

Leland Saunders
University of Maryland
Draft
30 May 2006

Reason and Intuition in the Moral Life: A Cognitivist Defense of Moral Intuitions

Abstract

It is common in moral philosophy to appeal to so-called “moral intuitions” about cases as a means of building and defending moral theories. Intuitions are often taken as the starting point for theories, and intuitions are tested against moral theories until reflective equilibrium is achieved. The process of reflective equilibrium is supposed to provide rational justification for moral theories and moral intuitions, but any explanation of how moral intuitions could be rationally grounded faces two distinct challenges. The first is that of moral dumbfounding, which seems to show that moral intuitions are arational emotional responses to situations that we then try to rationalize by appeal to socially accepted rules. The second challenge comes from the two-systems theory of reasoning, which generally holds that intuitions are not corrigible to explicitly reasoned theories, and therefore not subject to rational revision. This paper addresses both of these challenges, and draws upon a recent model of the mental architecture subserving norms to show “how possibly” the two-systems theory of rationality can actually support the view that moral intuitions can be subject to rational criticism and modified by explicit reasoning. This model, suitably filled out, demonstrates how reflective equilibrium can be psychologically realized, and thus, how moral intuitions can be rationally grounded.

It is a well-established phenomenon in moral philosophy that there are times when moral judgments seem to appear fully formed in the conscious mind, without any prior thought or reflection. Although it may not be commonly put this way, philosophers regularly appeal to this understanding of moral judgments when presenting artificial case studies, like the trolley problem, and attempting to elicit our moral intuitions about the situation. Such intuitions are often taken as the starting point for elucidating theory. For example, Thomson uses her intuitions about the trolley problems and its variants to draw the theoretical conclusion that deflecting harm is morally permissible, whereas causing harm by the direct laying on of hands is not. (Thomson, 1976.)

However, for those who hold that moral judgments must somehow be rationally grounded, the existence of pre-theoretic moral intuitions poses a serious *prima facie* challenge, because intuitions do not appear to be a product of any rational process. Intuitions come easily and automatically to mind, with an almost perception-like quality, which is quite different from the slow and deliberative process that characterizes carefully reasoned judgments. Furthermore, when we reason through a judgment, assuming it is not too complicated, we know how we arrived at the judgment, the logical inferences we made, etc. Intuitions, on the other hand, just seem to appear in the mind, and it takes some real philosophical work to figure out what principles, if any, underlie them.

To further complicate the picture, moral intuitions are not novelties that arise from time to time within an overall schema of carefully reasoned moral judgments. Rather, ordinary online

moral judgments of the sort we have while walking to class or sitting on the bus are much more like intuitions than deductions, coming quite easily and automatically to mind with a perception-like quality. When someone gives up his seat on the bus to an elderly passenger we *perceive* kindness, we do not deduce it. To adopt Williams's famous language, to have to deduce it might itself be a moral failing—it would be “one thought too many.” (1981)

If ordinary moral judgments are like intuitions, and intuitions themselves are not directly deduced by deliberative reasoning, then we must ask whether it remains possible to conceive of the moral life as having any basis in reason, or whether the intuitive nature of moral judgments shows that the moral life is largely arational. To put it another way, is there some way that moral judgments could be rationally grounded while conceding that moral judgments proceed largely by intuition? One highly influential model that suggests that intuitions can be rationally grounded is Rawls's reflective equilibrium. The process of reflective equilibrium begins with moral intuitions informing moral theory, and an ensuing give and take between intuition and theory when conflict arises until a steady state is achieved between intuitions and theory. Reflective equilibrium, therefore, can provide the rational ground for intuitions, because they are subject to rational criticism and refinement. Furthermore, the best research in psychology and cognitive science suggest that a process of reflective equilibrium is the correct modeling of the interplay between reason and intuition, and therefore the best account of how moral intuitions could be rationally grounded.

In this paper, using a model of architecture subserving norms, I shall show how the process of reflective equilibrium could be psychologically realized. But, before doing that, I must answer two *prima facie* challenges to the psychological possibility of reflective equilibrium. The first challenge denies that there is any rational basis for moral judgments. This position is best represented by Jonathan Haidt and his work with moral dumbfounding. (Haidt, 2001) Moral dumbfounding occurs when a person has a strong intuitive judgment about the permissibility of a certain action, for example, that it is prohibited, but is unable to rationally justify its wrong-making features. Haidt uses this phenomenon to argue that reason plays only a secondary role in the moral life—that of defending our intuitively arrived at judgments by appeal to socially accepted rules, which he calls social intuitionism. On this account, then, moral judgments are not grounded in reason, rather reason functions to provide a plausible post hoc justification for the judgment. Haidt's interpretation of moral dumbfounding presents a serious challenge to my thesis that the intuitive nature of moral judgment is consistent with a largely rational picture of morality. However, a more plausible explanation of moral dumbfounding actually supports the possibility of rationally grounded intuitions, so it is worth spending some time on the phenomena of moral dumbfounding and how it is to be interpreted.

Having discussed moral dumbfounding, I shall turn to the second, and more difficult, *prima facie* challenge to the possibility of a cognitivist account of intuition raised by the “two-systems” model of reasoning developed in economics and psychology. The challenge raised by this model is that it denies the possibility of intuitions as being rationally modified, because it views intuitions as largely incorrigible to explicit reasoning. Although the two-systems theory seems initially to undermine the possibility of rationally grounded intuitions, by applying it to a psychological model of the mental architecture subserving norms, it actually supports the process of reflective equilibrium, and thereby supports the possibility of rationally grounded intuitions.

I. Moral Dumbfounding

Haidt's example of moral dumbfounding begins with the following case study:

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 the very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom 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)

According to Haidt's research, most people judge this to be seriously morally wrong, but when pressed to explain why, they are unable to articulate the wrong-making feature of the case. Many start by making appeals to principles of harm, for example, that Julie and Mark will be emotionally traumatized by the experience, or that they will have defective children. Yet, this case excludes the possibility of harm on these fronts, and even when subjects are reminded of this, they nevertheless judge the actions of Mark and Julie to be seriously morally wrong.

This is an intriguing result, because it demonstrates a disconnect between people's judgments of the situation, and their reasoning about the situation. Haidt thinks that any account of moral judgment needs to take moral dumbfounding seriously and give an account of moral judgment that can explain how it happens, and that, he thinks, is what the social intuitionist model accomplishes. The social intuitionist model claims that moral judgments are essentially emotional responses to situations, and that when we reason about them we are seeking only to tie our emotional response to some socially accepted rule. And we only do this so as to gain social allies, and get along in our society. Haidt further argues that moral arguments do not so much seek to persuade others about the rightness of our judgment, rather, they are meant to produce in others the same feelings we have by pointing out new features that they perhaps did not initially consider. This model, he thinks, explains moral dumbfounding because people can have an emotional response to the case, and continue to hold it, even though there is no socially accepted rule that can vindicate it.

It is clear that Haidt's social intuitionist model can explain moral dumbfounding, but it has come under both methodological and philosophical criticism recently because it fails to explain anything else; e.g., that we change our minds about moral judgments after thinking more about the situation, and that moral arguments can be persuasive in ways that are characteristically rational.¹ (Saltzstein and Kasachkoff, 2004.) As such, the social intuitionist appears to be merely *ad hoc*—devised solely to explain the phenomenon of moral dumbfounding. However, it remains an open question whether it is the only or even the best way to make sense of dumbfounding. Furthermore, Haidt's argument against the rationality of moral judgments relies on a heavily idealized notion of moral reasoning. Haidt seems to assume that the only proper rational justification for any moral judgment is its direct deduction from universal principles of morality. That is, if a particular moral judgment cannot be shown to have been deduced from universal principles, its status as a rational moral judgment is questionable.

¹ Indeed, in another study co-authored by Haidt one respondent's answer seems to directly challenge the social intuitionist position: "one conservative woman... began by condemning homosexuality, but as she thought about the possibility that sexual orientation is innate rather than chosen [she said], 'If you get right down to it, then their act shouldn't be condemned either.'" (Haidt and Hersch, 2001; 218) She thus reasoned herself to a new position, in direct conflict with her intuition about the case.

Thus, what he has in mind for rationally grounded judgments is some sort of Aristotelian practical syllogism, such that we could say (1) Harm is morally bad; (2) Incest is harmful; (3) Therefore, incest is wrong.

Haidt has succeeded in showing that we do not *always* reason in the form of an Aristotelian syllogism starting from general principles of morality, but that alone is hardly sufficient to establish social intuitionism in place of a more cognitive account of moral judgment, because it rests on the mistaken assumption that syllogistic reasoning from universal principles would be rationally required for each and every moral judgment. That is simply too heavy a cognitive burden, even for the most committed reasoner. Yet, even if such thorough reasoning is not required for every moral judgment, other important moral phenomena, like the rational persuasiveness of moral arguments, indicate that moral judgments are somehow rational, even if not reasoned from universal principles. What we need, then, is an alternate explanation of moral dumbfounding that takes the rational nature of moral judgments seriously.

Consider again the incest case. Clearly we are not starting from universal principles to reason to a conclusion such as: (1) Harm is morally bad; (2) Incest is harmful; (3) Therefore, incest is wrong. If we did reason in this way, then we should not see dumbfounding, since premise (2) is obviously false. However, if online moral judgments are arrived at by reference to mid-level principles like, “incest is wrong,” instead of universal rules like “harm is morally bad,” we can readily account for moral dumbfounding. Let us assume, for the moment, that “incest is wrong” is a mid-level rule. The syllogism would proceed as: (1) Incest is wrong; (2) This is a case of incest; (3) Therefore, this is wrong. If this is what is occurring, then what moral dumbfounding demonstrates is that it is possible to detach mid-level moral rules from their initial justification by universal moral principles, while leaving the moral rule still in play. If the initial justification for a rule against incest is that it is harmful, that rule still issues in online moral judgments even in strangely contrived cases where incest is not harmful. Thus, people will judge that Julie and Mark’s behavior is seriously morally bad based on their mid-level rule against incest, even though if asked to justify it in this particular case, they cannot. This explanation, then, retains the rationality of moral judgments, while accounting for the possibility of moral dumbfounding. There are also, I think, independent reasons for thinking that this is the case, since detachment of this sort occurs in other domains as well, e.g., beliefs deriving originally from testimony or observation.

For example, say I listen to a reputable radio program that I have good reason to believe makes truthful reports, and when in error, takes pains to issue corrections. One day they report that scientists have just discovered two additional moons orbiting the planet Pluto, for a total of three. Given that this is a reputable program, and the possibility of scientists discovering new moons around Pluto does not conflict with any of my other relevant beliefs, I form the belief that Pluto has three moons. Furthermore, forming the belief that Pluto has three moons in this way seems to be *prima facie* rational. Years later, having forgotten all about the radio program and their report on Pluto’s moons, I get involved in a conversation that turns to astronomy, and someone says, “Pluto has one moon named Sharon.” I immediately correct him, and say, “No, Pluto has three moons.” Unsatisfied with this assertion, my interlocuter asks where I learned about Pluto’s moons. I have to admit that I cannot remember, but that I am certain it is correct. The question, then, is whether I am rational in maintaining my belief that Pluto has three moons, even though I cannot remember exactly how I came to form that belief. We are, I think, inclined to say yes, because being able to remember the processes or particular observations that justify a belief are not themselves part of the justification. We may well be justified in a holding a belief

when we have forgotten where we learned it, so long as we fulfilled the requirements of rationality when the testimony was accepted and the belief was formed. (Burge, 1993.)

Similarly, people who judge Mark and Julie to have done something seriously morally wrong are making rational judgments about the case, so long as we understand rational to mean consistent with their mid-level rule that incest is seriously morally wrong. It is highly plausible that participants in Haidt's study think that what the siblings are doing is seriously morally wrong simply *because* it is a case of incest. That is, they are making the judgment from their instantiated mid-level rule that incest is wrong without trying to derive it anew from general justifying principles of morality.

So, this model too accounts for moral dumbfounding, but does so within a broadly cognitivist framework, since the rule "incest is always seriously wrong," is rationally justifiable by reference to universal moral principles.² And the upshot is that it does so without minimizing or diminishing other of significant moral phenomena, e.g., the persuasiveness of moral arguments. Moral dumbfounding, therefore, does not pose a serious challenge to the claim that judgments can be largely intuitive while still being grounded in reason. Rather, it helps shed light on the relationships among theory, intuition, and reason. Intuitions proceed from mid-level rules, which are, in turn, justified by reference to universal moral principles, which is the process of reflective equilibrium.

On this model, intuitions inform theory by giving us the raw data on the sorts of acts that are and are not morally permissible by giving us insight into the sorts of mid-level rules we have already internalized. From these we derive basic principles of ethics by elucidating the wrong- and right-making features of situations, and can develop a full-blown moral theory based on those principles. Once our theory is in place, we may find that certain intuitions disagree with the requirements of our theory, and then we face a choice between modifying our intuitions and our theory so as to bring them into accord.

The process of reflective equilibrium maps onto the cognitive explanation of moral dumbfounding fairly well. However, if we accept reflective equilibrium as the correct account of intuitions and rationality, a new, much more difficult problem emerges. Much work has been done in recent years on the relationship between intuition and reason in human judgment, and a consensus has emerged around the "two-systems" model, which views intuition and reason as being realized in two different cognitive systems. (Evans and Over, 1996; Stanovich, 1999; Kahneman, 2002.) The problem is that the two-systems model stipulates a one-way relationship between intuition and reason: intuition can affect reason, but reasoning cannot affect intuition. If the two-systems model is correct, and there is good reason to think that it is, then it appears that reflective equilibrium is an unrealizable ideal, since reason could never have a direct influence on the sorts of moral intuitions we have. This places a cognitivist account of moral judgments in serious jeopardy, because without such an influence, the mid-level rules that issue in intuitive moral judgments would operate quite independently of our reasoned moral theories, and would not be sensitive to rational critique or refinement, which would, in turn, be a vindication of

² It might be objected that Haidt's incest case is a clear counterexample to the rule "incest is always seriously morally wrong," because his case is one where incest is committed, but nothing seriously morally wrong seems to have occurred. This, however, blurs the line between the rules by which we make quick, online moral judgments, and the general principles that ultimately justify those judgments. The rule that "incest is always seriously morally wrong," is justified by the moral principle to avoid harm and a posteriori facts about incest. Particular moral judgments, in turn, are justified by reference to the mid-level rule. Thus, the judgment that this case of incest is wrong just because it is a case of incest is a justified one.

Haidt's view that reasoning in morals is always just *rationalizing*. Since the two-systems model raises this difficulty, it is worth looking at in some detail to see how it might be answered.

II. Two-Systems Model: A Psychological Account of Intuitive Judgment

If there is any lesson to be drawn from experimental economics and psychology, it is that humans are far from ideal reasoners. For example, in a paradigm case from Tversky and Kahneman (1983), subjects were presented the following vignette:

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student she was deeply concerned with issues of discrimination and social justice and also participated in antinuclear demonstrations.

Subjects were then given options about Linda's current employment, and asked which was most likely based on what they knew about her past. Among the options were "Linda is a bank teller," and the conjunction "Linda is a bank teller and active in the feminist movement." Since the probability of a conjunction can never be more than any one of its conjuncts, it cannot be more likely that Linda is both bank teller and active in the feminist movement, than that Linda is just a bank teller. So, between these two options, the right answer is "Linda is a bank teller." Yet, the majority of respondents answered that Linda was more likely a bank teller *and* active in the feminist movement—a strikingly irrational response.

We might think such mistakes in logic and probability have to do with some fault in the respondent's education. Perhaps it is just not common knowledge that the probability of a conjunction can never be more than any of its conjuncts, so this is not so much a lapse in logic as it is a lack of the proper knowledge needed to solve the task. This is a tempting response, but in other experiments Kahneman (2002) found that even his highly educated colleagues made significantly and persistently erroneous judgments involving statistics, even though they possessed all the relevant knowledge to solve the tasks correctly. And this is not just an isolated case. Results like these have been reproduced across a variety of domains, (for a review see Brenner, Koehler & Rottenstreich, 2002) such that it is now a relatively robust finding that humans regularly and persistently reason incorrectly, even when they have all the resources needed to do so correctly.

This raises a provocative question: how to explain this gap between the normative requirements of reasoning and how people actually reason, especially given that education seems to have little influence? This question was the primary motivation behind the two-systems model of the human mind. In this model, System 1 reasoning is largely unconscious, involuntary, automatic, and driven by heuristics, where heuristics are understood as simple associative mental shortcuts. System 2 reasoning, on the other hand, is conscious, slow, and deliberative. (Stanovich & West, 2000; and Kahneman, 2002.) Furthermore, System 1 processes are often difficult or impossible to modify; that is, they are generally unaffected by learning or education, whereas System 2 reasoning is easily corrigible by education and learning. In ordinary language, System 1 reasoning is usually referred to as an intuition or a "gut reaction," i.e., an effortless, almost perception-like judgment of a situation; and System 2 is what we usually mean by explicit reasoning. As Kahneman (2002) puts it: "The operations of System 1 are fast, automatic, effortless, associative, and difficult to control or modify. The operations of System 2 are slower,

serial, effortful, and deliberately controlled; they are also potentially rule-governed.” Furthermore, System 1 processes, since they are unconscious, are not accessible by System 2.

These descriptions by themselves, however, are not enough to explain why people who know better continue to make erroneous judgments, for we could easily imagine System 1 and System 2 working in parallel, achieving different results, and the person well-versed in probabilities would rationally choose the answer arrived at by explicit reasoning, ignoring her “gut reaction.” So far there is nothing in these descriptions of the two systems that precludes this possibility, except that we know that this is precisely what is not happening. Those who know better continue to trust their gut, so there must be something more than just two systems in the mind working independently of each other—they must have some kind of relationship that prevents people from behaving in the way I suggested.

Indeed, there is. System 1 is so quick and effortless that it can arrive at a highly plausible answer before System 2 even gets started. So, it would be strange to think of them working in parallel when System 1 is so much faster than System 2. Rather, by default, they work serially, with System 1 issuing in quick intuitions, and System 2 providing oversight and correction of those intuitions. Thus, a System 1 intuition might be that Linda is most likely both a bank teller and active in the feminist movement, and System 2 is responsible, if you will, for signing off on that judgment. It can accept, reject, or modify System 1 intuitions as it pleases according to whatever explicit rules are brought to bear on the question. So, some people might remember their training in logic, and reject their intuition about Linda, and opt instead to say that she is just a bank teller, since a conjunction can never be more likely than any of its conjuncts. This is possible, but the fact that most people do not overturn their erroneous judgment about Linda suggests that System 2 oversight is somewhat lax. It seems that most people are, “often content to trust a plausible judgment that quickly comes to mind,” (Kahneman, 2002; p. 452) rather than laboriously and slowly refiguring the solution on the System 2 level.

Laziness may seem like a poor excuse for scanty System 2 oversight, but System 2 reasoning is subject to a number of constraints that affect its oversight ability. Some common constraints include time pressures, simultaneously working on different tasks, one’s mood, and strangely enough, the time of day (morning people tend to do perform worse on cognitive tasks in the evening, and vice versa for evening people). (Kahneman, 2002.) With all these constraints on System 2 reasoning, it is really no wonder that most people, most of the time, simply go with their System 1 intuition. However, people will overturn a System 1 intuition that violates a System 2 rule if they remember the rule quickly enough, which might be affected by other factors, e.g., how recently they last thought about that rule. (Kahneman, 2002.)

So, this model of two-systems where System 1 produces quick, intuitive judgments driven largely by mental shortcuts, and System 2 is the somewhat lax and gullible overseer explains the observed gap between normative requirements of rationality and how we reason in real cases. Furthermore, this model is highly plausible for independent reasons, because, as we said before, System 2 thinking is slow and laborious, whereas System 1 reasoning is quick, and for the most part, reliable *enough* to get by in most ordinary circumstances; so this model explains the computational tractability of judgment. Rarely do we have the opportunity to think carefully about the sorts of decisions before us, and System 1 heuristics, for the most part, point us towards judgments that are good enough for everyday purposes. Thus, the two-system model predicts a certain amount of mental efficiency, and it is hard to see how we could function without it.

If we apply this model to the process of reflective equilibrium, our moral intuitions are System 1 level judgments, and theory building is a System 2 level activity that issues in a commitment to act as if our theory were true. If a System 1 moral intuition conflicted with an expressly held System 2 moral rule, or if we had conflicting System 1 intuitions, it would give us cause to pause, and take time to reason through the situation more carefully. This would be consistent with what we see in actual moral practice, (Cohen, 2004; especially ch. 6) but it would only be part of the story. Remember that System 2 oversight is sloppy and inconsistent, so relying on it to constantly check our moral intuitions would be less than ideal. Furthermore, what reflective equilibrium requires is that our intuitions themselves align with our best moral theories, and that requires something far more complicated than this simple two-systems story will allow. It requires that our intuitions change, which, in turn, requires a change at the System 1 level—a change directed by our System 2 level reasonings and commitments. The two-system account given so far does not allow System 2 to have that kind of affect on System 1.³ In fact, it seems to prohibit such interactions outright. However, recent work on the psychology of morals may allow us a way forward at this point, so let us turn to that now.

III. The Psychology of Morals

In “A Framework for the Psychology of Norms,” Sripada and Stich (2006) draw on evidence from the fields of cognitive science and psychology to offer a very persuasive and compelling account of the cognitive architecture of human norm acquisition and norm deployment (Figure 1). In the figure, solid lines indicate links for which there is very strong evidence, and the dotted lines indicate links that are a bit more speculative. Aside from the wealth of information they draw on in their analysis, a real strength of this diagram is that it accounts for some of the most robust and basic facts concerning norms: that all ordinary children acquire norms from their environment; that all ordinary people make judgments using them; and that judgments of norm violation in particular feel a certain way to us.

In broad view, the norms system is divided into two parts: the acquisition system and the execution system. The norms acquisition mechanism allows children to acquire the rules of their community by being sensitive to the appropriate environmental cues, and parsing those cues in such a way that the right rules of conduct of that community are extracted. There is no exhaustive list of what the appropriate environmental cues are, but there is evidence to suggest that at least some of those cues are direct verbal admonitions like, “Don’t hit your sister,” and the observation of emotional responses in others. (Turiel, 1998.) For example, an act that elicits negative emotional responses from peers and caregivers is clear evidence that a norm has been violated, at least as clear as explicit statements. So far there is no indication of how much evidence a child needs of a rule before he or she acquires it, but it seems highly implausible that only one instance would be sufficient.

After a child parses the correct rule out of the social evidences available, that rule is stored in the norms database, and mapped to a set of perceptual inputs that should invoke that rule and issue in a judgment. For example, if the rule “do not steal” is stored in the norms database, it is mapped to the sorts of observable phenomena that count as instances of stealing. The database is also connected up to the emotion system, which attaches an affective valence to moral judgments. It seems uncontroversial to say that moral judgments feel a certain way—

³ Carruthers (ms) argues that, at least in the case of belief, System 2 beliefs can issue directly in System 1 beliefs.

judgments of wrongness have a negative valence, while judgments of rightness have positive valence.

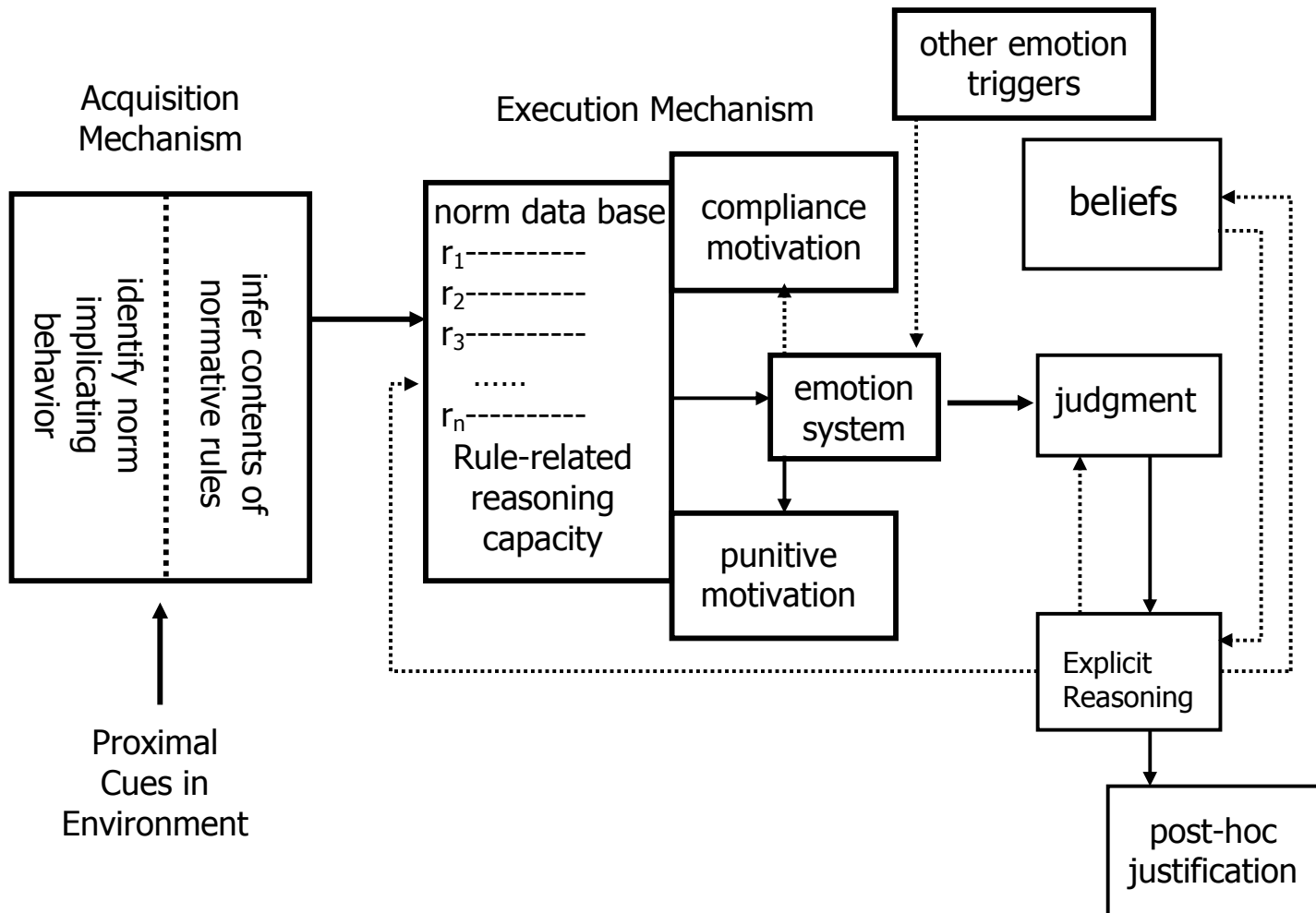


Figure 1. From Sripada and Stich (2006)

Furthermore, once a norm is acquired, the judgments that result have all the hallmarks of System 1; they are quick, involuntary, and automatic, and strike us with an almost perception-like quality. It seems likely that when we speak of our “moral intuitions,” what we are referring to are the judgments made in just this way. We simply “see” the situation as one where a norm is being violated, or where one is being exemplified. We can’t help but engage the world in these terms, and this again is a characteristic of System 1 type operations. Also characteristic of System 1 is that people often cannot say *why* they think something wrong, only *that* it is. The trolley problem, and fat man variation, are excellent examples of this kind of phenomenon: people commonly judge that it is morally permissible to deflect the trolley so that it kills one person rather than five; but then judge throwing a fat man in front of the trolley that would also save five as being morally impermissible. Most people cannot articulate the rules that give rise to

such judgments, and indeed, it takes quite a bit of philosophical work to figure it out, and the various different proposed reconciliations are highly controversial. (Thomson, 1976.)

Another analogy with language might be appropriate. In the language case, once we acquire a language, we can't help but hear certain strings of noises as language, and not only as a case where language is being used, but we immediately understand the meaning of those utterances (provided they have meaningful content). We do not need to reflect on the words being used, and think about the particular meaning of those words in order to understand. Rather, we often think of "perceiving" the meaning of a sentence—that is, we recognize that the meaning of sentences and utterances is perception-like, immediate, unconscious, and automatic. Furthermore, a common experiment in linguistics meant to derive the principles of grammar for a language is to put together a sentence in the language being studied, and then ask native speakers for their intuitions about the grammatical correctness of the sentence. Oftentimes, subjects know *that* a sentence is wrong, without ever being able to articulate the grammatical rule that makes it wrong. (see Chomsky, 1988; and Crain and Pietroski, 2001)

Going back to the Sripada and Stich diagram, it is only at the point that a particular moral judgment has been made, and the proper affective valence attached that the diagram moves to the System 2 level. On this level are explicit reasoning, and *post hoc* justification. As we saw before, System 2 is responsible for overseeing and checking the output of System 1, thus, at this level, the judgments arrived at by System 1 are consciously available, and at this point explicit reasoning can become involved. Sripada and Stich, however, leave it a partially open question what role System 2 reasoning plays in moral judgments, and in the moral life. For example, they do not go into detail whether System 2 reasoning can modify or reject the System 1 judgment. It seems from our discussion of the two-system model that it should, in principle, be possible. However, the only function they see as being well established is that of *post hoc* justification, i.e., the practice of attempting to justify our moral judgments by appeal to socially accepted rules, rather than attempting to check our judgments against an explicit moral theory. Other possible functions of System 2 reasoning in the moral life consist in the speculative connections from explicit reasoning to beliefs and the norms database. And, although this is not clear from the diagram, we should probably think of the beliefs box as consisting in part of System 1 beliefs and System 2 beliefs.

IV. Two-Systems in the Moral Life

The Sripada and Stich diagram was introduced as a way to solve the difficulty presented by the two-systems model, in which System 2 reasoning has no effect on System 1 intuitions. Reasoning occurring on the System 2 level, insofar as it conforms to requirements of rationality, can tell us what sorts of moral rules are rationally required, permitted, or forbidden. And, if moral rules arrived at the System 2 level can be instantiated at the System 1 level, as the diagram indicates might be possible and as reflective equilibrium requires, then it is also possible that System 1 level norms could be rationally grounded, and the intuitions they issue in, therefore, should also count as rational. It seems to me, at least, that this would count as moral cognitivism, and it would also be a psychological account of reflective equilibrium, because it could explain how we are psychologically able to achieve consonance between intuition and theory. The solution, I think, lies in working out "how possibly" the speculative connections between judgment, belief, and the norms database could function.

Before we launch into that discussion, an important distinction needs to be made between two different senses of the word “norm.”⁴ For psychologists and social scientists, the term norm is simply meant to refer to an action-guiding rule. It is a rather vague formulation of what it is to be a norm, but entirely consistent with the psychological and social project of attempting to explain why people do certain things, and one reason why people do certain things is because they have a certain set of action-guiding rules that have a strong and direct influence on behavior. So, in this first sense, a norm plays an *explanatory* role; i.e., it explains why people act in certain ways. This explanatory role also extends to the question of why we judge certain things in certain ways, i.e., because of the presence of some action-guiding rule. Norms, in this sense, are what anthropologists discover when studying cultures—they find the set of rules that guide actual behavior. So, this first sense of norm is meant to be purely descriptive. It makes no claim as to whether the action-guiding rule is the right one, only that it is the one present. When referring to norms in this sense, I will call them “psychological norms.”

Contrast this purely descriptive with the normative sense. In the normative sense, a norm is the standard by which we *should* judge the wrong and right making features of an action that make an act morally permissible or impermissible. In this sense, norms are *evaluative*, in that they allows us to evaluate our psychological norms. So, it is quite possible for someone to be acting in accord with a norm in the psychological sense, but not acting in accord with a norm in what I will term the “evaluative sense.” That is, we can explain why someone did what they did by referring to some psychological norm, and we can also evaluate that psychological norm as being morally wrong by reference to some evaluative norm. In the following discussion, the term “norm” when used without qualification will refer to psychological norms, and when referring to norms in the evaluative sense, I will use the term “moral norm.”

Moral theories are attempts to elucidate the correct moral norms, and there are at least four interlocking motivations for constructing them: (1) to systematize our moral judgments; (2) to check for consistency among our moral judgments; (3) to ascertain universal moral principles and maxims; and (4) to check our judgments against universal principles and maxims. This list is not meant to be exhaustive, rather, reflective of the motivations found in the philosophical literature going back as far as Plato. For instance, many of the Socratic dialogues are attempts on the part of Socrates to get his interlocutors to check for consistency of their moral judgments. Thus, when Socrates questions Euthyphro about his claim that he is morally required to prosecute his father, he finds that Euthyphro has a number of conflicting moral judgments that eventually lead to his famous paradox. Similarly, in *The Republic* Adiemantus and Glaucon give conflicting accounts and inconsistent answers about the nature of justice, which then provides the motivation for Socrates’ all-night exercise in moral theory building.

For others, a primary motivation for moral theory is to ascertain rationally required universal moral principles, and then ensure that our mid-level moral rules, or psychological norms, are consistent with them. Rawls’s theory of reflective equilibrium gives an account of how this can be realized. And here is where the defense of moral rationalism gets a grip. Rawls’s theory of reflective equilibrium is very much a cognitivist account of morality, and the two-systems process I am going to outline is no more than a psychological account of how this process might work.

Developing a moral theory is a System 2 process. It requires slow, deliberate reasoning, and sometimes university level courses in ethics. It is not an easy thing to do, which seems to be

⁴ I am indebted to Matthew King for making this point clear to me.

why many people never construct anything close to a full-blown moral theory. Even today, most people cannot say why they think it morally permissible to deflect the trolley to save five people, but think it impermissible to throw the fat man in the trolley's path to achieve the same end. So, judgments like these are the result of System 1 level processes. The question, then, is how do System 1 and System 2 interact when constructing moral theories?

Since the norms in the norms database are not consciously accessible to us, we cannot simply look inside for the data needed for our theorizing. However, the norms system does produce a plethora of consciously available data in the form of judgments. We may never know exactly what norms are in our database, or in what form, but we can start our theory building apart from such knowledge by relying on the sorts of judgments we already make. Thus, if I have many judgments of situations involving harm that say, "that is bad," then I can infer something about the content of my norms database, *viz.*, that things involving harm are bad, as well as begin to build theories in accordance with them. These judgments are the outputs of System 1 processes, so there is an obvious movement of things from System 1 to System 2, where System 2 can then reason about and build up theories.

There is nothing problematic in this account so far, because it is already well established that System 1 processes have an influence on System 2 processes in that they provide the fodder for System 2 reasoning. However, suppose that when constructing our moral theory we find that some set of our moral intuitions is seriously out of line with the requirements of our moral theory. Ideally, what we want to be able to do is make some change at the System 1 level such that in the future, when confronted with situations that normally evoked these problematic judgments, we make those judgments that conform to the requirements of our theory. That is, we want to be able to make the "right" judgments as quickly, unconsciously, and effortlessly as we currently make the wrong ones. If such a change could not be affected, we would be required to overturn those wrong judgments at the System 2 level, which we have already seen is an unreliable and lax supervisor. Furthermore, constantly overturning System 1 judgments in this way would require a tremendous amount of cognitive resources; resources that may not always be available. Thus, if the ideal is that our moral judgments be reliable, automatic, and effortless, they must issue from System 1, rather than System 2.

Yet, this an argument only that System 1 is better suited to the task of making consistent moral judgments than System 2, but says nothing about whether it is psychologically possible for System 2 reasonings to somehow become instantiated at the System 1 level. Generally, System 1 is incorrigible to System 2 operations. Thus, for many operations in the mind there is just a unidirectional relationship between the two-systems that always runs from System 1 to System 2, e.g., those for logical and probabilistic reasoning. So, we may ask is there anything about the norms systems that should give us reason to think that it is alterable? Indeed, there is if we consider that the norms system is a learning system—that is, the psychological architecture of the norms system may be universal, but the *content* of the norms system is not hardwired, rather, it must be acquired. And it is the content that matters when it comes to the sorts of judgments that are made, because our judgments issue from the rules that are instantiated in the System 1 level. Thus, if the acquisition mechanism in the Sripada and Stich diagram were to remain intact throughout a person's life, then it would be possible to alter the content of the norms database by acquiring new rules, say, the rule required by our normative theory, and once a new rule is acquired in the norms system the judgments that proceed from it will be consistent with the requirements of our theory.

It might be questioned whether we have any reason to believe that the norms acquisition mechanism remains intact throughout life. Perhaps an analogy with language might provide a way forward. Language is the paragon of learning systems, and some observations about language can probably teach us something about how learning systems function throughout life. Language, like norms, is acquired very early in life, assuming the right sorts of environmental stimuli are present. Furthermore, proper language acquisition needs to take place within a critical period early in life, such that failure to be exposed to the proper stimuli within that period means that competency can only be acquired with great difficulty, and even then not always perfectly.⁵ (Kuhl, 2000; Kuhl et al., 2005.)

However, even if a child is exposed to the correct stimuli at the right time, the learning mechanism becomes greatly attenuated at some developmental time, perhaps because of neural pruning. Thus, learning a second language later in life is a much more arduous and explicitly reasoned task than acquiring a first language early in life, requiring drilling, memorization, and lots of practice. Even so, students of second languages can attest to the fact that, once they become familiar enough with a second language, it becomes possible to “think” in that language. They are able to “switch” among the languages, and converse as easily in the second as they can in the first in a way that does not require them to explicitly reason through the rules of grammar that they assiduously studied. So, in some way, the language learning faculty does not shut off completely, but becomes greatly attenuated, and reliant on some process of explicit reasoning. It is still active, but it takes a lot more work to become fluent in that language.

Can an analogous process be at work in the norms acquisition system? If so, it would predict that coming to believe that one ought to act in accordance with a rule discovered by means of explicit reasoning could come to be instantiated in the rules database, but in order to do so it would require lots of practice with the rule, and significantly more time to become “second nature” than the rules acquired during childhood. Furthermore, just as in the language case, we would expect that during the process of acquisition, the rule would be wrongly or inconsistently applied.

Beyond this analogy with language, there are also independent reasons for thinking the norms acquisition mechanism remains intact throughout a person’s life. For example, norms can change within a society during one’s lifetime, and different norms can prevail in different environments, e.g., acceptable behavior as an undergraduate is not the same behavior that is acceptable as a professional. Also, when one becomes a member of a subgroup, say, a religious group, changing one’s norms to those of that group is expected, and it is expected, I think, in part, because it is possible. Lastly, being able to change one’s norms would be very adaptive. In early hunter-gatherer societies, women from conquered communities could be forcibly taken as wives, and being able to adjust to the new norms would ensure survival. It would be a short life for a captured woman who habitually and consistently breached the norms of her new community. However, simply being able to learn instrumental force-backed rules like, “If I disobey my new master, I shall get whipped,” would be enough to ensure survival in the sense of not getting killed, but the real question is of what would be the benefit of *internalizing* the norms of one’s new society. One possible benefit is integration into that society. These women, and their children, would more likely do better if they became fully-integrated members of their new

⁵ Generally, children who have not been exposed to language are neglected or horrifically abused, thus it is hard to separate out problems associated with language, and those associated with other psychological problems brought upon by such abuse and neglect. See, Snow and Hoefnagel-Höhle (1978).

society, rather than holding onto the norms of their previous community and doing just enough to stay alive.

Assuming that the acquisition mechanism remains intact, it becomes, in principle, possible to change our System 1 norms, and this provides the necessary feedback from System 2 that makes reflective equilibrium possible. At this point, work done by Frankish (2004) on two-systems and belief may provide a way forward. He argues that the best way to understand the differences between various kinds of beliefs, like occurrent versus dispositional, is to use the two-systems model. According to Frankish, System 1 beliefs are those that come to mind easily and unconsciously, whereas System 2 beliefs are those that have to be thought out, perhaps even deduced from System 1 beliefs. Since the two-systems model requires a sharp division between System 1 and System 2, Frankish argues that System 1 beliefs are different in kind than System 2 beliefs. On this account, a System 2 belief is a *commitment*. As Frankish intends it, a System 2 commitment is when we commit ourselves to act in the future as if our System 2 belief were the case, even if we continue to believe something else at the System 1 level. So, if on the System 1 level we do not believe in God, but on the System 2 level we do, since our System 2 beliefs cannot have an affect on our System 1 beliefs, what we have really done, according to Frankish, is commit ourselves to acting in the future as if God exists. So, we act as if p were the case, even if the System 1 belief is $\sim p$. Applying this to the moral case, when we develop our moral theory, we infer a number of mid-level moral rules that ought to guide our behavior, and even if on the System 1 level we have not acquired that rule, we can commit ourselves to acting in the future as if our theory-derived rule were the right one.

So, the first part of acquiring a new moral belief is forming a commitment to act in the future as if we had that moral belief such that if our initial judgment is $\sim p$, we act as though p were true. Obviously, this is a case of serious cognitive dissonance, and a situation that seems to me, unlikely to persist. Overriding our initial judgments requires serious attention and cognitive resources, and it would not take long for us to return to our initial *status quo*, just for sake of ease and cognitive peace. And sometimes this is exactly what we see. People make a serious commitment to act one way, even though they judge another, but it becomes too difficult, and they return to their original state. And this seems to me a fairly common phenomenon, and we have a number of ways to excuse ourselves from the requirements of our theory, like, “it’s just the way I am;” or “it’s just an ideal, I never really expected to live up to it.”

Yet, other times, the rule persists, and, it seems unlikely to think the rule persists where it remains only a commitment given the attention and cognitive resources it would require to consistently act in accord with a System 2 commitment. When a rule persists, it is more likely that the System 2 commitment becomes instantiated at the System 1 level, and this is how it might happen: when we make a judgment that conflicts with our second-order commitment, we override our initial judgment, and as we do so, we verbalize to ourselves our commitment. When invoked often enough, the self-verbalization becomes evidence of a new rule to the rule acquisition mechanism. That is, this practiced commitment itself becomes part of the moral environment whereby rules are adduced and added to the rules database. And this might be for no other reason than the fact that when we are thinking about our commitment, we are expressing to ourselves in a natural language, in effect, lecturing ourselves as our peers, parents, and whoever else might lecture us about certain norms. As we said before, admonitions expressed in a natural language are among the evidences available to the norms acquisition mechanism when parsing a rule out of the environment. So, as we repeat this commitment to ourselves in several cycles of moral judgments, the self-verbalization of our commitment can be

taken as evidence by the norms acquisition mechanism that there is a new norm that must be parsed and added among those already in the database.

Thus, the new rule is acquired in the rules database through the same mechanism as all other rules are, the only difference is that the initial evidence that a rule is in play proceeds from the self-verbalization of a rule, rather than the verbalization of a rule from those in the environment. Regardless of its origin, however, the rules acquisition mechanism takes that rule, and through its own operations, places that rule, in whatever form it might be stored, in the rules database. Once the new rule is acquired, System 1 judgments will be made in accord with it, just as with any other rule. And this, then, would result in integrity between the judgments and the commitment, and result in the sort of moral integrity we seek in our lives between our moral beliefs and moral judgments.

There are also other avenues to acquiring a new rule, based on the sorts of evidences available to the norm acquisition system. The above picture requires only that linguistic utterances be available as evidence to the norm acquisition system, but other sorts of evidences abound in the environment as well. Indeed, according to work by Nucci (2001), linguistic utterances seem to be rare for children, and rules rarely explicitly stated. Instead, children usually acquire rules by observing behaviors, especially emotional responses of approval, disapproval, and disgust.

Consequently, there is a second way we might think the acquisition process works that is a complement to the first process that draws on both admonishment and emotional cues. Again, we develop a set of rules at the System 2 level and commit ourselves to acting as though it were true. Perhaps through weakness of will, we perform an act that we have now committed ourselves to not performing, or *vice versa*, even though we quite possibly verbalize to ourselves, “don’t do it!”

When this happens, it is a violation of a very basic rule: Do what you have committed yourself to doing. This seems to me one possible formulation of the promise-keeping rule. So, now the rules system parses our behavior as an occasion where a promise has been broken, and this produces the judgment: “You broke your promise! That is wrong.” This produces a conflict of judgment, because the rule that produced the initial judgment said the action was permissible, but now, through breaking a commitment, there is a second judgment that something wrong has been done. You have one rule that says “you did something wrong” and another rule that says what you did was permissible. However, the judgment that we have done something morally impermissible, i.e., broken a promise, feels a certain way to us—it has a very negative emotional valence. In effect, we are disapproving of our behavior. At the same time, we are quite aware as to why we disapprove of our behavior, because, most likely, we feel bad about not doing what we committed to, and say things to ourselves like, “I can’t believe I did ~p, when I know the rule I arrived at says p is right and ~p wrong.”⁶ Now we have provided two evidences to the rules acquisition system: emotional disapproval, and self-verbalization.

However, this process does not rely entirely on negative feedback in the form of emotional disapproval. If we act in accord with our System 2 commitment, this provides us with the positive emotional approval that goes along with keeping our promises, along with self-verbalization of the rule. In time, this process of negative and positive emotional responses to the breaking and keeping of the rule respectively, along with self-verbalization, provides the rules acquisition mechanism enough evidence to parse the right rule of action, and inserts it in the

⁶ Indeed, 18th Century British moralists thought that this feeling of self-disapproval was the very heart of moral judgment. See Hutcheson (1728) and Shaftsbury (1711).

rules database. Once that happens, System 1 level judgments proceed in the normal way by reference to that rule.⁷

One complication, however, needs to be addressed: if the newly instantiated rule conflicts with an already existing rule in the norms database it would seem to follow that, in the presence of the correct stimuli, both rules would be invoked and issue in conflicting judgments of the situation when what we want is a single judgment in line with our considered moral theory. What is needed, then, is some way for norms to be erased from the database. At present, I can only hint at how this might be accomplished, but looking back at the Sripada and Stich diagram there are two obvious places for the connection between a rule in the norms database and the judgment it issues in to breakdown. The first, less likely scenario, is a breakdown between an observed instance of a violation or fulfillment of the rule and the emotional valence attached to it before the actual judgment. If the breakdown occurred here, then we would make judgments with the old rule, but the judgment would fail to feel like a moral judgment. The second possibility is at the level of mapping incoming stimuli to the rule. If the mapping were altered such that incoming stimuli were no longer mapped to the old rule, then that rule would no longer issue in judgments. Since the mapping between stimuli and rule is part of the learning system, this seems the most likely way for a rule to be functionally erased.

A final corollary to this theory of how rules can be acquired is that it should be much easier to acquire new norms when part of a community that endorses it, because the community, through verbal and emotional cues, will increase the sheer volume of evidence available to the norms acquisition system. For instance, if one becomes a member of a religious community later in life, because it is a community of people that shares similar norms (in some respects) it will be much easier to acquire those norms than attempting to do so alone.

If it is, therefore, possible to acquire a System 1 norm based on System 2 reasoning through these processes, then we have what we set out to give, *viz.*, a psychological account of reflective equilibrium, for we have shown how it is possible to reach consonance between theory and judgment in a way that takes rational theory-making seriously. Furthermore, if a rule is arrived at through a rational process, then it seems plausible that the judgments that proceed from that rule in the appropriate circumstances are themselves rational. That is, the judgments inherit the rationality of the rule. Furthermore, rules not acquired in this way, but endorsed by our best rational theories are thereby rationalized, that is, they are deemed rational because a belief, *however acquired*, counts as rational once it has been rationally justified. This does not mean that every moral judgment will, on reflection, be rational, only that commitment to rationally accepted rules is rational, and therefore it is a rational policy to adhere to them. Thus, moral cognitivism need not be construed as rationally arriving at each and every moral judgment, rather, it can be construed as adhering to rationally arrived-at rules. If we can consider moral cognitivism in these terms, then this psychological account of reflective equilibrium vindicates moral cognitivism.

V. Conclusion

In this paper I have shown how moral intuitions need not force us to the conclusion that moral judgments are largely arational, rather, that the best work in cognitive science and psychology are completely consistent with, and provide independent support for, the rational

⁷ Furthermore, according to Brown (1991), promise-making and promise-keeping are cultural universals (p. 139), so it is, at least in theory, possible for the process outlined here to be at work for all people. The upshot is that rationally grounded moral judgments can be a universal feature of the moral life, across cultures.

nature of moral judgments. I have answered objections from Haidt that reason cannot play any substantive role in the moral life, and showed how it might be possible for System 2 reasoning to interact with our System 1 intuitions in a way required by Rawls's theory of reflective equilibrium using the framework provided by the work of Sripada and Stich. Furthermore, since reflective equilibrium is a rational process, and the judgments it issues in are also rational, the psychological account of reflective equilibrium given in this paper is also a defense of moral cognitivism.

References:

- Brenner, Lyle; Koehler, Derek; and Rottenstreich, Yuval (2002), "Remarks on Support Theory: Recent Advances and Future Directions," in Gilovich, T, Griffin, D, and Kahneman, D (eds.), *Heuristics and Biases*. New York: Cambridge University Press.
- Brown, Donald (1991), *Human Universals*. Philadelphia: Temple University Press.
- Burge, Tyler (1993), "Content Preservation," *Phil Review* 102(4):457-488.
- Carruthers, Peter (ms), *The Architecture of Mind*.
- Chomsky, Noam (1988), *Language and the Problems of Knowledge*. Cambridge, MA: MIT Press.
- Cohen, Stephen (2004), *The Nature of Moral Reasoning*. New York: Oxford University Press.
- Crain, Stephen and Pietroski, Paul (2001), "Nature, Nurture, and Universal Grammar," *Linguistics and Philosophy* 24:139-186.
- Evans, J and Over, D. (1996), *Rationality and Reasoning*. Psychology Press.
- Frankish, Keith (2004), *Mind and Supermind*. Cambridge, UK: Cambridge University Press.
- Haidt, Jonathan (2001), "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment," *Psychology Review* 108(4):814-834.
- Haidt, Jonathan; and Hersch, M.A. (2001), "Sexual morality: The cultures and reasons of liberals and conservatives." *Journal of Applied Social Psychology* 31: 191-221.
- Hutcheson, Francis (1728), *Essay on the Nature and Conduct of the Passions With Illustrations on the Moral Sense*. Dublin.
- Kahneman, Daniel (2002), Maps of Bounded Rationality: A Perspective on Intuitive Judgment and Choice. Nobel laureate acceptance speech. Available at: <http://nobelprize.org/economics/laureates/2002/kahneman-lecture.html>. Last accessed December 27, 2005.
- Kelly, Daniel; Stich, Stephen; Haley, Kevin; Eng, Serena; Fessler, Daniel (forthcoming), "Harm, Affect, and the Moral/Conventional Distinction."
- Kuhl, Patricia (2000), "A New View of Language Acquisition," *Proceedings of the National Academy of Sciences* 97(22):11850-11857.
- Kuhl, PK; Conboy, BT; Padden, D; Nelson, T; Pruitt, J (2005), "Early Speech Perception and Later Language Development: Implications for the 'Critical Period,'" *Language and Learning Development* 1(3&4):237-264.
- Nucci, Larry (2001), *Education in the Moral Domain*. New York: Cambridge University Press.
- Rawls, John (1972), *A Theory of Justice*, Revised ed. Cambridge, MA: Harvard University Press.
- Saltzstein, Herbert and Kasachkoff, Tziporah (2004), "Haidt's Moral Intuitionist Theory: A Psychological and Philosophical Critique," *Review of General Psychology* 8(4):273-282.
- Shaftesbury, Lord (1711), *Characteristics of Men, Manners, Opinions, and Times*.

- Snow, Catherine and Hoefnagel-Höhle, Marian (1978), "The Critical Period for Language Acquisition," *Child Development* 49(4):1114-1128.
- Sripada, CS and Stich, Stephen (2006), "A Framework for the Psychology of Norms," in Carruthers, P., Laurence, S., and Stich, S. (eds.), *The Innate Mind: Culture and Cognition*. Oxford University Press.
- Stanovich, Keith (1999), *Who is Rational? Studies in Individual Differences in Reasoning*. Laurence Erlbaum.
- Stanovich, Keith; and West, Richard (2000), "Individual Differences in Reasoning: Implications for the Rationality Debate?" *Behavioral and Brain Sciences* 23:645-726.
- Sunstein, Cass (2005), "Moral Heuristics," *Behavioral and Brain Sciences* 28:531-573.
- Thomson, Judith Jarvis (1976), "Killing, Letting Die, and The Trolley Problem," *Monist* 59:204-217.
- Turiel, Elliot (1998), "The Development of Morality," in Damon, W, and Eisenberg, N (eds.) *Handbook of Child Psychology*, 5th ed.: Vol 3 Social, Emotional , and Personality Development. Hoboken, NJ: Wiley and Sons.
- Tversky, A and Kahneman, D (1983), "Extensional vs. Intuitional Reasoning: The Conjunction Fallacy in Probability Judgment," *Psychology Review* 90:293-315.
- Williams, Bernard (1981), "Persons, Character, and Morality," in Williams, Bernard *Moral Luck*. Cambridge: Cambridge University Press.