Two Theories about the Cognitive Architecture Underlying Morality

Daniel Kelly∗ & Stephen Stich§
Department of Philosophy, Rutgers University

Abstract

In this paper we compare two theories about the cognitive architecture underlying morality. One theory, proposed by Sripada and Stich (forthcoming), posits an interlocking set of innate mechanisms that internalize moral norms from the surrounding community and generate intrinsic motivation to comply with these norms and to punish violators. The other theory, which we call the M/C model was suggested by the widely discussed and influential work of Elliott Turiel, Larry Nucci and others on the “moral/conventional task”. This theory posits two distinct mental domains, the moral and the conventional, each of which gives rise to a characteristic suite of judgments about rules in that domain and about transgressions of those rules. We give an overview of both theories and of the data each was designed to explain. We go on to consider a growing body of evidence that suggests the M/C model is mistaken. That same evidence, however, is consistent with the Sripada and Stich theory. Thus, we conclude that the M/C model does not pose a serious challenge for the Sripada and Stich theory.

1. Introduction

In recent years, many cognitive scientists and empirically oriented philosophers have turned their attention to questions about morality.1 Among the issues that have been actively discussed are the nature of the cognitive mechanisms subserving various aspects of moral cognition, and whether or to what extent those mechanisms are innately specified (Dwyer 1999, forthcoming; Greene & Haidt 2002; Haidt 2001; Hauser forthcoming; Nichols 2004; Prinz forthcoming; Sripada & Stich forthcoming). In this paper we will compare two accounts of the cognitive architecture underlying morality. The first of these, which was proposed by Sripada and Stich (forthcoming) posits an interlocking set of innate mechanisms that underlie the acquisition of moral norms from the surrounding community and the generation of characteristic motivations to comply with those norms and to punish others who violate them. In section 2 we’ll give a brief sketch of the Sripada and Stich (S&S) model.

∗ dankelly@rci.rutgers.edu  § stich@ruccs.rutgers.edu

1 For overviews of this work, see Doris & Stich (2005 & forthcoming).
The second account has a more complicated provenance. For the past twenty-five years, some of the most influential work in moral psychology has been aimed at exploring and explaining the distinction between moral and conventional rules. Inspired by the pioneering work of Elliot Turiel, researchers in this tradition have published over sixty papers in which they investigate the emergence of the distinction in children and study its contours in an impressive range of subject populations. In section 3, we’ll present an overview of this research and some of the important conclusions that have been drawn from it. Researchers in this tradition have devoted relatively little effort to proposing explicit accounts of the psychological mechanisms and processes that underlie people’s ability to draw the moral / conventional distinction. So, in section 4, we will suggest one sort of psychological model that might be posited to explain the experimental results described in section 3 and the conclusions drawn from them. That model, which we’ll call the M/C model, is dramatically different from the S&S model and, as we will argue in section 4, the two models lead to very different predictions. Since it promises to explain a vast array of empirical findings, the M/C model is also, arguably, the best supported competitor to the S&S theory.

In section 5, our stance turns critical. Though there are many studies compatible with the conclusions about the moral / conventional distinction assembled in section 3, we believe there is mounting evidence that points in the other direction, suggesting that those conclusions are in fact false and thus that the M/C model, which is designed to explain those conclusions, is untenable. However, as we’ll argue in section 5, this evidence is all comfortably compatible with the S&S model. So the conclusion for which we’ll be arguing is that the M/C model does not pose a serious challenge to the S&S theory.

2. The S&S Theory of the Psychological Mechanisms Underlying Norms

Norms are a ubiquitous and important element of morality and of social life in general. In “A Framework for the Psychology of Norms,” Sripada and Stich (forthcoming) offer a theory about the innate cognitive architecture that gives rise to many of the individual and social level facts about norms. In this section we’ll begin by recounting some of those facts. We’ll then sketch some of the central elements of the S&S model, focusing on those that are most important when comparing the S&S model with the M/C model.2

S&S argue that norms are a theoretically important class of behavior regulating social rules characterized by the following features:

---

2 This paper is available on-line at: [http://www.rci.rutgers.edu/~stich/Publications/publications2.htm](http://www.rci.rutgers.edu/~stich/Publications/publications2.htm).

3 For further details along with an extended discussion of the evidence supporting the empirical claims made in this section, which is drawn from a number of different disciplines, see Sripada & Stich (forthcoming).
• **Independent normativity**: norms are rules which specify behaviors that are required or forbidden independently of any legal or social institution or authority, though of course some norms are also enforced by laws or other social institutions.

• **Punishment supported stability**: violations of norms result in a variety of punitive attitudes – including anger, condemnation and blame – directed at rule violators, and these attitudes sometimes lead to punitive behavior; the presence of these punitive attitudes in members of the community contributes to a norm’s long term stability.

• **Universal presence**: all human societies have norms and sanctions for norm violations; this includes human groups that have been in long standing isolation from other groups.

• **Ubiquity and Importance**: in virtually all societies, norms regulate a vast array of day-to-day behaviors, including behavior in a large number of quite important domains such as social exchange, status relationships, sexual behavior, mate choice, diet, and a host of others.

• **Reliable pattern of ontogenesis**: all normal children appear to have knowledge of some norms by the age of 3-5, and much of the cross-cultural diversity of normative rules among adults in different societies is already present and stable by the age of 9.

• **Cultural conformity**: children typically acquire the normative rules which prevail in their cultural group, regardless of their own biological heritage.

• **Substantial cross-cultural diversity**: the specific behaviors required or forbidden by norms vary dramatically from culture to culture.

Together, these last two features of norms, cultural conformity and substantial cross-cultural diversity, strongly suggest that norm development is significantly culturally determined.

Another important pair of properties of norms are the *motivational effects* they have on agents. Philosophers have long emphasized that from a subjective perspective, norms present themselves with a unique kind of authority that differs from standard instrumental motivation. Sripada and Stich argue that this philosophic tradition is largely correct. More specifically, they maintain that norms generate robust and reliable motivation to *comply* with norms and to *punish* those who violate them. Moreover, this motivation does not depend on the agent’s beliefs about the social or personal consequences of compliance or non-compliance.

Let’s now consider what sort of psychological architecture might explain the features of norms that we’ve assembled. The facts that norms are universally present in all societies, that they differ dramatically from one society to another, and that they
exhibit a reliable pattern of ontogenesis, suggest the existence of innate mechanisms dedicated to norm acquisition. The function of these mechanisms is to locate and internalize the norms prevailing in the surrounding society. Once a normative rule is acquired, it gives rise to reliable and robust intrinsic motivation to comply with the norm and to punish those who violate it. It is worth emphasizing that this pair of motivations sharply distinguish norms from other rules or information that may be mentally represented elsewhere in an agent’s cognitive system. This suggests that norm utilization is subserved by its own, dedicated “execution” mechanism, and that this mechanism, too, is innate. Thus a first pass at characterizing the psychological architecture subserving the acquisition and utilization of norms might looks like the system labeled with black type in Figure 1.

4 Figure 1, we should stress, is only a first pass. In the last section of their paper, S&S develop a much more complicated model, aimed at accommodating a significantly larger collection of empirical findings. We focus on the simplified model in Figure 1 because it makes it easier to see the differences between S&S’s model and the M/C model that we’ll elaborate in section 4.
enforced normative rule prevails in the local cultural environment, inferring the content of the rule, and passing that information on to other cognitive mechanisms for storage and utilization. On the S&S account, the acquisition mechanism operates automatically – a person does not decide to turn it on and cannot decide to turn it off, though it may be the case that the acquisition mechanism gradually turns itself off starting at some point late in adolescence. The mechanism for executing norms performs a set of functions that includes maintaining a data base of the normative rules that were identified and passed along by the acquisition mechanism, generating intrinsic motivation to comply with those rules, detecting violations of the rules, and generating intrinsic motivation to punish the violators.

Of course, people also accept and follow many behavior governing rules that they do not treat as norms. The motivation for following these rules varies, and can include considerations of prudence, fear of social sanctions, and a variety of other factors. These rules, it is plausible to assume, are stored and executed by a variety of different mental mechanisms, represented by the black boxes in the lower right of Figure 1. What distinguishes this heterogeneous set of rules from norms, according to the S&S theory, is that they are not acquired by the innate norm acquisition mechanism and they do not automatically engender either the compliance motivation or the punitive motivation associated with norms.

It is important to note that the architecture depicted in Figure 1 allows considerable variation with respect to the sorts of rules that the norm system can acquire and the sorts of punishments these rules can motivate. The normative rule data base can contain rules governing a wide variety of behaviors including harming others, sexual practices, food preparation and consumption, burial rituals, and so on. Moreover, rules can include information about the people to whom they apply, and different rules can apply to different groups of people. Some might apply to everyone, while others might apply only to more narrowly circumscribed groups like adult women, unmarried men, or members of a specific religion or caste, or even menstruating women in one’s own tribe or village. And while all rule violations lead to punitive attitudes, the rules themselves can specify how serious a transgression is and what sorts of punitive behavior is appropriate.

3. An Overview of Research on the Moral / Conventional Distinction

Common sense sanctions a vague but intuitively appealing distinction between two quite different sorts of rules that govern behavior, namely moral rules and conventional rules. On the one hand, prototypical examples of moral rules include those prohibiting killing or injuring other people, stealing their property, or breaking promises. On the other hand, prototypical examples of conventional rules include those prohibiting wearing gender-inappropriate clothing (e.g. men wearing dresses), licking one’s plate at the dinner table, and talking in an elementary school classroom when one has not been

---

5 See, however, Sripada and Stich (forthcoming) section 5.6 for a discussion of the various ways in which the contents of the data base might be constrained or biased.
called on by the teacher. This intuitive difference has caught the attention of philosophers of various orientations. Many have attempted to clarify the distinction, some by specifying those features that are distinctive of moral rules (Mill 1863; Rawls 1971; Gewirth 1978; Dworkin 1978), others by giving an account of systems of conventions and the rules that are embedded within them (Lewis 1969, Searle 1995). Despite (or perhaps due to) the wide range of approaches philosophers have taken to this issue, no single account has been widely accepted.

Psychologists have taken an interest in the distinction as well. Starting in the mid-1970s, a number of developmental psychologists, following the lead of Elliott Turiel, have offered their own characterization(s) of the intuitive distinction between moral and conventional rules. Moreover, they have gone on to argue that the distinction as they characterize it is both psychologically real and psychologically important (Turiel 1979; Turiel 1983; Turiel et al. 1987; Smetana 1993; Nucci 2001). Let us start with the proposed characterization of the distinction. Though the details have varied over time and from one author to another, the core ideas that researchers in this tradition have advanced about moral rules are as follows:

- Moral rules have an objective, prescriptive force; they are not dependent on the authority of any individual or institution.
- Moral rules hold generally, not just locally; they not only proscribe behavior here and now, they also proscribe behavior in other countries and at other times in history.
- Violations of moral rules involve a victim who has been harmed, whose rights have been violated, or who has been subject to an injustice.
- Violations of moral rules are typically more serious than violations of conventional rules.

By contrast, the following are the core features of conventional rules according to the account proposed by researchers in this tradition:

- Conventional rules are arbitrary, situation-dependent rules that facilitate social coordination and organization; they do not have an objective, prescriptive force, and they can be suspended or changed by an appropriate authoritative individual or institution.
- Conventional rules are often local; the conventional rules that are applicable in one community often will not apply in other communities or at other times in history.
- Violations of conventional rules do not involve a victim who has been harmed, whose rights have been violated, or who has been subject to an injustice.
Violations of conventional rules are typically less serious than violations of moral rules.  

Having offered a characterization of the distinction between moral and conventional rules, Turiel and his associates then set about developing an experimental paradigm to explore the psychological status of the distinction they had described. Experiments were designed to test the hypothesis that the moral / conventional distinction, characterized in this way, is both psychologically real and psychologically important. In these experiments (employing what has come to be called the “moral / conventional task”), subjects are presented with examples of transgressions of both prototypical moral rules and prototypical conventional rules, and are then asked a series of probe questions. These questions are designed to elicit subjects’ judgments about the transgressions along a number of significant dimensions, often called “criteria”. More specifically, “criterion judgments” were elicited from subjects to determine the following:

(i) whether the subjects consider the transgressive action to be wrong, and if so, how serious it is;

(ii) whether the subjects think that the wrongness of the transgression is “authority dependent,” i.e. does it depend on the existence of a socially sanctioned rule or the on the pronouncement or endorsement of an authority figure? (For example, a subject who has said that a specific rule-violating act is wrong might be asked: “What if the teacher said there is no rule in this school about [that sort of rule violating act], would it be right to do it then?);

(iii) whether the subjects think the rule is general in scope; is it applicable to everyone, everywhere, or just to a limited range of people, in a restricted set of circumstances?

(iv) how the subjects would justify the rule; in justifying the rule, do subjects invoke harm, justice, or rights, or do they invoke the fact that the rule prevails locally and/or that it fosters the smooth running of some social organization?

Results from the initial experiments using this paradigm supported the claim that the moral / conventional distinction, as characterized by Turiel and his associates, is indeed psychologically significant. They indicated that subjects’ responses to prototypical moral and conventional transgressions differed systematically, and in just the way suggested by the characterization given above (Nucci & Turiel 1978; Smetana 1981; Nucci & Nucci 1982). More specifically, transgressions of prototypical moral rules

---

6 Although there seems to be general agreement that violations of moral rules are typically less serious than violations of conventional rules, some authors downplay the importance of seriousness in their formal characterization the moral / conventional distinction. For example, Smetana (1993: 117) maintains that “severity of the transgression is not considered to be a formal criterion for distinguishing moral and conventional rules and transgressions.”
(almost always involving a victim who has clearly been harmed) were judged to be wrong and to be more serious than transgressions of prototypical conventional rules, the wrongness of the transgression was judged not to be “authority dependent,” the violated rule was judged to be general in scope, and these judgments were justified by appeal to harm, justice or rights. Subjects judged transgressions of prototypical conventional rules quite differently. They were judged to be wrong but usually less serious, the rules themselves were judged to be authority dependent and not general in scope, and the judgments were not justified by appeal to harm, justice, and rights. Adding to the case that the distinction thus characterized is psychologically real was the fact that the pattern of replies appeared to be quite robust. The pattern was not significantly affected, for instance, by the way in which transgressions were presented to subjects, the wording of the questions, or the order in which the questions were asked.

Supporting the contention that this pattern of results – along with the moral / conventional distinction as characterized by Turiel and his followers – is psychologically important is the prevalence of the pattern across a wide range of subject populations. During the last twenty-five years, the same pattern reported in the initial studies has been found in an impressively diverse set of subjects ranging in age from toddlers (as young as three and a half years) to adults, with a substantial array of different nationalities and religions. The pattern has also been found in children with a variety of cognitive and developmental abnormalities, including autism (Blair 1996; Blair et al. 2001; Nucci & Herman 1982; Smetana et al. 1984; Smetana et al. 1999). The pattern is notably absent, however, in both psychopaths and in children exhibiting psychopathic tendencies (Blair 1995; 1997). While many researchers see significance in this latter finding, no single explanation yet enjoys a consensus.

This large and prima facie striking set of experimental results seems laden with psychological implications. So it is hardly surprising that researchers in the moral / conventional tradition have drawn ambitious conclusions from their work. Here again the details of those conclusions have varied over time and from one author to another and, unfortunately, some of the crucial notions appealed to in those conclusions have not been explained as carefully as one might like. Nevertheless, it is clear that a majority of investigators in this research tradition would likely endorse something like the following collection of conclusions:

(C-1) The Clustering of Criterion Judgments: In moral / conventional task experiments subjects typically exhibit one of two signature response patterns. In the first signature pattern rules are judged to be authority independent and general in scope; violations are wrong and typically judged to be serious, and judgments are justified by appeal to harm, justice and rights. We call this the signature

---

7 For a study that included 3½ year old children, see Smetana & Braeges 1990. Among the cultural and religious groups studied were Chinese preschoolers (Yau and Smetana 2003), Korean children (Song et al. 1987), Ijo children in Nigeria (Hollos et al. 1986), Virgin Islander children, teens, and adults (Nucci et al. 1983), Roman Catholic high school and university students (Nucci 1985), Amish and Mennonite children and teens, and Dutch Reformed Calvinist children and teens (Nucci & Turiel 1993). For reviews, see Smetana 1993; Tisak 1995; Nucci 2001.
moral pattern. In the second signature pattern, rules are judged to be authority dependent and not general in scope; violations are wrong but usually less serious, and judgments are not justified by appeal to harm, justice, or rights. We call this the signature conventional pattern. Moreover, these signature response patterns are what philosophers of science sometimes call “nomological clusters” – there is a strong (“lawlike”) tendency for the members of the cluster to occur together.

(C-2) Response Patterns and Transgression Types: Not only do criterion judgments cluster into two distinct response patterns, but each pattern is reliably evoked by a certain type of transgression. Specifically, (a) transgressions involving harm, justice, or rights evoke the signature moral pattern, while (b) transgressions that do not involve harm, justice, or rights evoke the signature conventional pattern.

(C-3) Universality: The regularities described in (C-1) and (C-2) are pan-cultural, and they emerge quite early in development.

4. Explaining the Results: The M/C Model

As we noted in the Introduction, we are skeptical about these conclusions, but in this section, we propose to bracket that skepticism. Instead, we will assume that (C-1), (C-2) and (C-3) are true and ask: what sort of cognitive architecture could explain these (putative) facts? Researchers who work on the moral / conventional distinction maintain that their results can be explained by the hypothesis that moral rules and conventional rules belong to two quite different conceptual “domains.” By way of clarifying this hypothesis, these researchers highlight several important characteristics of the domains, maintaining that they are distinct and independent from each other, that they underlie subjects’ capacity to differentiate between different types of rules, and that they are present cross-culturally and in place quite early in development.

According to Nucci, for example, “[t]hese two forms of social regulation, morality and convention, are both part of the social order. Conceptually, however, they are not reducible to one another and are understood within distinct conceptual frameworks or domains (Nucci 2001, page 7. emphasis added). Turiel similarly claims that “social convention and morality a) constitute two distinct conceptual domains, which b) develop independently of each other” (Turiel 1979, p. 77). While they are sometimes hard to interpret, advocates of the domain hypothesis also suggest that the differences between the conceptual domains has an important role to play in explaining the criterion judgments elicited from subjects on the moral / conventional task. The nature of that role is often left vague because advocates emphasize subjects’ ability to differentiate different kinds of social rules, rather than spelling out the alleged role of the domains in explaining the ability. For example, Smetana remarks: “Children have been asked to make judgments along a set of dimensions that are hypothesized to differentiate moral and conventional rules. … In general, this research has indicated that children across a wide age range distinguish between moral and social-conventional rules and transgressions in
their reasoning and judgments” (Smetana 1993, pp. 114-115). Nucci more directly connects this ability to the domains, and to the specific criterion judgments elicited in the m/c task experiments: “[w]hat we have learned through research over the past twenty-five years is that people in general … reason very differently about matters of morality, convention and personal choice. More specifically, these conceptual differences become apparent when people are asked to evaluate different actions in terms of criteria [like those set out above] (Nucci 2001, p. 6). Nucci also makes the following remarks regarding the explanatory link between the domains and performance on the m/c task experiments:

In order to gain clear-cut answers to whether or not people make distinctions between morality and convention, researchers have asked people to make judgments that would constitute prototypical examples of moral or conventional issues [sic].... Consistent with the assumptions of domain theory, children and adults distinguish between morality and convention on the basis of these criteria” (2001, p. 10, emphasis added).

In elucidating the (putative) relationship between subjects’ performances on the m/c task and the hypothesized conceptual domains, comments such as these suggest a cognitive architecture like the one we are about to propose. Finally, advocates of the moral / conventional domain theory hold that these domains are cross-cultural, and in place early in psychological development. Nucci maintains that “in all cases, children and adolescents have been found to treat moral issues entailing harm and injustice in much the same way,” (2001, p. 12) and that “the domain of morality is structured around issues that are universal and nonarbitrary” (p. 19). Yau and Smetana hold that “[r]esearch in diverse cultures has shown that children across a wide age range differentiate morality from social convention” (2003, p. 654).

While the moral / conventional domain theorists do not go on to offer explicit cognitive models like those proposed by S&S, the details of their domain hypothesis suggest what such a model might look like. If the fact that a rule belongs to a particular domain is to explain the pattern of responses that subjects offer when presented with questions about the rule and transgressions of the rule, then a domain is best thought of as a functionally distinct component of the mind that stores rules (or representations of rules). In addition to its proprietary set of rules, each distinct domain would also contain a proprietary body of information. The information stored in each domain would lead subjects to respond as they do to questions about the rules stored that domain, and also to questions about transgressions of those rules. The information stored in the moral domain, for example, would indicate that rules stored therein are authority independent and general in scope; it would also indicate that those rules can be justified by appeal to harm, justice or rights and that transgressions of those rules are typically serious. Furthermore, in order to explain facts such as those described in (C2a), which claims that the signature moral response pattern is only evoked by rules that deal with harm, justice and rights, the domain hypothesis must also insist that the component of the mind that we’re calling the moral domain is restricted in such a way that it contains only rules of
that sort. Figure 2 is our attempt to capture the essential features of the domain hypothesis. We will call it the M/C model.

Figure 2
The M/C Model of the Psychological Mechanisms Underlying Performance on the moral/conventional task
The M/C model depicted in Figure 2 raises two important questions. First, where does the information in the domains come from? Second, what explains the fact that only rules dealing with harm, justice and rights end up being stored in the moral domain, while only rules not dealing with harm, justice and rights are stored in the conventional domain? Several answers to these questions have been proposed. First, though they are often hard to interpret, many researchers in the Turiel tradition suggest that the information about moral and conventional rules in the two domains is “constructed,” by which they seem to mean that it is not conveyed to them by other people. Rather that information is acquired via individual learning as the child interacts with the social environment. Researchers in this tradition also apparently believe that particular features of these interactions with the social environment enable the child to figure out which rules belong in which domain.8 Others, most notably Susan Dwyer (1999, forthcoming), impressed by the claim that the information contained in the domains is both pan-cultural and available early in development, argue that the information is innately specified. Dwyer may also believe that some of the rules in the moral domain are innately specified as well. In support of this view, she offers a version of the “poverty of the stimulus” argument commonly found in discussions of linguistic knowledge: It is hard to see how the information that the child ends up with could possibly be inferred from the limited information available in the child’s physical and social environment.9 Finally, Shaun Nichols (2002, 2004) has recently offered a rather different account in which both social transmission and innate predispositions play a role. On Nichols’ hypothesis, the content of both moral and conventional rules is acquired via social transmission. However people are innately disposed to have affective responses to actions with certain sorts of consequences, and rules proscribing those actions evoke the signature moral response.

Obviously, each of these alternatives needs to be spelled out in greater detail. That’s not a project we propose to undertake here, however. Nor need we take a stand on which alternative is more plausible. For it is our view that the architecture proposed in the M/C model is seriously mistaken. To put the point bluntly, we don’t believe that the psychological domains posited by the M/C model exist. If we are right, then questions about where the information in the domains comes from and how particular rules get assigned to one domain or the other are otiose.

8 For instance, Turiel (1983, p. 9) says that “thought is organized and … it is constructed out of the child’s interactions with the environment.” See also Turiel (1979, p. 108): “the child’s conceptual knowledge is formed out of his actions upon the environment: To form concepts about objects and events the child must act upon them. Thus conceptual development is a constructive process stemming from individual-environment interactions.” In response to the second question, what explains the fact that only rules dealing with harm, justice and rights come to be stored in the moral domain, while only rules not dealing with harm, justice and rights come to be stored in the conventional domain, domain theorists appeal to the (putatively) distinctive and intrinsic features of actions that violate moral rules. Rules dealing with harm, justice and rights end up in the moral domain because transgressions of those rules, in contrast to transgressions of conventional rules, are marked by distinctive and intrinsic features, namely “consequences such as harm inflicted upon others, violation of rights, effect on general welfare” (Turiel 1979, page 80).

9 For another discussion of the innateness of the moral / conventional distinction, see Wilson (1993), 141 ff.
Before setting out our case against the M/C model, it will be useful to underscore the differences between that model and the S&S model, and to draw out some of the ways in which the models lead to quite different predictions. Since the M/C model was designed to explain (C-1), (C-2) and (C-3) – the major conclusions that researchers in the Turiel tradition have drawn from moral / conventional task experiments – it is no surprise that the M/C model is comfortably compatible with those conclusions. But if the S&S model is correct, we should expect each of those conclusions to be false.

To see why, let’s focus first on (C-1), the clustering of criterion judgments. The claim here is that the two signature response patterns in moral / conventional task experiments are nomological clusters, and thus that the members of each cluster will typically occur together. On the M/C model, this is just what we should expect, since responses to moral / conventional task questions are guided by the information in the domain where the rule being investigated is stored. On the S&S theory, on the other hand, no such nomological clustering is to be expected. According to the S&S theory, any rule in the normative rule data base will generate reliable and robust intrinsic motivation to comply and to punish violators. Since these motivations are intrinsic, they do not depend on authority, or on the existence of social rules or on fear of social sanctions. So, for any rule stored in a subject’s normative rule data base, we would expect the subject to judge the rule to be authority independent when given the moral / conventional task, since the subject is feels motivated to comply and to punish violations whether or not the rule is sanctioned by an authority. However, the S&S theory gives no reason to think authority independence will regularly be accompanied by any other specific criterion judgment. On the contrary, rules stored in the normative rule data base can vary in how general they are, how serious transgressions are, and what their justification is. Thus, we should not expect that rules judged to be authority independent will also be judged to be applicable to everyone, that their transgressions will be judged to be serious, or that they will be justified by appeal to harm, justice or rights.

The S&S theory also maintains that lots of different sorts of behavior regulating rules will be stored outside the normative rule data base – in the black boxes in Figure 1. Though some rules stored there might evoke an authority independent response, many will not. Moreover, rules stored outside the normative rule data base may evoke any pattern of answers on the seriousness and generality questions. So if the S&S model is on the right track, there should be no nomological clustering of the signature response patterns. Indeed, the S&S theory leads us to expect that responses in the moral / conventional task could occur in just about any combination.

(C-2) deals with the alleged correlation between response patterns and transgression types. More specifically, it maintains that transgressions involving harm, justice or rights will evoke the signature moral pattern, while transgressions not involving harm, justice or rights will evoke the signature conventional pattern. And here again, of course, this is just what the M/C model would predict, since on that model only rules involving harm, justice or rights can be stored in the moral domain, and only rules not involving harm, justice or rights can be stored in the conventional domain. On the S&S model, by contrast, neither rules involving harm, justice and rights nor rules not involving
harm, justice or rights constitute a distinctive psychological category. Some rules from each group may find their way into the normative rule data base, and others may be stored in other components of the mind. So, for example, on the S&S account, it is entirely possible that a rule prohibiting harm of a certain sort would be stored outside the normative rule data base and thus that a transgression of that rule would evoke an authority dependent response. It is also possible that a rule prohibiting behavior that does not involve harm, justice or rights would be included in the normative rule data base, and thus that a transgression of that rule would evoke an authority independent response.

Finally, according to (C-3), the regularities described in (C-1) and (C-2) are both pan-cultural and early emerging. The M/C model, as we have developed it, predicts that the patterns will be pan-cultural, though it does not explain why they emerge early in development. The S&S theory need not worry about the patterns being pan-cultural or early emerging, since, as we’ve just seen, the S&S theory predicts that the patterns do not exist at all!

Clearly, there is no shortage of empirically testable disagreements between the two models. Let’s now ask which one fares better in accommodating the data.

5. The Models and the Evidence

In section 3 we gave an overview of some of the findings that have led many researchers in the Turiel tradition to advocate conclusions (C-1) – (C-3). Not everyone has been persuaded by these conclusions, however. Most of the dissenters have been impressed with the diversity in the sorts of behaviors that different cultures “moralize” by treating them as wrong in an authority independent way. These researchers have focused on rules and transgressions that do not involve harm, justice, or rights. (C-2b) predicts that such transgressions should evoke the signature conventional response pattern. But, the dissenters maintain, there are many societies in which such transgressions evoke one or more of the signature moral responses. If this is correct then not only is (C-2b) false, but so is (C-3) – the claim that the regularities described in (C-1) and (C-2) are pan-cultural.

For example, in an pioneering and influential study Haidt et al. (1993) employed much of standard moral / conventional task methodology, and showed that low SES groups in both Brazil and the USA judged activities such as privately washing the toilet bowl with the national flag and privately masturbating with a dead chicken to be generally and seriously wrong, and that this judgment did not depend on any authority figure or explicit rule prohibiting these activities. In addition to the standard probe questions, Haidt et al. added another question that allowed subjects to explicitly specify which transgressions they took to be harmless. Even when the low SES groups acknowledged that no one was harmed by a particular sort of behavior, those groups still judged that many of the harmless transgressions had most of the features of the signature

10 To the best of our knowledge, advocates of moral / conventional domain theory have never offered an explanation of the (putative) fact that the patterns emerge early in development.
moral response pattern. Other researchers employing the moral / conventional task methodology have reported similar results. In a study of children in traditional Arab villages in Israel, Nisan (1987) found that all the transgressions tested evoked most of the signature moral response pattern, including such transgressions as mixed-sex bathing and addressing a teacher by his first name – behaviors that clearly do not involve harm, justice, or rights. In another study, Nucci and Turiel reported that orthodox Jewish children in the USA judged a number of religious rules to be authority independent even though the rules did not deal with harm, justice, or rights (Nucci & Turiel 1993; see also Nucci 2001, chapter 2 for discussion).

Perhaps most interestingly, Nichols (2002, 2004) showed that for a particular subset of etiquette rules, namely those that prohibit disgusting actions, American children judged transgressions to be serious, authority independent, and general in scope. American college students judged transgression of those same etiquette rules to be serious and authority independent, though they did not regard the rules as general in scope. Like the other studies just described, Nichols’ work clearly raises problems for claim (C-2b). However, Nichols’ results are unique in that they also pose a particularly clean challenge to (C-1), the claim about the clustering of criterion judgments. In Nichols’ study, not only do transgressions that do not involve harm, justice, or rights evoke most of the elements of the signature moral response pattern contrary to what (C-2b) predicts, but the putative nomological clusters posited in (C-1) come apart in two different ways. Indeed, Nichols finds 3 different sets of responses to rules that do not involve harm, justice, or rights, and finds that adults and children respond differently to the same rules.

Taken together, we think the findings just cited pose a significant challenge to (C-1) – (C-3), and thus to the M/C model which predicts those conclusions. Since the S&S theory does not predict that transgressions not involving harm, justice or rights will exhibit the signature conventional response pattern, and does not expect criterion judgments to exhibit any systematic pattern or nomological clustering, all of the findings we’ve just cited are comfortably compatible with the S&S theory. Moreover, we suspect that the results described in the previous two paragraphs may be only the tip of the iceberg. For a variety of reasons, researchers using the moral / conventional task have only looked at a relatively narrow range of transgressions which do not involve harm, rights or justice. However, the literature in cultural psychology and anthropology, as well as reports in the popular press lead us to expect that if researchers using the moral / conventional task were to study a more extensive range of transgressions in a wider range of cultural groups, they would find (C-1) – (C-3) massively disconfirmed. For example, we would expect that a vast majority of Americans, along with people in many other cultures, would judge that consensual sibling incest is wrong, and that the wrongness of incest is authority independent.12 We would expect much the same judgment about

11 The third pattern that Nichols found was the only one predicted by (C-2b): etiquette rules prohibiting actions that are not disgusting inducing evoke the signature conventional pattern.

12 Haidt (2001) reports a study in which university age subjects could not justify their strong moral condemnation of a case of consensual sibling incest in which the couple used two forms of birth control.
homosexual sex from the 55% of the American public who tell opinion researchers that homosexual behavior is a sin.\textsuperscript{13} We are also prepared to bet that in traditional societies where taboo violations and failure to respond appropriately to “polluting” acts like being touched by a low caste person are taken very seriously, these violations would not lead to the full set of signature conventional responses that would be predicted by the M/C model.\textsuperscript{14}

It is noteworthy that none of the studies we have described as posing a challenge to (C-1) – (C-3) use transgressions involving harm, justice, or rights. Nor have we been able to find any other study in the literature that contradicts (C-2a) by demonstrating that transgressions involving harm, justice, or rights do not evoke the signature moral pattern. One possible explanation for the absence of such studies in the literature is that (C-2a) is both true and pan-cultural. Perhaps transgressions involving harm, justice, or rights do reliably and cross-culturally evoke the signature moral response pattern. However, we think there are at least three reasons to be skeptical of this explanation. First, though there are many studies employing the moral / conventional task paradigm, the range of transgressions involving harm that have been included in these studies is remarkably narrow. Early work using the paradigm was done by developmental psychologists and was focused on young children. Thus the examples of harmful transgressions studied were all behaviors that would be familiar to youngsters, such as pulling hair and pushing someone off a swing. In the intervening years, the moral / conventional task has been used with a number of different subject populations, and the set of transgressions that do not involve harm, justice or rights broadened somewhat as well. Though we know of no study that asked subjects to consider incest, homosexuality or taboo violations, some of the transgressions described in more recent work were behaviors that might not be familiar to young children. Oddly, however, all of the harmful transgressions studied have been of the “schoolyard” variety, even when the experimental subjects were incarcerated psychopathic murderers (Blair 1995)! As a result, little is known about how people respond to a broader range of harmful transgressions in the moral / conventional task. Second, philosophical views like Bernard Williams’ “relativism of distance” and the sophisticated version of moral relativism defended by Gilbert Harman encourage the speculation that there may be many moral rules – including those prohibiting slavery, corporal punishment, and treating women as chattel – that people do not generalize to other cultures or other historical periods (Williams 1985; Harman 2000). Though these philosophers offer only anecdotal evidence, we think these speculations have considerable intuitive plausibility. Third, our informal sampling of public discussion about recent news stories dealing with issues such as the treatment of detainees at the U.S. military base in Guantanamo Bay suggests that a significant number of people do


\textsuperscript{14} See Shweder et al. (1987) and Shweder et al. (1997) for some suggestive discussion of norms governing polluting acts, and Fessler & Navarrete (2003) for very useful material on taboos.
not consider rules prohibiting harmful treatment in such cases to hold independently of authority.

In order to explore the possibility that many harmful transgressions that are not of the schoolyard variety would not evoke the signature moral response pattern, we recently designed a web-based study, in collaboration with Kevin Haley, Serena Eng and Daniel Fessler, in which participants were asked about a number of such transgressions (Kelly et al., forthcoming). For example, to explore whether rules prohibiting use of corporal punishment are judged to be authority independent, participants were presented with the pair of questions in Box 1. The results were quite dramatic: 8% of participants said it was OK to spank the boy in response to question (A) and 48% said it was OK to spank the boy in response to question (B). Similar results were found when the questions, appropriately modified, were asked in the opposite order. So for a very substantial number of respondents, it appears that the rule against spanking is not authority independent. Five other scenarios were used to explore whether rules prohibiting serious harms would be judged to be authority independent, and in each case the results indicated that, for a significant number of subjects, they were not.

15 Pooling the two orders, 5% judged that spanking was OK in response to question (A) and 44% judged that it was OK in response to question (B). p = 0.000.

16 For the full text of all questions used in this study, along with all of the data, contact the Daniel Kelly or Stephen Stich.
Box 1
A Pair of Questions Designed to Determine Whether Participants Judged a Rule Against Corporal Punishment to Be Authority Independent

(A) It is against the law for teachers to spank students. Ms. Williams is a third grade teacher, and she knows about the law prohibiting spanking. She has also received clear instructions from her Principal not to spank students. But when a boy in her class is very disruptive and repeatedly hits other children, she spans him.

Is it OK for Ms. Williams to spank the boy?

YES  NO

On a scale from 0 to 9, how would you rate Ms. Williams' behavior?

Not at all bad  0  1  2  3  4  5  6  7  8  9  Very bad

(B) Now suppose that it was not against the law for teachers to spank students, and that Ms. Williams' Principal had told her that she could spank students who misbehave if she wanted to.

Is it OK for Ms. Williams to spank the boy?

YES  NO

On a scale from 0 to 9, how would you rate Ms. Williams' behavior?

Not at all bad  0  1  2  3  4  5  6  7  8  9  Very bad

The pair of questions in Box 2 was designed to determine whether participants judged rules prohibiting harmful behavior to be temporally universal. Are actions that are judged to be wrong now also judged to be wrong in the past? Once again the results were quite dramatic, clearly confirming Williams’ claims about the “relativism of distance.” In response to question (A) 52% of participants said that it was OK to whip a drunken sailor 300 years ago, but only 6% said it was OK to do it today!\(^\text{17}\) A second pair of questions asked subjects to judge the wrongness of slavery in the American South and in ancient Greece and Rome. In this case, too, significantly more subjects judged slavery to be wrong long ago and far away.

\(^{17}\) Asking the questions in the opposite order had no significant effect. When the results from the two orders were pooled, 51% said whipping was OK in response to (A) and 10% said it was OK in response to (B). \(p = 0.000\).
We believe that the Kelly et al. experiment poses a serious challenge to (C-2a), which claims that harm norms evoke the signature moral pattern. Rather, it seems, when we go beyond the narrow range of schoolyard transgressions that have been used in previous studies, many subjects think that rules prohibiting harmful actions are neither authority independent nor general in scope. In directly challenging the conclusion (C2a) these findings significantly add to the case against the M/C model, which was designed to predict that conclusion and explain why it was true. As we noted earlier, the S&S model, in contrast with the M/C model, accords harm norms no special status. According to the S&S theory, some harm norms may be stored in the normative rule data base, and those

---

**Box 2**

A Pair of Questions Designed to Determine Whether Participants Judged a Rule Against Corporal Punishment to Be Temporally General

A) Three hundred years ago, whipping was a common practice in most navies and on cargo ships. There were no laws against it, and almost everyone thought that whipping was an appropriate way to discipline sailors who disobeyed orders or were drunk on duty.

Mr. Williams was an officer on a cargo ship 300 years ago. One night, while at sea, he found a sailor drunk at a time when the sailor should have been on watch. After the sailor sobered up, Williams punished the sailor by giving him 5 lashes with a whip.

Is it OK for Mr. Williams to whip the sailor?

| YES | NO |

On a scale from 0 to 9, how would you rate Mr. William' behavior?

Not at all bad | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Very bad | 9

B) Mr. Adams is an officer on a large modern American cargo ship in 2004. One night, while at sea, he finds a sailor drunk at a time when the sailor should have been monitoring the radar screen. After the sailor sobered up, Adams punishes the sailor by giving him 5 lashes with a whip.

Is it OK for Mr. Adams to whip the sailor?

| YES | NO |

On a scale from 0 to 9, how would you rate Mr. Adams' behavior?

Not at all bad | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Very bad | 9

---

We believe that the Kelly et al. experiment poses a serious challenge to (C-2a), which claims that harm norms evoke the signature moral pattern. Rather, it seems, when we go beyond the narrow range of schoolyard transgressions that have been used in previous studies, many subjects think that rules prohibiting harmful actions are neither authority independent nor general in scope. In directly challenging the conclusion (C2a) these findings significantly add to the case against the M/C model, which was designed to predict that conclusion and explain why it was true. As we noted earlier, the S&S model, in contrast with the M/C model, accords harm norms no special status. According to the S&S theory, some harm norms may be stored in the normative rule data base, and those
that are will be judged to be authority independent, though they may be of limited
generality. Others may be stored in other components of the mind, and those may be
judged to be both authority dependent and of limited generality. So the Kelly et al.
results are fully compatible with the S&S theory.

6. Conclusion

Our goal, in this paper, has been to assess the merits of two competing accounts
of the cognitive architecture underlying morality: the S&S model, which was designed to
account for a range of findings in a variety of disciplines, and the M/C model, which was
designed to explain the main conclusions drawn from a large body of work using the
moral / conventional task. We’ve tried to shape the discussion in a way that emphasizes
the differences between these two models and highlights the fact that they are
incompatible with each other: they make divergent predictions about a wide range of
moral judgments, including the sorts of judgments that are central to the m/c task. The
view we’ve been arguing for is that the S&S model is clearly superior, especially in
light the growing body of evidence indicating that the conclusions (C-1), (C-2) and (C-3),
which the M/C model was designed to explain, are themselves very problematic. A
leitmotif in our critique of the conclusions drawn from moral / conventional task studies
is that these studies have focused on a very narrow range of rules and transgressions. As
researchers have begun to explore people’s judgments about a broader and more varied
class of rules and transgressions, the shortcomings of the conclusions drawn from earlier
work using the moral / conventional task have become increasingly apparent.

While the focus of this paper has been largely restricted to two specific accounts
of cognitive architecture, there is reason to think that if correct, our grim assessment of
the conclusions drawn from studies using the moral / conventional task has implications
of much wider relevance. In recent years, a number of psychologists and philosophers
have assumed that the moral / conventional task tells us something important about moral
psychology, and they have used this assumption in arguing for a variety of important
claims. For example, the philosopher Shaun Nichols (2004) has claimed that the capacity
draw the moral / convention distinction “reflects the ability to appreciate the distinctive
status of morality” (p. 4), that it “plumbs a fairly deep feature of moral judgment” (p. 6),
and that it can be used “as a measure of moral cognition” (p. 196). And the psychologist
James Blair (1995, 1996, 1997; Blair et al. 2001) has used the test to draw conclusions
about the moral capacities of psychopaths and individuals with autism. We’ve argued
that the evidence reviewed above shows the M/C model of cognitive architecture is false.
That evidence also suggests that the moral / conventional task itself is not a good assay
for the existence of a psychologically important distinction. If that’s right, then the
reasoning behind claims like Nichols’ and Blair’s merits very careful scrutiny.
References


