* threatens harm warrants

an attitude with this psychic profile, just as evidence that *p* obtains warrants an attitude with belief's psychic profile.

(8) Intuitions and reasons. On this picture, an attitude on my part can be a warranted response to reasons—reasons for belief, fear, amusement, etc.—without my *judging* that I have these, or any, reasons to have it. A crawling infant's fear of the neighbor's unamiable cat, and belief that it is fast approaching, can arise from experience, be fully warranted, and appropriately guide his urgent, if clumsy, attempt to find a way to escape, without the child being able to construct a higher-order representation of these attitudes or their appropriateness. Such spontaneous thought- and action-guiding attitudes thus can qualify under conditions (i)-(iv) as "intuitive".

Does this prevent us from seeing them as forms of *knowledge* or as aptly *responsive to reasons*? It is hardly controversial that everyday perceptual belief is a form of direct, "non-inferential" knowledge. Should we treat fear differently? Not if Aristotle is our model, surely, since part of his point about the role of feeling in virtue is that the virtuous man—knower and responder to reasons *par excellence*—need not engage in the sort of deliberate, step-wise weighing of reasons or consultation of rules found when the merely continent individual struggles to figure out and do what's best. Indeed, Aristotle expresses skepticism about whether rule-based reasoning of this sort, even if assiduously carried out, could be fully and aptly responsive to all of the reasons at stake in a given situation, or sufficient to yield wisdom in the practical sphere. "This is why a young person is not fitted to hear lectures on political science, since our discussions begin from and concern the actions of life, and of these he has no experience" (*NE* 1095a). Practical reasoning itself begins not with a rule or norm, but with an end or desire; and it concludes not in a normative judgment ("I *ought* to *F*") but rather in "the

beginning of the action" (*DA* 433a). Otherwise, the conclusion would not genuinely be practical, and we would need yet another, properly practical faculty to reason our way to action.

This point can be expanded. Suppose that no thought process could count as aptly responsive to reasons unless it involved a conceptual recognition of this responsiveness. Then we would have introduced a new step, a "middle term", into the process, bringing with it a demand for recognition of the reason for it, or of the reason it affords. Regress threatens. To avoid this, we should allow that properly attuned mental processes—of perceiving, believing, fearing, or inferring—can effect direct transitions of thought that count as apt responses to reasons even though they are not, and sometimes could not be, themselves be spelled out or grounded deliberatively. Here, then, is a perfectly general and indispensable job for "intuition" in sense (i)-(iv)—not simply as an alternative to explicit deliberation, but also as necessary for it. Interestingly, it also has a venerable pedigree, connecting with Aristotle's claim that "demonstration cannot be the originative source of demonstration", which must instead be a non-demonstrative "intuition" or "comprehension" (*PA* 100b), and with Kant's appeal to "intuition" in theoretical reason since "judgment cannot always be given yet another rule by which to direct its subsumption [of a case under a rule] (for this would go on to infinity)" (*CS* 8:275) and to the "moral feeling" in the practical sphere, as an "antecedent predispositio[n] on the side of *feeling*" essential to avoid a regress of "obligation to duties" in virtuous action (*MM* 6:399, 402).

Another way to put these points is to say that reliance upon spontaneous intuition at some level is found wherever there is thinking or acting for reasons. What distinguishes the virtuous individual is simply how fluently and reliably his spontaneous intuitions are attuned to the reasons at hand, and how inventively and seamlessly they translate into action.

(9) Affect and evaluation. Why would *affect*—"feeling"—be the natural place for Aristotle to look for an account of spontaneous, fluent attunement to reasons in the practical sphere, or for Kant to look for an account of the direct "mental attunement" to value found in the "moral feeling" (*CJ* 5:267-268)? Affective attitudes, such as fear, or its complement, confidence, are suited to play the role of attuning us to reasons for thought and action because of their direct sensitivity to experience and their capacity to translate this sensitivity directly to attention, cognition, motivation, and action. Psychologists unembarrassedly speak of perceptually-based affect as *evaluation* of one's situation, because it plays to role in adjusting one's thought and action to the situation that evaluation would play. The pioneering psychologist Robert Zajonc, for example, offers this characterization of the human affective system:

> The capacity for emotional reaction, thus, is *the capacity to discriminate between and respond adaptively to present and anticipated conditions that are likely to be harmful or beneficial to the individual or his/her community*. [1998, 592]

Emotions present the world to us in quasi-evaluative terms. Liking and loving, hoping and trusting, respecting and admiring, are positive in valence, directly inducing attraction, approach, acceptance, and credence. Fear and distrust, anger and disgust, hatred and contempt, are negative in valence, directly inducing aversion, avoidance, rejection, and disbelief. Emotions entrain an immediate yet *integrated* response to experience. Some, like fear, surprise, and anger, involve arousal, shifting our psyche from "business as usual" to heightened and refocused activity. These tend, when consciously felt, to have a distinctive phenomenology. Others are unaroused, "default" emotions, like confidence, interest, and satisfaction, which underwrite everyday living and learning by encouraging us to carry on in the world, largely trusting our senses and companions, learning as we go, and continuing to pursue our goals. Default

emotions tend to have a "thin", non-distracting phenomenology. They are noticeable primarily when they fail—as when a depressive loses confidence to a point that virtually paralyzes action, or when our lawyer found her usual confidence in her manner of conducting a legal defense undermined, leaving her with a restless, anxious, urgency about her situation, which eventually caused her to lose heart and goaded her to change.

It is easy enough to imagine an evolutionary story behind these and other basic emotions, along Zajonc's adaptivist lines, and it certainly seems that many of our basic emotions can be found not only in articulate adult humans, but in very young children and our nearest animal kin. This would help account for why, like the lawyer's initial fear, they can be present but not registered as such in self-conscious thought, and why, like her eventual profound loss of confidence, which certainly *did* register, they can be "recalcitrant" (D'Arms & Jacobson, 2003) to higher-order judgment.

Proper attunement of these two forms of affect—confidence and fear—is as essential to daily life as courage is to the warrior on the battlefield. Excessive confidence can render us foolhardy, socially insensitive, and oblivious to evidence and to the feelings of others, while excessive fear can render us withdrawn, irresolute, suspicious, defensive, and mean-spirited. Out-of-tune affective responses often are a much greater obstacle to successful navigation of the physical and social world than poor sense perception or physical coordination.

(10) Affective primacy. We are finally in a position to connect our discussion of "intuition" and affect with recent developments in empirical moral psychology. As late as 1992 a survey of 20th century ethical theory could conclude by bemoaning the lack of engagement of moral philosophy with serious empirical research (Darwall *et al.*, 1992). Today, though many moral

philosophers continue to question the relevance of empirical research, it cannot be denied that such research in general, and "Experimental Philosophy" in particular, are making themselves felt.

Psychology itself has been undergoing a series of yet more dramatic changes. In 2007, the psychologist Jonathan Haidt announced the beginning of a "new synthesis" of ethics and science, writing:

> … the key factor that catalyzed the new synthesis was the "affective revolution" of the 1980's… . … [S]ocial psychologists have increasingly embraced a version of the "affective primacy" principle … [in light of] evidence that the human mind is composed of an ancient, automatic, and very fast affective system and a phylogenetically newer, slower, and motivationally weaker cognitive system. … [The] basic point was that brains are always and automatically evaluating everything they perceive, and that higher-level thinking is preceded, permeated, and influenced by affective reactions (simple feelings of like and dislike) which push us gently (or not so gently) toward approach or avoidance. [Haidt 2007, 998]

A "dual-process" model of the mind has emerged in which perceptual signals entering the brain first pass through specialized sensory cortices, which supply some structure to the percepts (e.g., constructing the "edges" and three-dimensionality of objects, or the constancy of objects through change in motion or perspective), but then the perceptual stream immediately interacts with "System 1", the areas of the brain specialized for learning, encoding, and experiencing reward and affect. Crucially, this is *prior* to the time the information reaches "System 2", higher-order, declarative thought. As a result, an affective response to new perceptual

information is already underway before self-conscious thought and reasoning come into play. This affective response is capable both of directly potentiating an array of rapid, "implicit", or "automatic" behavioral responses that need not draw on higher cognition—Haidt mentions approach and avoidance—and also of directly influencing what is salient to high-order, conscious cognition, and how it will be received or interpreted.

The architecture of "affective primacy" makes sense because it is the regions of "System 1" above all that keep track of the evaluative information associated with stimuli—for example, whether a given pattern in perception is associated with particular harms or benefits to the organism, and with what frequency or latency. Rather than process all incoming information neutrally, and then ask what importance or bearing or urgency it might have—a hugely inefficient plan—the brain within the first half-second appears on the strength of learned expectations to triage incoming information as good news, bad news, or no news, associating with it a corresponding valence and priming relevant memory, inference, and motivation. Something like this may account for why a parent can turn and scoop up a child the moment a dog snaps, "even before I had a chance to think about it". And at the same time, it may help account for why certain other situations *do* cause us to stop in our tracks and think, even before we had a chance to stop and think about whether to do so.

Because affect is the brain's principal way of representing value, "affective primacy" in perception suggests a natural psychological model for Aristotelian "evaluative perception". First, it proposes a neural architecture in which evaluative information is incorporated into the perceptual process itself; second, it suggests how an individual's "way of seeing things" could directly yield action without need for reasoning or additional motivation; and third, it coheres well with Aristotle's view that learning how to act virtuously requires above all the training of

perception and feeling. We can see here as well the beginnings of an understanding of how complex "intuitive" action, such as our trail lawyer's improvised summation, might be possible.

(11) Affect, representation, and information. Of course, to have satisfactory explanations of this kind we would need to be able see how "System 1" could exhibit *intelligent, structured* action-guidance—guidance that is adequately sensitive to relevant evidence, able to represent the diverse values or costs at stake, and capable of synthesizing this information to guide novel, often complex, behavior appropriately. A tall order for an "ancient, automatic, and very fast system" (Haidt, 2007), which has been variously characterized as "hard-wired", "quick and dirty", "button-pushing", and "point-and-shoot".

Any chance? Surprisingly, yes. The last two decades of research has provided a picture of the affective system as a perceptually-sensitive, cognitively-rich information-processing system in its own right (Schwartz & Clore, 1983, 2007). It appears to form the neural substrate for the principal forms of experience-based learning, houses core long-term memory, and is massively interconnected with—in many cases, functionally integral with—regions of the brain directly responsible for coordinated action, goal-pursuit, weighing costs and benefits, and decision-making (Pessoa, 2008). In those individuals who suffer affective disorders, such as depression or mania, or who experience damage to interconnections between affective regions and the higher cortices, deliberation and decision-making is often *less* rather than more rational, consistent, or effective (Barg & Chartrand, 1999).

Perhaps the clearest evidence of the affective system's information-processing capacities has come in the last decade, during which an increasingly wide array of techniques for assessing brain function have been brought to bear on the affective system, ranging from behavioral tasks

and lesion studies to functional magnetic resonance imagining, single neuron electrode recording, and microinjections to disrupt or stimulate particular chemical pathways. The principal subjects for this research have been intelligent animals, including primates, whose affective architecture we largely inherit. But there has been no shortage of behavioral and neuroscientific studies of humans as well.

We might begin with the evolutionary ecologists' reminder that all animals, our ancestors included, "run on batteries". This means that they underwent eons of selection for efficiency and effectiveness in the main tasks of life. Searching for food, for example, becomes greatly more efficient if the brain can develop an accurate representation of its environment, the types and quality of available foods, the varied effort required to obtain them, the reliability with which they are present, and the risk of predation. So effective are foraging animals such as birds and mammals at learning about their environment, and forming and following such representations, that they tend over time to exhibit near-optimal foraging behavior—solving through reinforcement learning a complex linear programming problem bringing together energy gain, energy expenditure, nutritive needs, and risks. Similar effectiveness can be found in various forms of social behavior, including group formation and dispersal, cooperation in hunting, mate searching, cheating, and shared responses to predation (Dugatkin, 2004).

Very impressive. But why speak of "representation" of this information rather than simple trained behaviors? It would be preposterous to imagine that a foraging rat, for instance, carries about in its tiny brain an elaborate map of its environment, bearing detailed cost/benefit information, suitably discounted by risk. Rats are stimulus-bound creatures that learn by conditioning and act through instinctual or habitual responses.

Except that they aren't—and what seems preposterous is true. Rats possess dedicated "place" neurons that build up map-like representations of their immediate environment, as well as a sophisticated learning system, centered in affective regions of the brain, that keeps careful track of successes and failures in all the key areas of rat life, gains and losses, variations and probabilities, continually computing newly revised expectations that guide behavior while remaining sensitive to subsequent discrepancies between expectation and outcome that induce further, discrepancy-reducing revisions of expectations.

Instead of rigid, conditioned motor habits, we find experience-sensitive systems of neurons whose functional organization, selective activation, and firing rates separately reflect: the identity and magnitude of varied potential benefits; the relative value associated with a stimulus (e.g., the changing value of food in relation to satiation) as well as the absolute value (e.g., its relation to recurrent needs); the probability or expectation of a given positive or negative outcome; the pairing of outcomes with potential behaviors; the occurrence of a better- or worse-than-expected outcomes; and the expected value (probability **x** value) and absolute risk associated with an outcome (see, among many others, Schultz, 2002; Tobler, *et al.* 2006; Preuschoff *et al.* 2006; Berridge & Aldridge, 2008; Rolls *et al.*, 2008; Craig, 2009; Kringelbach and Berridge, 2009; Singer *et al.* 2009; Grabenhorst and Rolls, 2011). These neural systems, classified as parts of the affective system and its cortical interfaces, are found in more developed forms, and appear to function similar ways, in the modern human brain (Quartz, 2007).

Recent experiments with rats learning mazes not only shows them to build up a mental map of the maze as they explore it, but also to rehearse this map through repeated brain activation during REM sleep, consolidating memory and improving performance (Louie &

Wilson, 2001). And as the rat passes through the maze, brain activation will spread down the alternative paths in this map *ahead* of its actual location, prospecting the different levels of value associated with each choice-point, and guiding behavior accordingly (Johnson & Redish, 2007). It is this capacity to represent the geometry and rewards of their world that explains the surprising early observation of the great experimentalist, Karl Lashley, that rats escaping from the start box in his T-shaped maze quickly scampered *diagonally* across to the top of the maze, directly to the food location (Lashley, 1929). Rats, it emerges, are more predictable if we can see what evaluative representations they hold of their world and its prospects.

How do we square this large body of research with the fact that humans are notoriously bad at assessing conditional probabilities and calculating the expected value of outcomes or actions? In "System 2" declarative reasoning, we are indeed easily led astray by effects such as "framing", "anchoring", and "representativeness" (Kahneman & Tversky, 2000). But this does not measure the representational capacity of "System 1", and a number of lines of experimentation suggest that, in a variety of complex choices, humans do better when relying upon "intuitive" rather than calculated decisions (Bargh & Chartrand, 1999). A series of experiments by Dijksterhuis, *et al.* (2006) suggests that when decisions involve more than three or four dimensions of variation, a forced, "intuitive" decision is better able to keep track of all the dimensions than considered, deliberate choice, and more likely to yield a choice with which one is satisfied. And our deliberative reasoning and decision-making notably improve when the affective system can be brought to bear "analytic reasoning cannot be effective unless it is guided by emotion and affect" (Slovic *et al.*, 2004).

(12) Social learning, simulation, and empathy. Can the affective system attune us to more complex social values, such as those at stake in our lawyer's situation? Empathy—

understood as a capacity to imagine and understand the attitudes of others "as from their point of view"—has recently taken a central place in the literature on social cognition, moral development, and individual prudence. The affective system appears to play a crucial role underwriting this capacity. Experiencing a mild shock, anticipating the arrival of such a shock, watching another undergo such a shock, and imagining inflicting such a shock upon another appear to activate extensively overlapping regions of the affective system (Ruby & Decety, 2001; Decety & Ickes, 2009). A leading hypothesis is that we tend "automatically"—without need for conscious attention—to *simulate* the mental states of those around us and as well as our own possible future states, deploying our affective system as a "test bed". On this account, the hostility we "see" in the face of another, for example, which we tend to find aversive and alarming, is registered in us by activating an internal simulation of the angry attitude behind it—in part, perhaps, owing to a kind motor mimicry (Niedenthal, 2007). The negative and disturbing character of the simulated attitude then informs our response to the angry individual. We do not simply "mirror" the attitude of the other, becoming angry ourselves, because the simulation is "off line" rather than "on line", a difference that appears to be neurally correlated with self- vs. other-representation in general (Ruby & Decety, 2001).

Given the thorough-goingly and often tightly social conditions under which humans and their immediate ancestors have come into the world, grown to maturity, lived, and reproduced, it would not be surprising if we possessed a capacity to be attuned "automatically" to the thoughts, feelings, attention, and likely behavior of those around us—no more requiring conscious deliberation than the perception of physical objects.

Empathic simulation is thought to help us compete as well as cooperate effectively, since it is capable of representing accurately attitudes quite at odds with our own current state or self-

conscious construal of the situation. What we call "good social intuition" and "good self-protective instincts" require this ability, and those with profound deficits in empathic capacities find such "intuitive" social or prudential knowledge exceptionally difficult to acquire or act upon effectively (Baron-Cohen *et al.*, 1997; Decety & Ickes, 2009).

In the case of the trial lawyer, empathy as unconscious, automatic simulation would provide a plausible mechanism for explaining how information about the feelings of the jurors, so unrelated to her own view of how the case was going, could begin to make itself felt, unsettling her. Accurate "off line" empathic simulation on her part could also make salient the specific character of jurors' greatest need—to *reconnect* with her emotionally—so that her reliance upon "intuition", once begun, could be effective. Sensitivity to moral considerations, similarly, often involves seeing a situation from others' points of view. Given the inevitable influence of our own viewpoint and interpretation of events upon self-conscious moral deliberation, automatic empathic simulation, which influences thought, feeling, and behavior through non-deliberative means, affords a valuable counterweight.

The theories of moral development that dominated the early days of cognitive psychology emphasized levels of abstraction and generality in moral reasoning (Kohlberg, 1971), but these measures failed to show a systematic connection to differences in moral conduct. A capacity to empathize, and activation or inhibition of this capacity in a given instance, appear by contrast to be important predictors of moral behavior (Hoffman, 2001; Baron-Cohen, 2011).

(13) Summing up so far. The fact that, as Haidt (2007) emphasizes, our affective system is "ancient" and "automatic" and "fast" should not be taken as ground for thinking that it affords only a crude means of evaluating incoming information, limited to immediate likes and dislikes

or to a few hard-wired basic emotional "buttons", or dominated by simple "heuristics", or confined to a few rudimentary "scripts". Although the intuitive operation of the affective system is grounded in "innate skills that we share with other animals", as Daniel Kahneman puts it, it does not follow that this system is purely qualitative in operation or "has little understanding of logic and statistics" (2011, pp. 21, 24). Statistical learning systems, such as the affective system, have strengths and weaknesses. Fortunately, humans possess self-conscious reasoning processes that can offset some of these weaknesses—most importantly, self-conscious reasoning permits self-representation of one's category system, and thus questioning and revising it. But one weakness is not inherent lack of subtlety, or incapacity to learn and change, or inability to represent and process information analytically and in depth, or incapacity to balance costs and benefits.

The account I have sketched of how "moral intuition" might be grounded in our affective system and empathic capacities suggests that these assessments involve "domain general", widely-distributed capacities for perceiving, evaluating, learning, and acting—rather than, say, a dedicated "moral faculty". Accumulating evidence supports this suggestion, though some have drawn a conclusion opposite to the one I would urge. Amitai Shenhav and Joshua Greene (2010), have provided important experimental evidence indicating that intuitive moral assessments deploy "domain general" regions of the brain and are highly sensitive to variations in the magnitude and probability of outcomes. They conclude:

> If, as the present results suggest, the neural mechanisms we use to think about complex, life-and-death moral decisions are in fact mechanisms originally adapted primarily for other purposes (e.g., foraging for food), then it becomes more likely that such decisions are made suboptimally. [Shenhav & Greene, 2010, 674]

I would argue that, precisely because these "other purposes" have for eons involved "complex, life-and-death" decisions, with both individual and social dimensions, intuitive moral assessments emerging from such domain general capacities may reflect a great deal of important, contextually-nuanced information and implicit social understanding relevant to moral evaluation. Knowing more about psychic processes, and their strengths and limitations, will help us in a "broad reflective equilibrium" assessment of how much confidence, if any, we should place in particular moral intuitions. As no more than a philosophical observer, the general lesson I draw is that recent decades of research suggest how much potential for attunement to reasons lies in "intuition". That is good news, if, as argued above, properly-attuned intuition is needed for the full range of thought, demonstrative as well as non-demonstrative.

With this cheerful, if overly-simple and overly-optimistic preliminary conclusion in mind, let us, then, quickly revisit some of the well-known examples in the empirical literature on moral intuition.

(15) Julie and Mark. Let us suppose now that our trial attorney is asked what she makes of the following scenario (from Haidt, 2001):

> *Julie and Mark are brother and sister. They are traveling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At very least it would be a new experience for each of them.*

How might this information be processed by her "ancient" and "automatic" affective system? Given her life experience—she is, let us suppose, not just an attorney who has seen cases of incest, but also an active member of the community, a mother of two teenagers, and a first-

class poker player—as her mind races ahead and evaluates this input, what sort of expected value is likely being assigned to Julie and Mark's "interesting and fun" idea?

Well, her affective system might be simulating Julie and Mark's thinking and feeling as they discuss this question and reach their conclusion. It would also be simulating forward into various outcomes of this decision and their action, viewed from their perspective and the perspective of others, drawing upon her own personal and professional experience.

My guess is, by this point in the scenario, her affective system is already coding Julie and Mark's idea as highly risky, liable to hurt one or the other, or both, of them—and unlikely to yield the simple hedonic reward they appear to expect. The net expected value for acts of this kind is therefore strongly negative—they could cause considerable, possibly long-lasting harm for small gain, and hurt family and others in ways the two young people do not anticipate.

The scenario continues:

> … *Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think about that, was it OK for them to make love?* [Haidt, 2001]

Updating on this information, our lawyer's initially negative evaluative representation might soften a bit—at least, improbably, they are said to have gotten off lightly. That's a surprise, and in fact not very credible—any skilled psychiatrist or social worker would feel suspicious of this seemingly innocent and untroubled feeling of "closeness" afterwards, and its wider ramifications in their psyches. Even so, would any of this do much to change our lawyer's initial assessment

of whether the act made sense? Not really. The act was a poor, risky idea, showing them insensitive to, and insufficiently motivated by, the lasting harm they might have caused to each other. Intuitively, then, "Not OK".

Attitudes and types of behavior, not chance outcomes of individual acts, seem to be the locus of much moral thought. We might compare:

> **John**: *John dangled his baby out the sixth-floor window. The baby loved it. And John's grip, he knew, was excellent. The baby never fell. What do you think, was it OK for John to do this?*

Haidt's subjects largely thought Julie and Mark's incest "Not OK", but had trouble explaining the ground of this "intuitive" judgment. They cited information that is relevant to incest in general—psychological harm, genetic consequences, etc.—but deemed "not applicable" to this case because the description ruled them out. From our perspective, general information about incest is quite relevant to assessing the permissibility of their "interesting and fun idea"—just as general information about accidental injuries to babies is quite relevant to assessing the permissibility John's interesting and fun idea. So in this example we have a non-deliberative, "automatic" affective response from an "ancient" system of "social intuition", made on the spot, "pushing" us in the direction of rejecting Julie and Mark's idea. Does this show that the response was not in fact grounded in a sensitivity conditioned by the complex values and risks at stake?

(16) In the executive suite. A widely-studied pair of examples first discovered by Josh Knobe (2010) have suggested that people's intuitive assessments of the *causal* structure of an act are in fact confounded by their *moral* assessment of it, contradicting the "lay scientist"

picture of ordinary cognition. Let's consider Knobe's examples, but only after contemplating two similar ones:

> **Goat 1:** *You are a herder living in a remote village on a hillside. Above you is a neighbor with an orchard of olive trees. One day you are browsing the shelves in the agricultural supply store in town, when you overhear your neighbor being told by the man at the register, "Yes, you could use this spray on your trees, and it would kill all the bugs, but when the rain comes it will wash off and run down onto your neighbor's fields—poisoning her goats when they eat the grass."*

A keen interest on your part will attach to your neighbor's next words. What does he say?

> *"I don't care at all about poisoning her goats, I just want to kill those bugs. Give me the spray."*

Your neighbor is about to use a spray that he knows is likely to kill your goats. How do you now feel about your neighbor's attitude toward you? Would you characterize this as a *neutral* attitude toward you and your well-being? Live and let live? And are *you* now prepared to live and let live, or will you now be highly vigilant about your neighbor's spraying and do what you can to stop him? And can you expect others in the village to share your reaction?

Now, by contrast, consider:

> **Goat 2:** *You are a herder living in a remote village on a hillside. Above you is a neighbor with an orchard of olive trees. One day you are browsing the shelves in the agricultural supply store in town, when you overhear your neighbor being told by the man at the register, "Yes, you could use this spray on your trees, and it would kill all the bugs, but when the rain comes it will wash off and run down onto your neighbor's fields—killing the bugs that have been destroying her grass."*

In this case, too, you would listen intently for his answer:

"I don't care at all about helping her grass, I just want to kill the bugs on my olive trees. Give me the spray."

Well, how does *that* feel? Your neighbor is about to use a spray that will help your goats, but he doesn't give a damn about that. Not a great attitude—an indifference toward helping you is a problem of sorts in a neighbor. But live and let live. This is miles away from the highly problematic attitude you have your hands in the first scenario. If you were to run to others in the village to complain of this man's lack of interest in benefiting you, could you expect much sympathy? Is this so far from normative social expectations?

Now consider Knobe's two scenarios:

Boardroom I: *The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.' The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was harmed.*

Boardroom II: *The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, and it will also help the environment.' The chairman of the board answered, 'I don't care at all about helping the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was helped.* [cf. Knobe, 2006]

The result Knobe obtained, to the surprise of many, is that subjects intuitively judge that the chairman in the first scenario *intentionally* harms the environment, while the chairman in the second scenario *does not* intentionally help it. What are we to make of this?

What I hope **Goat 1** and **Goat 2** suggest is that indifference to harming someone is not a "neutral", live-and-let-live attitude, while indifference to helping someone more nearly is. Why is that? There is too much to be said on this head, but the crystallized social knowledge this asymmetry represents reflects understanding of a vital condition for living together peaceably and successfully—in a family, in a neighborhood, in a village, or in a society. The neighbor's apparent attitude in **Goat 1**, manifest in knowingly and willingly harming a neighbor—even if this was not the *aim* of his action—reveals in him an underlying attitude that is strongly at odds with living together in this way, an attitude that is *congruent* with inflicting harm. In **Goat 2** the neighbor's behavior reveals an underlying attitude that is not sharply at odds with shared social life, but also not particularly congruent with helping. This understanding of the role of underlying attitudes in explaining why people would permit themselves to make certain choices, translated from the village to the boardroom, would enable us to understand why the environmental harm in **Boardroom I** is intuitively seen as intentional in a way the environmental benefit in **Boardroom II** is not.

A recent series of experiments lends support to this idea. By considering a range of carefully matched variants of **Boardroom**-like scenarios, Chandra Sripada was able to compare moral vs. other factors influencing mental state attribution (Sripada, in press). Sripada found that a "deep self" model, in which intentionality judgments are mediated by the imputation of different implicit, underlying value structures, better predicts the variation found across

scenarios than does the hypothesis that moral judgments as such affect the structure of people's folk-psychological explanations (Knobe, 2010).

Hume more than anyone drew our attention to the importance in intuitive moral assessment of the quality of the attitudes imputed *behind* others' behavior, and P.F. Strawson more recently has revived a similar idea in a Kantian setting, developing an account of "reactive attitudes" intimately associated with notions of responsibility (1974). The asymmetries in **Goat** and **Boardroom**, if the sort of interpretation suggested here can be sustained, might be taken as lending some support to this general orientation in understanding commonsense moral assessment—rather than calling into question either the asymmetric intuitions themselves or our ordinary understanding of folk-psychological explanation.

(17) Bus and Trolley. We are all familiar with the asymmetry in intuitive moral judgment between **Switch** vs. **Footbridge** versions of the "trolley problem". Thanks to the brain-imaging work of Joshua Greene and others (Greene, *et al.*, 2001), we know something about the neural correlates of this difference—it appears that, in deciding whether to push the "fat man" in **Footbridge**, certain areas of the brain involved in affect are considerably more active than they are in deciding whether to pull the lever to divert the tram in **Switch** cases (though they are active to some degree in both cases).

Let's consider a scenario I have used with my students, before giving them either trolley problem:

> **Bus:** *You live in a city where terrorists have in recent months been suicide-bombing buses and trains. The terrorists strap explosives to themselves under their clothing, and, at busy times of the day, spot a crowded bus or train and rush aboard, triggering the bomb instantly to avoid being stopped. You are on a very crowded bus at 5:10 pm, and are struggling to get to out the door at*

your stop. The doors are starting to close and you won't be able to get off unless you jostle the slow-moving obese gentleman trying to exit at the same time. Suddenly you notice a man rushing up to bus and forcing his foot into the doorway, wedging it between the fat man and the door frame. He is reaching with one hand under his coat and a gap between the buttons reveals to you what look like explosives strapped around his chest. You can't reach this man, but if you push the corpulent gentleman beside you hard *in his direction right now, this man will fall directly on top of the seeming bomber and both will end up on the empty sidewalk, while you fall backwards into the bus as the doors snap shut.*

—So, if you push hard, and this man is not *a bomber, then the bus will leave behind two very annoyed men on the sidewalk, and you will be left on the bus, covered with embarrassment. But if he* is *a bomber, the bus will be spared, and you with it, but the fat man killed as the bomber explodes underneath him.*

—On the other hand, if you simply squeeze off the bus alongside the corpulent gentleman and do nothing more, and the other man is *a bomber, then many people on the bus will be killed while you and the corpulent gentleman are safe on the sidewalk. But if this man is* not *a bomber, then no one on the bus will be hurt and you simply will have jostled a corpulent gentleman while exiting a bus, and you can apologize to him on the sidewalk.*

Whatever happens, you will not be killed if there is a bomb and it goes off—you will either be on the bus when it explodes on the sidewalk, or on the sidewalk when it explodes on the bus.

Should you (a) shove the corpulent gentleman hard right now, or (b) squeeze off the bus, jostling the corpulent gentleman but doing nothing else?

The most recent time I tried this, and then presented students with the traditional **Switch** and **Footbridge** scenarios, I received the following results (with thanks to Warren Harold):

Bus: Push the corpulent gentleman, (a)?	67% Yes	33% No
Footbridge: Push the "fat man" off the bridge?	29% Yes	71% No
Switch: Pull the lever on the switch?	72% Yes	28 % No

What might this pattern of judgments tell us? It might tell us that morally irrelevant details influence people's judgments. Or it might tell us something rather different.

Try the following mental experiments. First, can you imagine yourself pushing the corpulent gentleman off the bus? Vividly—under what feels like life-or-death pressure, actually feeling the heft of his body as you shove with all your might, aiming as best you can squarely at the seeming bomber? Now, can you imagine yourself pushing the fat man off the bridge? Vividly—under what feels like life-or-death pressure, actually going up to this stranger, hoisting his body with all your might, and sending it over the rail, aiming as best you can to hit the track squarely? I am guessing that for many of you, the answer to the first question is that you *can* credibly simulate this imaginatively, but in the second case you *cannot*—even if you are inclined to *judge* that the fat man should be pushed. When I try to imagine pushing the man off the bridge, I can at best find myself doing it mechanically, with almost muscular reluctance. Not so in the case of the bus.

Try now some further mental experiments, involving imagining the standpoints of others, who can see and understand your situation. Can you imagine the people on the bus shouting, "Push him, for God's sake!"? Can you imagine the five track-workers seeing the trolley hurtling

toward them and shouting to you, up on the bridge, "Push him, for God's sake!"? Now try to imagine the aftermath. For each case, imagine that you have successfully pushed the weighty individual. In **Bus**, the man was indeed a bomber, but his self-explosion was blanketed by the corpulent gentleman. Both die a gory death. Can you imagine the passengers on the bus, or a bystander, yelling, "Grab the man by the door—don't let him get away!"? In **Footbridge**, the runaway streetcar was indeed brought to a gory halt, killing the "fat man" but saving the workers. Can you imagine the workers or a bystander yelling, "Grab the man on the bridge—don't let him get away?"!

And now consider the longer-term aftermath, the haunting feeling you would have, in either case, of having pushed a man to his death. You bear this as a fact of your life. In **Bus**, do you feel *shame* or *disgrace* as well as guilt? In **Footbridge**? Do you feel in either case that you can look your friends and co-workers, or strangers, in the eye while recounting this episode? Can you imagine what reaction you might expect, when moving into a new community, this episode became known? Could you in such a case expect sympathy from others, as you bear your lonely burden? And can you imagine how you would face the family of the individual you pushed, taking condolences to them, standing beside them at a funeral? You might feel forever as if you owed them something—and perhaps you do. Would you also, in either case, feel as if you deserve forever their wrath and contempt, or eventual forgiveness? And can you imagine being a member of the family, and how forgivingly you would look upon the person who pushed your loved one to his death?

If you are like me, you will find asymmetries in attitudes felt in these imaginings paralleling the original asymmetry in attitude between pushing in **Bus** vs. pushing in **Footbridge**. This suggests, if I am right, that the asymmetry at work here is relatively standpoint independent—it

has to do with an implicit bit of *shared* social knowledge, and has little to do with the situation of the person who pushes being one of "up close and personal" or "ME HURT YOU" violence (compare Greene *et al.*, 2001; Greene & Haidt, 2002), since this is present in **Bus** as well as **Footbridge**. And there is too much imaginative richness and complementarity in the attitudes that *surround* these cases for the asymmetry to be rooted in "roughly … violations a chimpanzee can appreciate" (Greene & Haidt, 2002).

I'm sure I cannot fully articulate the implicit, shared social knowledge that makes pushing someone to a gory death in **Bus** virtually indistinguishable in moral acceptability from pulling a lever in **Switch**, and so different from **Footbridge**. The fact that the passengers in **Bus** are nearly as proximate to you as the corpulent gentleman may help equalize the salience of their possible fate if one does not act, perhaps making empathy more equally distributed as well. This might make **Bus** a fairer test of "doing vs. allowing" intuitions than **Footbridge**. **Bus** seems to me to frame your dilemma in terms of "social self-defense", even though *you* are not strictly exercising self-defense, since you will escape death whether or not you push. And so I can imagine pushing in **Bus** to conform to informal norms that might spring up in a community threatened by a spate of bombings. **Footbridge** invites no such framing, though it would not strictly be impossible to see pushing in this case in that light. And it seems improbable at best that informal norms of pushing people under streetcars would emerge in a community plagued with trams with bad brakes. Penetrating further into these questions calls for something more than my armchair speculation, but my suggestion is that it might be worth doing, and could show that we are wrong to read the original trolley asymmetry in terms of morally irrelevant features or atavistic aversions, rather than pushing harder to uncover implicit social knowledge and value structures.

(18) Intuition and authority. Unlike those who have pioneered the important developments in moral psychology and experimental philosophy, I have nothing but intuition to offer. And my question has been, what should I make of it? The examples we have considered all share certain features. For people to live together with some measure of mutual trust, cooperation, and understanding, they should not decide on flimsy grounds to engage in incest with a loved one, or spray an insecticide while knowing it will destroy a neighbor's livestock and perhaps his livelihood, or push an innocent person off a bridge in cold blood to prevent an accident, or be unwilling to play their part, even if grisly, in social self-defense. These cases all involve attitudes that are in tension with our intuitive sense of being the kinds of people we want to be or live with.

Contemporary psychology and neuro-science suggest that "intuition" could be rooted in an affective and empathetic system adept at attuning us to the many and conflicting values at stake as we live our separate and shared lives. We all know certain people whose advice and counsel we especially value and trust. What is it about these people that gives them such authority for us? Is it that they hold moral principles we share? Many people do that, yet we do not turn to *them*. Indeed, the people to whom we turn often are people with whom we have some degree of difference on moral principles.

What is it, then? Might it be their *intuitive grasp* upon situations? Their attunement, probably only articulable in part, to a situation's many dimensions, the various values at stake, and the many perspectives from which it might be viewed. And their ability to synthesize this information into a "sense" of what is appropriate—a sense we ourselves could come to share.

(19) Glimpses ahead? Experimental philosophy has done a great deal, I think, for the field. It has shaken our naïve confidence in "intuitions", or in knowing what "common sense" thinks. It has begun to look seriously at the *How?* and the *Why?* of moral evaluation and action, and to hold our philosophical feet to the fire of psychological realism when we theorize.

The contributions of empirical psychology, too, are much needed. Jonathan Haidt's "Social Intuitionist" model of moral judgment, first articulated in "The Emotional Dog and Its Rational Tail" (2001), strikes me as an important step forward, since it brings together behavioral and neuro-scientific data, and tries to locate moral judgment within a wider picture of human thought and action. I have tried to suggest here that "emotional" vs. "rational" might be a misleading way to put things—as are a number of other ways now current, such as "affective" vs. "cognitive" or "automatic" vs. "analytic". Our affective system is integral to our cognition and analysis of information, and while it is "automatic" it is not reflexive. It is a core part of our capacity to respond aptly to reasons—our *rationality*, construed in the way that matters most. And that is the rational tale of the affective dog.[i]

References:

Aristotle (*DA*). *De Anima*. Trans. by W.S. Hett. Cambridge: Harvard University Press, 1936.

Aristotle (*NE*). *Nichomachean Ethics*. Trans. and ed. by Roger Crisp. Cambridge: Cambridge University Press, 2000.

Aristotle (*PA*). *Posterior Analytics*. Trans. by G.R.G. Mure. In R. McKeon, ed. *The Basic Works of Aristotle*. New York: Random House, 1941.

Bargh, J.A. & Chartrand, T. (2009). "The Unbearable Automaticity of Being". *American Psychologist* **54**: 462-479.

Baron-Cohen, S., Cosmides, L. & Tooby, J. (1997). *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge: MIT Press.

Baron-Cohen, S. (2011). *The Science of Evil: On Empathy and the Origins of Cruelty*. New York: Basic Books.

Behrens, T.E., *et al.* (2006). "Associative Learning of Social Value". *Nature* **456**: 245-249.

Beierholm, U.R. & Shams, L. (2009). "Bayesian Priors are Encoded Independently from Likelihoods in Human Multisensory Perception". *Journal of Vision* **9**.

Berridge, K.C. & Aldridge, J.W. (2008). "Decision Utility, the Brain, and Pursuit of Hedonic Goals". *Social Cognition* **26**: 621-646.

Craig, A.D. (2009). " 'How do you feel—now?' The Anterior Insula and Human Awareness". *Nature Reviews Neuroscience* **10**: 59-70.

D'Arms, J. and Jacobson, D. (2003). "The Significance of Recalcitrant Emotion (or, Anti-Quasijudgmentalism)". *Philosophy* **52** (suppl.): 127-145.

Decety, J. & Ickes, W. (2009). *The Social Neuroscience of Empathy*. Cambridge: MIT Press

Dijksterhuis, A., *et al.* (2006). "On Making the Right Choice: The Deliberation-without-Attention Effect". *Science* **311**: 1005-1007.

Dugatkin, L.A. (2004). *Principles of Animal Behavior*. New York: Norton.

Gintis, H., *et al.* (2005). *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life.* Cambridge: MIT Press.

Gobet, F., *et al.* (2001). "Chunking Mechanisms in Human Learning". *Trends in Cognitive Science* **5**: 236-243.

Grabenhorst, F. & Rolls, E.T. (2011). "Value, Pleasure, and Choice in the Ventral Prefrontal Cortex". *Trends in Cognitive Sciences* **15**: 56-67.

Greene, J., *et al.* (2001). "An fMRI Investigation of Emotional Engagement in Moral Judgment". *Science* **293**: 2105-2108.

Greene, J. & Haidt, J. (2002). "How (and Where) Does Moral Judgment Work?". *Trends in Cognitive Science* **6**: 517-523.

Haidt, J. (2001). "The Emotional Dog and Its Rational Tail". *Psychological Review* **108**: 814-834.

Haidt, J. (2007). "The New Synthesis in Moral Psychology". *Science* **316**: 998-1002.

Hare, R.D. (1993). *Without Conscience: The Disturbing World of the Psychopaths Among Us.* New York: Guilford.

Hoffman, M. (2001). *Empathy and Moral Development: Implications for Caring and Justice.* Cambridge: Cambridge University Press.

Johnson, A. & Redish, A.D. (2007). "Neural Ensembles in CA3 Transiently Encode Paths Forward of the Animal". *Journal of Neuroscience* **27**: 12176-12189.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus, and Giroux.

Kahneman, D. & Tversky, A. (2000). *Choice, Value, and Frames*. New York: Russell Sage.

Kant, Immanuel (*CS*). *On the Common Saying : That May Be Correct in Theory, but It Is of No Use in Practice*. Trans. by M.J. Gregor. Cambridge: Cambridge University Press, 1996.

Kant, Immanuel (*CJ*). *Critique of Judgment*. Trans. by W.S. Pluhar. Indianapolis: Hackett, 1987.

Kant, Immanuel (*MM*). *Metaphysics of Morals*. Trans. by M.J. Gregor. Cambridge: Cambridge University Press, 1996.

Knobe, J. (2010). "Person as Scientist, Person as Moralist". *Behavioral and Brain Sciences* **33**: 315-329.

Knobe, J. (2006). "The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology". *Philosophical Studies* **130**: 203-231.

Kohlberg, L. (1971). "From Is to Ought: How To Commit the Naturalistic Fallacy and Get Away with It in the Study of Moral Development." In T. Mischel (ed.), *Cognitive Development and Epistemology*. New York: Academic Press.

Kringlebach, M.L. & Berridge, K.C., eds. (2009). *Pleasures of the Brain*. Oxford: Oxford University Press.

Lashley, K. (1929). *Brain Mechanisms and Intelligence*. Chicago: University of Chicago Press.

Louie, K. and Wilson, M.A. (2001). "Temporally Structured Replay of Awake Hippocampal Ensemble Activity during Rapid Eye Movement Sleep". *Neuron* **29**: 145-156.

Niedenthal, P.M. (2007). "Embodying Emotion". *Science* **316**: 1002-1005.

Pessoa, L. (2008). "On the Relationship between Emotion and Cognition". *Nature Reviews Neuroscience* **9**: 148-158.

Preuschoff, *et al.* (2006). "Neural Differentiation of Expected Reward and Risk in Human Subcortical Structures". *Neuron* **51**: 381-390.

Quartz, S.R. (2007). "Reason, Emotion, and Decision-Making: Risk and Reward Computation with Feeling". *Trends in Cognitive Sciences* **13**: 209-215.

Ross, W.D. (1930). *The Right and the Good*. Oxford: Claredon.

Rolls, E.T., *et al.* (2008). "Expected Value Reward Outcome and TD Error Representations in Probabilistic Decision Task". *Cerebral Cortex* **18:** 652-663.

Ruby, P. & Decety, J. (2001). "Effect of Subjective Perspective Taking during Simulation of Agency: a PET Investigation of Agency". *Nature Neuroscience* **4**: 546-550.

Sidgwick, H. (1907/1981). *The Methods of Ethics*. 7th ed. Indianapolis: Hackett.

Schutz, W. (2002). "Getting Formal with Dopamine and Reward". *Neuron* **36**: 241-263.

Schultz, W. (2010). "Dopamine Signals for Reward Value and Risk". *Behavioral and Brain Functions* **6**:24, 1-9.

Schwartz, N. & Clore, G.L. (1983). "Mood, Misattribution, and Judgments of Well-Being: Informative and Directive Functions of Affective States". *Journal of Personality and Social Psychology* **45**: 513-523.

Schwartz, N. & Clore, G.L. (2007). "Mood as Information: Twenty Years Later". *Psychological Inquiry* **14**: 296-303.

Singer, T., *et al.* (2009). "A Common Role of Insula in Feelings, Empathy, and Uncertainty". *Trends in Cognitive Science* **13**: 334-340.

Slovic, P., Finucane, M.L., Peters, E., & MacGregor, D.G. (2007). "Risk as Analysis and Risk as Feelings: Some Thoughts about Affect, Reason, Risk, and Rationality". *Risk Analysis* **24**: 311-322.

Strawson, P.F. (1974). *Freedom and Resentment*. London: Methuen.

Sripada, C. (in press). "Mental State Attributions and the Side Effect Effect". *Journal of Experimental Social Psychology*.

Tobler, *et al.* (2006). "Reward Value Coding Distinct from Risk Attitude-Related Uncertainty Coding in Human Reward Systems". *Journal of Neurophysiology* **97**: 1621-1632.

Zajonc, R. (1998). "Emotions". In Gilbert, D.T., Fiske, S.T., and Lindzey, G., eds. *The Handbook*

of Social Psychology. Fourth edition. **II**: 591-632.

[i] I owe a debt to many individuals and groups for helping me think through the issues discussed in this paper. <acknowledgements>