Reason and Intuition in the Moral Life: A Dual Process Account of Moral Justification
Leland F. Saunders

Are moral judgments rational? Many of us think so, and according to one influential research tradition in moral reasoning and moral judgment, they are (Piaget, 1932; Kohlberg et al., 1983; Kohlberg, 1984). On this view, moral judgments are the conclusions of deductions from consciously held moral principles, such as: Harm is morally bad; \( x \) is harmful; therefore, \( x \) is morally bad. Call this the *deductive model* of moral judgment. The deductive model of moral judgment is intended as both a normative theory and a theory about the psychological processes that underlie moral judgment. And notice that both components are necessary to show that moral judgments are rational. We need a normative theory of moral justification that shows that morality is a rational enterprise, and a psychological theory that shows that the normative model is actually implemented in human psychology. A normative theory whose requirements are not psychologically possible for humans cannot establish the rationality of our moral judgments, nor can a psychological theory of moral judgment apart from some normative theory that rationally justifies them.

According to the deductive model, the criteria for the rationality of moral judgments are their logical validity and their relationship to general moral principles. The deductive model has been a dominant psychological paradigm for research on human reasoning in a number of domains (for a review see Evans, 2002), including moral reasoning and judgment in the tradition of Piaget and Kohlberg. Moral intuitions, however, raise serious difficulties for the deductive model, because they show that moral
judgments can arise in ways ruled out by the deductive psychological theory. Moral intuitions are the sorts of easy and immediate moral judgments that we constantly make without any conscious effort and that strike us with a perception-like quality—characteristics that distinguish them from ordinary deductions that are comparatively slow and effortful. Moreover, unlike moral judgments in the deductive model, the sources of our moral intuitions are opaque, and we do not know what, if any, principles might underlie them. The very term “moral intuition” is, in fact, meant to distinguish them from the kind of reasoned moral judgment given by the deductive model. These phenomenological differences alone, however, do not show that the deductive model cannot explain moral intuitions, since they may be the result of very quick and perhaps enthymemetic deductions.

The real challenge comes from recent findings that show that moral intuitions can be “dumbfounded” (Haidt, 2001; Haidt and Hersch, 2001; Hauser, 2006; Hauser et al., 2007). I will explain the details of this research shortly, but in summary it indicates that moral intuitions are not always immediately responsive to reasons, since they can persist even after subjects recognize compelling reasons to modify or drop them. This is a puzzling result if moral intuitions are thought of as the conclusions of deductions from consciously held moral principles, because deductive conclusions are easily modifiable, especially if one of the premises is shown to be false or in need of modification, as is the case with moral dumbfounding experiments. Moral dumbfounding strongly suggests, then, that the deductive model cannot explain the rational basis of moral intuitions, and by extension it casts doubt on the rationality of morality as a whole.
The challenge, then, is to show how morality can be rational if moral intuitions are resistant to rational reflection. Again, there are two parts to this question, one normative and the other psychological. The normative problem is whether there is a model of moral justification that can show that morality is a rational enterprise given the facts of moral dumbfounding. Appealing to the model of reflective equilibrium for the rational justification of moral intuitions, I argue, fairly easily solves this problem. Reflective equilibrium views the rational justification of morality as a back-and-forth balancing between moral theory and moral intuition, and so does not require that individual moral intuitions be directly responsive to rational reflection. The psychological problem is whether human psychology actually implements the processes required for reflective equilibrium. The psychological problem is far more difficult, and will require appealing to a dual-process view of moral judgment that regards moral intuitions and moral theories as belonging to different mental systems. This allows us to explain the potential gap between moral intuition and moral theory demonstrated by moral dumbfounding, and more importantly, allows us to show how moral intuitions and moral theory can satisfy the requirements of reflective equilibrium.

Moral Dumbfounding and Rational Justification

A striking example of how moral intuitions can be dumbfounded is found in Haidt’s research using the following vignette:

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)

Most participants in Haidt’s research think not, and judge Mark and Julie’s actions to be seriously morally wrong. However, when pressed to explain why it is wrong, they are unable to give coherent reasons in support of their judgment. 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. The judgment of moral wrongness persists even after the participants find that there are no reasons supporting the judgment; a phenomenon Haidt labels ‘moral dumbfounding’. Though subjects are able to offer some reasons for their judgments, the reasons offered are often confabulated and lack coherence or relevance to the actual vignette—a finding supported by Hauser and colleagues (Hauser, 2006; Hauser et al., 2007).

These findings raise serious difficulties for the psychological theory of moral judgment posited by the deductive model. On the deductive model, even if the initial
judgment is arrived at through some unconscious or enthymematic reasoning process, the judgment would be responsive to conscious and explicit reasons, because we could consciously entertain the premises, recognize the error in our reasoning, and then modify our judgment accordingly. Yet, in the case of moral dumbfounding, people consciously recognize reasons against their initial judgment, but fail to modify it—something that should be psychologically impossible on the deductive model. Furthermore, if rational justification depends on there being a proper relationship between principles and judgment, then moral intuitions show it is not always psychologically possible to demonstrate the logical validity of that relationship, since we do not know what, if any, moral principles might underlie the moral intuition. The deductive model fails the psychological test, because it cannot provide a plausible psychological theory that implements its normative requirements. If moral judgments are rational, some other model of justification is needed.

One model of moral justification that can quite naturally account for moral dumbfounding is reflective equilibrium (Rawls, 1971). According to reflective equilibrium, we use our moral intuitions as a starting point for moral theorizing. From particular moral judgments we begin to theorize about what they have in common—for example, whether they point to any common wrong-making feature or principle. Moral theories are then constructed, guided by constraints of consistency, coherence, and completeness. Of course, not all moral intuitions will accord with the resultant theory, and when intuitions conflict a person will have to decide whether to stick with their theory and revise their intuitions, or to stick with their intuitions and revise their theory. There is no hard-and-fast rule about when to choose theory over intuition or vice versa,
but if the intuition is strongly held, and the change to the theory required to accommodate it is minimal, then it seems reasonable to modify the theory. Conversely, if the theory is general and wide-ranging, and the intuition only mildly held, it seems reasonable to revise the intuition to accord with the theory. The goal of all this back-and-forth is to achieve maximum consonance between theory and intuition—that is, a reflective equilibrium. Rawls argues that reflective equilibrium counts as rational justification because, “[Rational] justification is a matter of the mutual support of many considerations, of everything fitting together into one coherent view.” (Rawls, 1971, p. 502).

At first glance it may seem that reflective equilibrium faces the same problem with moral dumbfounding as the deductive model, since it too requires that moral intuitions be responsive to reasons to achieve equilibrium with theory. This is true, but a little too simple. Reflective equilibrium does not require that moral intuitions be immediately corrigible by reasoned reflection, rather it requires only that, over time, a person’s moral intuitions can be made to conform to their moral theory. This process of reflective modification and endorsement also provides rational justification for occurrent moral intuitions, because the moral intuitions that are modified or endorsed upon reflection inherit the rationality of the process. Thus, if someone reflectively modified or endorsed a moral intuition prohibiting incest in accordance with their moral theory, that intuition is rational, even if, on a given occasion, reasons for it cannot be produced. Moral dumbfounding, therefore, does not raise any difficulties for reflective equilibrium, because it does not require that we be able to articulate reasons for our moral intuitions or
that we be able to overrule them; only that they are justified through a rational process of modification and endorsement.

Reflective equilibrium is a plausible normative account of the rationality of moral judgments that is widely accepted (see, for example: Rawls, 1971; Smith, 1994; Daniels, 1996; Kamm, 2007), and I propose that we accept it as well. With our normative account of moral judgments now in place, we need to show how the process can be implemented in human psychology.

**Two-Systems and The Possibility of Reflective Equilibrium**

Part of what makes reflective equilibrium attractive as a normative theory of moral judgment is that it recognizes a potential gap between moral intuitions and a consciously held moral theory—a gap strikingly demonstrated by moral dumbfounding. Any plausible psychological account of reflective equilibrium needs to explain how that gap is possible. It is important to note that disparities between intuitions and consciously held theories occur with respect to many cognitive domains, not just morality. For example, people often make judgments about probabilities that are in direct conflict with consciously held norms of conjunction (Tversky and Kahneman, 1983). It is now a robust finding that humans regularly and persistently make judgments that violate their consciously held norms, even when they have all the resources to reason through a problem carefully and correctly (For a review see Brenner, Koehler & Rottenstreich 2002).

The potential disparity between intuitions and consciously held norms cannot be explained by a unified theory of judgment (Gilbert, 1999), and the theory of reflective equilibrium also assumes that there are two distinct methods for arriving at a moral
judgment—one that is quick, automatic, and nonconscious, and another that is slow, deliberate, and conscious. This fits nicely with dual-process theories of reasoning, according to which humans have two reasoning systems, distinguished by their functional characteristics. System 1 processes are fast, automatic, and nonconscious, whereas System 2 processes are slow, deliberate, and conscious (Evans and Over, 1996; Sloman, 1996, 2002; Stanovich, 1999; Kahneman, and Frederick, 2002; Wilson, 2002). Beyond this general description, there is little consensus on how best to draw the line between the two different kinds of systems. For example, some theorists have suggested that the two systems differ with respect to evolutionary age, arguing that System 1 is evolutionarily old and shared with other animals, whereas System 2 is evolutionarily recent and distinctively human (Evans and Over, 1996; Stanovich, 1999). There are good reasons to be skeptical of a cut-and-dried division of the two systems in terms of evolutionary age (Evans, 2008), and it is quite possible that some, if not all, of the processes underlying moral intuitions are evolutionarily recent and distinctively human (Hauser, 2006). However, nothing of what follows depends on adopting any particular view about the evolutionary age of the two systems.

Another common claim is that System 1 processes are associative, and System 2 processes are rule-based (Sloman, 1996, 2002; Kahneman and Frederick, 2002; Strack and Deutsch, 2004). But here too there is good reason for thinking that the characterization of the two systems on this basis is too simplistic, and that at least some System 1 operations are indeed rule-governed (Carruthers, 2006, this volume; Evans, 2008). For example, many researchers claim that System 1 processing involves heuristics, which are not merely associative operations, but rather ‘rules of thumb’ (Evans, 2008).
Some System 1 operations may indeed be associative, but not all, and, as we shall see, it is plausible to think that much moral reasoning involves processes of a broadly System 1 kind (fast, automatic, nonconscious) that are nonetheless rule-based.

However, even if the two systems can both employ rules, it is likely that the rules in question will differ in content, and that they are stored in separate memory systems—System 1 rules in an nonconscious implicit system, and System 2 rules in an explicit system, whose contents are available to consciousness (Smith and DeCoster, 2000; Evans and Over, 1996). The difference in memory systems also has consequences for learning. It takes very little time for System 2 to learn and apply a new rule, for example a rule of logic about how to manipulate symbols to produce valid conclusions, but it takes a great deal of time and repetition for System 1 processes to learn and apply a new rule, if ever. Another consequence of this division is that the two systems can come to different conclusions about the same situation because they draw upon different rules.

When the two systems reach conflicting conclusions, it is possible, in the right circumstances, for a System 1 intuition to be overridden by System 2 reasoning. Indeed, many dual process theorists characterize System 2 in terms of its role in scrutinizing System 1 intuitions, and either approving or overriding them (Kahneman and Frederick, 2002; Gilbert, 2002; Stanovich and West, 2000; Evans, this volume; Stanovich, this volume). System 2 is not always activated, however, especially if a highly plausible intuition is readily available (Kahneman and Frederick, 2002). Moreover, System 2 is easily disrupted by time pressures, multi-tasking, mood, and time of day (Bodenhausen, 1990; Finucane et al., 2000; Kahneman and Frederick, 2002). And, even when activated and not disrupted, System 2 cannot always override a strong System 1 intuition (Bargh,
1999; Wilson et al., 2000). A System 1 intuition can persist, even when people recognize good reasons to drop or modify it.

A dual process theory, then, can explain the gaps between moral intuitions and consciously held moral theories. We can think of moral intuitions, like any other kind of intuition, as System 1 judgments, and consciously and explicitly developed moral theories can be thought of as the outcomes of System 2 processes. This suggestion gains further support when we consider possible conflicts between moral intuitions and consciously derived moral judgments. If a moral intuition conflicts with a System 2 moral judgment, or if two moral intuitions conflict, this gives us reason to pause and think through the situation more carefully. These sorts of conflicts are characteristic triggers for reasoning through tricky moral situations, or for that matter even knowing that one is in a tricky moral situation (Cohen, 2004).

Furthermore, since dual process theories recognize that it is possible for System 2 to override a System 1 intuition, but that override is difficult and often fails, dual process theories provide an explanation for moral dumbfounding, where moral intuitions do not change in response to explicit reasoning about a situation (at least in the short term). So, moral dumbfounding does not present a problem for a dual process theory of moral judgment; on the contrary, a dual process theory of moral judgment explains the facts of moral dumbfounding quite naturally.

Now we need to show how a dual process theory of moral judgment can satisfy the requirement of reflective equilibrium that our moral intuitions can change to align with our consciously held moral theories. This requirement is much less demanding, in terms of a normative theory, than the deductive model, but it complicates the
psychological picture since we now need to show that System 2 reasoning can direct changes over time in the System 1 processes to produce moral intuitions that conform to it. This type of change in moral intuitions is different from simply overriding them—it changes the kinds of moral intuitions we are disposed to have by changing the norms employed by System 1. We have already seen that many dual process theories allow for the possibility of learning new rules, over time, by System 1 processes, but we now need to provide a model for how that is possible in the moral case. If we can do that, it will show how reflective equilibrium is psychologically possible for minds like ours. To do that, we first need some account of the psychological mechanisms that underlie moral intuitions; that is, a specification of the System 1 processes that give rise to moral intuitions.

**A Psychology for Moral Intuitions**

If the System 1 operations that give rise to moral intuitions are rule based, then moral intuitions must arise from the application of a set of internalized moral norms. Moral norms have been studied extensively, so there are some fairly robust findings that any psychological model of the System 1 processes underlying moral intuitions must be able to explain. For example, every human society has moral norms, but there is variation in the content of those norms (Brown, 1991). Children reliably acquire the prevailing norms of their community from a very early age, and can recognize specifically moral norms between the ages of three and five (Nucci, 2001; Turiel, 1983). Moral norms have a practical force, which is, perhaps, mediated by emotions, since emotions are somehow involved in, or implicated by, moral intuitions (Greene et al., 2001; Wheatley and Haidt,
2005). And lastly, as we have seen, moral intuitions can be dumbfounded (Haidt, 2001; Hauser, 2006; Hauser et al., 2007), indicating that at least some of our operative moral norms are not consciously accessible.

Sripada and Stich (2006) offer a plausible psychological model of moral intuitions that synthesizes these findings. Figure 1 shows the outlines of the system. In the figure solid lines indicate links for which there is very strong evidence, and the dotted lines indicate links that are more speculative though supported by some evidence.

[Insert Figure 1 around here].

According to this model, 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 inferring the norms of their community from them. Sripada and Stich offer no exhaustive list of what the appropriate environmental cues are, but there is evidence that at least some of them are direct verbal admonitions such as, “Don’t hit your sister,” and the observation of negative emotional responses in peers and caregivers (Edwards, 1987; Turiel, 1998). They also give no indication of how much evidence a child needs concerning the content of a rule before he or she acquires it, but it seems highly implausible that only one instance would be sufficient.

After a rule has been inferred, it 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 inputs representing the sorts of observable phenomena that count as instances of stealing. Although this example is given in terms of an imperative, Sripada and Stich make no claims about how the rule is represented in the norms database. They offer as possibilities that norms could be represented as prototypes, stereotypes, or narratives. It is also possible that they could be represented as heuristics (Sunstein, 2005). The database is also connected to the emotion system, which attaches an affective valence to moral judgments—judgments of norm violation have a negative valence, while judgments of norm satisfaction have a positive valence.

There are some limitations to the model that must be mentioned. First, it is incomplete. For example, there is evidence that judgments of intentionality influence our moral intuitions (Núñez and Harris, 1998), and that is not included in the Sripada and Stich model. Second, the model is general in scope. That is, the precise details of how the links between various components are accomplished, and what, if any intervening factors may exist are not given. These points are not meant as criticisms, because Sripada and Stich note that their model is a framework, not an exhaustive account, but it is important to recognize that these limitations exists. For our purposes, a plausible framework for the System 1 operations underlying moral intuitions is sufficient, and that is what the Sripada and Stich diagram provides.

Moreover, this framework suggests that the psychology of morality is consistent with a dual process architecture—a possibility that Sripada and Stich recognize, but do not develop.
We suspect that some version of the two sets of books hypothesis [that is, dual process theory] is correct… If the hypothesis is true, it would go a long way towards explaining the commonplace observation that while people do recognize inconsistencies in their moral beliefs and rationally revise certain of them, those changes are often superficial; automatic, intuitive reactions to real-world cases are still governed by the old, inconsistent norms (2006, p 297).

That is all Sripada and Stich say about this possibility, but I think we can do considerably more to explore this suggestion as a working hypothesis, and develop their model in terms of a dual process theory that could explain the possibility of reflective equilibrium. Our first step is to determine which parts of the diagram belong to System 1, and which parts belong to System 2. Figure 2 shows the likely division between the two systems on the Sripada and Stich framework.

[Insert Figure 2 around here.]

The acquisition mechanism most likely belongs to System 1, since young children acquire norms automatically as early as three to five years of age (Nucci, 2001; Turiel, 1983). 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. 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. Often it takes real philosophical work to figure out what rules might underlie such judgments, and often such theories are highly controversial. So, it is plausible that the left part of the diagram up to and including the initial judgment belong to System 1.

The System 2 portion of the diagram is explicit reasoning and *post hoc* justification. These operations are conscious and comparatively slow and effortful. Sripada and Stich leave some open questions about what role System 2 plays in moral judgments and in the moral life more generally. For example, they do not discuss whether System 2 reasoning can result in the modification or rejection of a System 1 intuition. It seems from our discussion of dual process models that it should be possible at least some of the time, but that it is difficult and often fails, resulting in moral dumbfounding. The only role they see as being well established by the evidence for System 2 reasoning is *post hoc* justification, that is, the practice of attempting to justify our moral intuitions by confabulating reasons for them, rather than attempting to derive a justification from an explicit moral theory. *Post hoc* justification and confabulation are not unique to the moral domain, and some dual-process theorists take the view that System 2 generally operates to provide reasons after the fact for System 1 intuitions (Evans & Over, 1996; Wegner 2002; Wilson, 2002; Stanovich, 2004).

Another possible function of System 2 reasoning that Sripada and Stich see as speculative is that it could have some effect on the contents of the System 1 norms database. This is the critical connection for our purposes, because it could provide the necessary psychological basis for reflective equilibrium. For suppose that we find that some set of our moral intuitions is seriously out of line with the requirements of our
moral theory. What we want is to make some System 1 change such that in the future, when confronted with situations that evoke these problematic moral intuitions, we instead come to have moral intuitions 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 effected, we would be required to overturn those wrong judgments by System 2 processes—a practice which, as we have seen, is effortful and prone to failure. Furthermore, constantly overturning System 1 judgments in this way would require significant cognitive resources—resources that may not always be available. Thus, if we want our moral judgments to be reliable, automatic, and effortless, then they must issue from a System 1 process, rather than a System 2 one. If the norms in the database can be changed by explicit reasoning to produce intuitions that conform to an explicitly held moral theory, then we will have shown how reflective equilibrium is psychologically possible.

One final point before doing that: Sripada and Stich intend the beliefs box in their diagram to represent explicit beliefs that we can consciously entertain and can manipulate by System 2 processes. It is important to note, however, that explicit reasoning is likely affected by implicit beliefs, which are stored in different memory systems than those beliefs accessible to explicit reasoning (Smith and DeCoster, 2000). Such implicit beliefs are not consciously accessible, and so do not directly enter into explicit reasoning, but they can influence how we frame and undertake deliberation, since they play a role in producing consciously accessible judgments through some System 1 processes (Evans, this volume). To preserve Sripada and Stich’s intent, the beliefs box should be thought of as belonging to System 2, since the content will draw on explicit memory, but it is
important to keep in mind that implicit beliefs, though they do not figure into explicit reasoning directly, can influence the directions explicit reasoning takes.

**Dual Process and Moral Justification**

In the preceding section, the term ‘norm’ was used rather loosely, but to be clear about what kind of effect System 2 reasoning can have on the norms database we need to distinguish between two senses of the term: one psychological and the other evaluative. In the psychological sense, the term ‘norm’ refers to an action-guiding rule that explains a person’s behavior and judgment. Norms, in this sense, are purely descriptive, making no claim as to whether the action-guiding rule is the right one, only that it is the one accepted by the agent. I shall refer to norms of this kind as ‘psychological norms.’ In the evaluative sense, norms are the standards by which we should act and judge. I shall refer to norms of this kind as ‘evaluative norms.’ Although psychological norms and evaluative norms can prescribe the same behavior, they can also diverge. It is quite possible to act according to a psychological norm that is contrary to an evaluative norm (See Bargh, 1999).

The goal of reflective equilibrium is the maximum consonance between psychological and evaluative norms, which requires, *inter alia*, the reflective endorsement or modification of psychological norms. It also requires that we engage in some process for elucidating the correct evaluative norms for morality, which is the goal of moral theory. From a dual process perspective, developing a moral theory is a System 2 operation, requiring slow, deliberate reasoning, and sometimes university level courses in ethics. It is not an easy thing to do, which is probably why many people never
construct anything close to a full-blown moral theory. Quick, intuitive moral judgments, on the other hand, reflect the agent’s psychological norms, and issue from System 1 operations. However, our psychological norms are not consciously accessible to us through introspection; we cannot simply look inside the norms database to find the data needed for our System 2 theorizing or to see whether there is a mismatch between our psychological and evaluative norms. So, it may seem that a dual process model of moral reasoning cannot satisfy the psychological requirements of reflective equilibrium.

That conclusion is too quick, because even though the contents of the norms database are not consciously available, the norms system does produce a plethora of consciously available data in the form of moral intuitions. Such intuitions will not take the form of universal rules, such as “Harm is bad,” but will rather be particular, and often take the form “This is bad.” But, from many instances of such particular judgments in similar situations, we can infer the underlying rules in our norms database—for example, that things involving harm are bad. Inferring these rules is a System 2 process, as is reasoning about the rules to build moral theories in accordance with our reflective judgments. But what effect, if any, can our System 2 moral theory have on our System 1 moral intuitions? Do we have any reason to think that our psychological norms are alterable?

Indeed there is, if we consider that the norms system is a learning system—that is, although the architecture of the norms system may be hardwired and universal, its content is not, but must be acquired. And it is the content that matters when it comes to the sorts of moral intuitions that are made. Thus, if the norms acquisition mechanism were to remain intact and active throughout a person’s life, then it would be possible to alter the
content of the norms database by acquiring new rules, including ones required by our normative theory. Once such rules were acquired, the norms system would make judgments consistent with the requirements of our theory.

There are some similarities here with language, which is the paradigm of a learning system, that could illuminate how it is possible for System 2 reasoning to alter System 1 rules, since it is possible to acquire a second language that becomes just as natural and easy to use as the first. Moreover, just like the norms system, the architecture of the language system is universal and hardwired, but the lexical content must be acquired from the environment (Chomsky, 1988; Crain and Pietroski, 2001; Laurence and Margolis, 2001), which occurs early in development, usually by around age 3 (Kuhl, 2004). Proper language acquisition also needs to take place within a critical period early in life, and failure to be exposed to the proper stimuli within that period means that competency can only be acquired with difficulty (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, perhaps because of the neural pruning that occurs around three to four years-old and then again in early adolescence (Webb et al., 2001). 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 it, they are able to “think” in that language. They are able to “switch” between 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, there is some learning
mechanism that allows the explicitly studied rules of grammar to modify a System 1 language process, though it takes a significant amount of work to accomplish.

Can an analogous process be at work in the norms acquisition system? If so, it would predict that an evaluative norm endorsed by System 2 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. There may be one important disanalogy, though, between adult language acquisition and adult norm acquisition. In the case of acquiring a second language, it is possible that there is a different acquisition mechanism in adults than in children (See, for example, Dornyei, 2005). While the same may be the case in acquiring new norms as adults, it is more likely that the same norms acquisition mechanism remains from childhood to adulthood, though possibly attenuated.

Among the reasons for thinking that the norms acquisition mechanism remains intact throughout a person’s life, and that it is, perhaps, even more flexible than the language system is that norms can change within a society during one’s lifetime, and different norms can prevail in different social environments. For example, some behaviors acceptable when one is an undergraduate are not acceptable as a professional. Also, when one becomes a member of a subgroup, it is expected that one will change one’s norms to those of the group—an expectation people would not have if it were not possible to change one’s norms. 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 be necessary for survival. It would be a short life for a captured woman who habitually and consistently breached the norms of her new community. It might be objected that simply being able to learn instrumental force-backed rules such as “If I disobey my new master, I shall be punished,” would be enough to ensure survival, but there would be significant benefits to adopting the new norms. One possible benefit is integration into that society. These women, and their children, would be more likely to become fully integrated members of their new society if they adopted the norms of the community, rather than retaining their previous norms and doing just enough to stay alive. These advantages would increase if the rule could be adopted by a System 1 process, since that would produce quicker, more natural reactions with the appropriate emotional valence without the heavy cognitive load that would be required if the rule were adopted only by System 2.

Assuming the norms 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. It should be noted, though, that it would also be possible to acquire a new System 1 norm without System 2 involvement. For example, if a person joins a new community, she may acquire the norms of her community unconsciously. However, these cases are not interesting for the possibility of reflective equilibrium, which requires a conscious, rational process for justifying moral theories and beliefs. What we are looking for is some way for our System 2 reasoning to exploit, perhaps indirectly, the norms acquisition mechanism. At this point, work done by Frankish (2004) on belief may help. He argues that the best way
to understand the differences between various kinds of beliefs, such as occurrent (beliefs you are currently entertaining) versus dispositional (beliefs you are disposed to assent to in the right circumstances) is to adopt a two-strand theory of belief that treats them as different kinds (“basic beliefs” vs. “superbeliefs”). According to Frankish, basic beliefs are those that influence our behavior automatically and unconsciously, while superbeliefs are commitments to reason and act in the future as if the proposition accepted were the case, even if our basic beliefs remain unchanged. While Frankish does not make an explicit link between his two-strand theory and dual process theories, the characteristics he describes maps quite easily onto it, with basic beliefs being implicit System 1 beliefs and superbeliefs being explicit System 2 beliefs. On this view, our System 2 beliefs cannot have a direct effect on our System 1 beliefs. Suppose we have a System 1 belief that God exists, but form a System 2 belief that God does not exist. Then what we have really done, on this view, is commit ourselves to reasoning and acting in the future as if God does not exist, even though we continue to have a System 1 belief that God does exist.

Applying this to the moral case, when we develop our moral theory we settle on a number of evaluative norms, and even if they have not been acquired by System 1 operations, we can commit ourselves to reasoning and acting in the future as if our evaluative norms are correct. And there may be a good psychological motivation for forming such moral commitments, perhaps due to a System 1 belief-desire pair—a desire to commit to sound moral principles and a belief that the moral principle in question is a sound one. Or, perhaps simply a System 1 desire to be rational as defined across the whole person (Stanovich, 2004; Over and Saunders, this volume).
So, the first part of acquiring a new moral belief via this reflective route is forming a commitment to act in the future as if we had that moral belief such that even if our initial judgment is \( \sim p \), we nevertheless act as though \( p \) were true, since we want to follow through on our commitment. Obviously, this is a case of serious cognitive dissonance, and a situation that seems 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. Perhaps this is just one way of being akratic. This seems to me fairly common, and we have a number of ways to excuse ourselves from the requirements of our theory, such as, “It’s just the way I am;” or “It’s just an ideal, I never really expected to live up to it.” Recall from the Sripada and Stich quote above, it is just these sorts of excuses that they think a dual process theory of moral judgment should explain.

Yet, at other times, we really do succeed in living up to the requirements of our theory, even when it conflicts with our original intuitions, and over time our intuitions themselves begin to change. Here is how it might happen: when we make a judgment that conflicts with our second-order commitment, we attempt to override our initial judgment, and as we do so, we verbalize our commitment to ourselves, either vocally or silently. This self-verbalization is important, as it provides System 1 with new perceptual inputs in the form of natural language sentences that it can then manipulate in accordance with its own subconscious processing and inferential rules (Carruthers, 2006, this volume; Frankish, 2004) So, our self-verbalization of a consciously held moral rule would provide
evidence to the System 1 norms acquisition mechanism that there is a new rule in the environment that needs to be inferred and added to the norms database. That is, this practised commitment itself becomes part of the moral environment whereby rules are inferred 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 it to ourselves in a natural language, in effect, lecturing ourselves, as our peers or parents might lecture us about certain norms. As already noted, admonitions expressed in natural language are among the evidences available to the norms acquisition mechanism when inferring a rule from the environment. So, as we repeat this commitment to ourselves in cycles of moral judgment, the self-verbalization of our commitment serves as evidence to the norms acquisition mechanism that there is a new norm that must be added to those already in the database. This process is outlined in Figure 3.

Thus, the new rule is acquired through the same mechanism that underlies the acquisition of all other rules, the only difference being that the initial evidence that a rule is in play comes from the self-verbalization of that rule, rather than the verbalization of the rule from others. Regardless of its origin, however, the rules acquisition mechanism takes that rule, and through its own operations, places it in the rules database. Once the new rule is acquired, System 1 judgments will be made in accordance with it, just as with any other rule. And this explains how rules in the norms database can be changed, over time, to conform to the requirements of an explicitly held theory.

[Insert Figure 3.]
There is a limitation to this method of acquiring new rules. According to work by Nucci (2001), rules are not always explicitly stated to children, and so linguistic utterances may not be the most effective type of evidence for norm acquisition. Instead, children usually acquire rules by observing behaviors, especially emotional responses of approval, disapproval, and disgust. So, children between three- and five-years-old are probably incapable of acquiring new rules in this reflective fashion, since they are insensitive to explicitly stated rules. However, this fact also points to another possible route for reflective rule acquisition that relies on admonishment and emotional cues.

On this route, we develop a set of System 2 rules and commit ourselves to reasoning and acting as though they 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 as we do so 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 like one possible formulation of the promise-keeping rule, which is most likely a System 1 rule, given its ubiquity among cultures and early appearance in childhood morality (Brown, 1991). The suggestion here is that commitments to act according to some System 2 moral rule are like promises, albeit promises we make to ourselves, but promises nonetheless. In fact, we often talk of having promised ourselves to do or refrain from some behavior. For example, we promise ourselves to exercise more, or never to take seconds on dessert, and it seems that we can make promises or commitments to ourselves in inner-speech, just as we can make them to other people in overt speech. For psychological purposes, both have the same status, and so, too, do our
commitments to explicit System 2 moral beliefs: they are promises made to ourselves in episodes of inner-speech.

This has important consequences, namely, that when we fail to live up to our own moral commitments, our System 1 norms system interprets our behavior as an occasion where a promise has been broken, and this produces the judgment: “You broke your promise! That is wrong.” This creates a conflict of judgment, because the System 1 rule that produced the initial judgment determined that the action was permissible, but now, through breaking a commitment, there is a second System 1 judgment that something wrong has been done. However, the judgment that we have done something morally impermissible, that is, 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 of why we disapprove of our behavior, because we feel bad about not doing what we committed to, and say things to ourselves such as, “I can’t believe I did that, when I told myself I shouldn’t do it.” Indeed, 18th Century British moralists thought that this feeling of self-disapproval was the very heart of moral judgment (See Hutcheson, 1728; Shaftsbury, 1711). The importance of self-disapproval for our purposes is that it provides evidence to the rules acquisition system in the form of an emotional response, not just self-verbalization.

However, this process does not rely entirely on negative feedback in the form of emotional disapproval. If we act in accordance 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, these negative and positive emotional responses to the breaking or adhering to the rule, along with self-verbalization,
provides the norms acquisition mechanism with enough evidence to infer the correct norm, and insert it in the rules database. Once that happens, System 1 judgments proceed in the normal way by reference to that rule. Furthermore, it seems that promise making and promise keeping are cultural universals (Brown, 1991), so this mechanism of rule acquisition is likely to be universally available. The upshot is that reflective equilibrium can be a universal feature of the moral life, across cultures.

The process of reflective rule acquisition sketched above fulfills both predictions made earlier in this section about modifying psychological norms. The first was that changing intuitions to accord with an explicit theory takes time and practice, and now we can see why. A single cycle of self-verbalization or self-judgment is not sufficient to acquire the new rule, so reflective equilibrium is not something that can be achieved immediately. The second prediction was that during the acquisition period, the new norm would be applied inconsistently, and again this model explains why. The process requires that people consciously entertain their System 2 moral beliefs on those occasions when System 1 produces a judgment that conflicts with them. If the System 2 moral belief is temporally forgotten, it will not trigger episodes of self-verbalization and self-judgment. Since people often fail to recall their commitments, whether to themselves or to others, when a rule is being acquired it will be applied only imperfectly, in starts and stutters, until it is acquired by the System 1 norms system.

Moreover, the process of reflective rule acquisition can account for moral dumbfounding. There is nothing in the process that requires any occurrent moral judgment to be overturned by System 2 processes. There may be some moral intuitions that are so strong—perhaps because of some connection to powerful emotions like
disgust—that they persist even when they are no longer warranted. That is perfectly consistent with the process outlined here. The System 1 intuition is corrigeable by explicit reasoning, but it will take time and practice, and in the meantime the powerful moral intuition will persist.

A final corollary to this theory of how System 1 rules can be acquired is that it should be much easier to acquire new norms when one is part of a community that endorses them, because the community, through verbal and emotional cues, will increase the volume of evidence available to the norms acquisition system. For instance, if one becomes a member of a religious community later in life it will be much easier to acquire those norms than attempting to do so alone. Not only are the external cues increased, but being part of a community will most likely increase how much attention one pays to the explicitly endorsed theory, so there will be fewer occasions when the commitment is overlooked. There would thus be an increase in both external and internal cues, allowing the norms acquisition mechanism to infer the correct norm much more quickly, with less practice. This is shown in Figure 4.

[Insert Figure 4.]

One complication, however, needs to be addressed before concluding: if the newly acquired System 1 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. This would not satisfy the requirements of reflective equilibrium, which requires a consonance between
theory and intuition, not merely consonance with theory and one of many intuitions. So, to realize reflective equilibrium fully there must be some way for previously held System 1 norms to cease to issue in intuitions. There are two possibilities here. One is that norms are deleted from the database. That is, once a new rule has been added that conflicts with an existing rule, the old rule is simply erased, and so there is no possibility of conflict. There are reasons to be skeptical of this account. It is unclear that the inferential capacities of the norms database are sensitive to conflicts of this sort, because potential conflicts between norms are fairly common—for example, breaking a promise to meet a colleague for lunch in order to save a drowning child. It is unlikely, therefore, that the inferential mechanism of the norms database will simply delete norms that could potentially conflict.

Another option is that the old norm is functionally erased; that is, it remains in the norms database but it is no longer used to produce moral intuitions. Recall that one role of the norms database in the Sripada and Stich framework is to connect norms to sets of perceptual inputs that would invoke them. So, a norm against lying would be connected to many different sets of perceptual inputs indicative of lying, and when any one set is present, the norm is invoked. This would obviously be a very large and diverse collection of sets, and Sripada and Stich do not provide details of how the norms database is connected to perceptual inputs, nor of how the inputs would trigger the activation of a norm to produce a moral intuition. However, it is reasonable to think that for every set of perceptual inputs, there is only one norm with which it is uniquely linked, since if identical inputs triggered more than one norm it would be hard to see how the norms system could fulfill its action-guiding role, as it would issue conflicting advice. It is
likely, then, that the inferential mechanisms of the norms database would be sensitive to this kind of conflict, and, when it occurred, would map the inputs to the norm for which there is continuing evidence in the environment, and remove the mappings from the one for which there is not such evidence. So, when a norm is being acquired the environmental evidence available to the norms acquisition mechanism shows that the new rule governs the current social landscape, and so gives it priority when mapping perceptual inputs. The old norm remains, but it is superseded, and thereby functionally erased, because it is no longer mapped to perceptual inputs that would activate it.

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, namely, a psychological account of reflective equilibrium, for we have shown how it is possible to reach consonance between theory and judgment. And, if reflective equilibrium is psychologically possible for minds like ours, then we have shown how moral judgments can be rational, since reflective equilibrium provides a normative theory of moral justification. Thus, moral intuitions and the facts of moral dumbfounding do not force us to the conclusion that morality cannot be rationally justified.

Acknowledgements

I would like to thank Peter Carruthers and Keith Frankish for their detailed comments on several drafts of this chapter, which were extraordinarily helpful, and no doubt made it much better. I would also like to thank Jonathan Evans, David Over, and Clare Saunders for their comments.
References


Shaftesbury (1711/[2004]), *Characteristics of Men, Manners, Opinions, and Times Part Three*. Kessinger Press, Whitefish MT.


Figure 1. Reprinted from Sripada and Stich (2006).
Figure 2.

System 1

Acquisition Mechanism

infer contents of normative rules

identify norm implicating behavior

Proximal Cues in Environment

Execution Mechanism

rule-related reasoning capacity

norm data base

r1

r2

r3

......

rn

compliance motivation

emotion system

punitive motivation

System 2

other emotion triggers

beliefs

judgment

explicit reasoning

post-hoc justification
Figure 3.

Acquisition Mechanism

Identify norm implicating behavior

Infer contents of normative rules

Execution Mechanism

norm data base

$r_1$

$r_2$

$r_3$

$r_4$ new rule

$r_n$

Rule-related reasoning capacity

compliance motivation

punitive motivation

emotion system

beliefs/commitments

explicit reasoning

judgment

other emotion triggers

Self-verbalization

Proximal Cues in Environment
Figure 4. Acquisition of norms