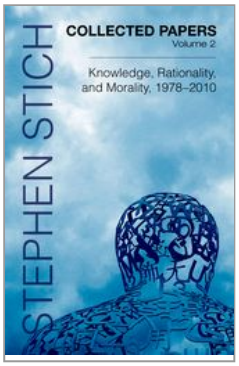

Collected Papers, Volume 2: Knowledge, Rationality, and Morality, 1978-2010
Stephen Stich

Print publication date: 2012
Print ISBN-13: 9780199733477
Published to Oxford Scholarship Online: September 2012
DOI: 10.1093/acprof:oso/9780199733477.001.0001

A Framework for the Psychology of Norms

Chandra Sekhar Sripada
Stephen Stich

DOI:10.1093/acprof:oso/9780199733477.003.0012

Abstract and Keywords

This chapter offers an account of the psychological mechanisms and processes underlying norms that integrate what is known and can serve as a framework for future research. The chapter is organized as follows. Section 1 offers a preliminary account of what norms are. Sections 2 and 3 assemble an array of facts about norms and the psychology that makes them possible, drawn from a variety of disciplines. Though the distinction is not a sharp one, Section 2 focuses on social level facts, while Section 3 focuses on how norms affect individuals. Section 4 provides a tentative hypothesis about the innate psychological architecture subserving the acquisition and implementation of norms, and explains why an architecture like the one proposed can explain many of the facts assembled in Sections 2 and 3. Section 5 focuses on open questions —important issues about the cognitive science of norms that the account in Section 4 does not address.

Keywords: norms, psychological mechanisms, psychological processes, psychology, cognitive science

> No concept is invoked more often by social scientists in the explanations of human behavior than norm.
>
> —*Encyclopedia of the Social Sciences*

Humans are unique in the animal world in the extent to which their day-to-day behavior is governed by a complex set of rules and principles commonly called *norms*. Norms delimit the bounds of proper behavior in a host of domains,

providing an invisible web of normative structure embracing virtually all aspects of social life. People also find many norms to be deeply meaningful. Norms give rise to powerful subjective feelings that, in the view of many, are an important part of what it is to be a human agent. Despite the vital role of norms in human lives and human behavior, and the central role they play in explanations in the social sciences, very little systematic attention has been devoted to norms in cognitive science. Much existing research is partial and piecemeal, making it difficult to know how individual findings cohere into a comprehensive picture. Our goal in this essay is to offer an account of the psychological mechanisms and processes underlying norms that integrates what is known and can serve as a framework for future research.

In section 1, we'll offer a preliminary account of what norms are. In sections 2 and 3, we'll assemble an array of facts about norms and the psychology that makes them possible, drawn from a variety of disciplines. Though the distinction is not a sharp one, in section 2, we'll focus on social level facts, while in section 3, our focus will be on how norms affect individuals. In section 4, we'll offer a tentative hypothesis about the innate psychological architecture subserving the acquisition and implementation of norms, and explain why we believe an architecture like the one we propose can explain many of the facts assembled in sections 2 and 3. Section 5, the last and longest section, focuses on open **(p. 286)** questions—important issues about the cognitive science of norms that our account in section 4 does not address. In some cases, we've left these issues open because little is known about them; in other cases, more is known but crucial questions are still very much in dispute. Though we are acutely aware that our account of the psychology of norms leaves many important questions unanswered, we hope that the framework we provide will contribute to future research by clarifying some of those questions and offering an overview of how they are related.[1]

1. A Preliminary Characterization of Norms

We'll begin with an informal and provisional account of what we mean when we talk of *norms*. As we use the term, a norm is a rule or principle that specifies actions that are required, permissible, or forbidden independently of any legal or social institution. Of course, some norms are *also* recognized and enforced by social institutions and laws, but the crucial point is that they needn't be. To emphasize this fact, we'll sometimes say that norms have *independent normativity*. Closely linked to the independent normativity of norms is the fact that people are motivated to comply with norms in a way that differs from their motivation to comply with other kinds of social rules. Very roughly, people are motivated to comply with norms as *ultimate ends*, rather than as a means to other ends; we'll refer to this type of motivation as *intrinsic motivation*, and we'll have much more to say about it in section 3. People can *also* be motivated to comply with a norm for instrumental reasons, though intrinsic compliance motivation adds a substantial additional motivational force. Violations of norms,

when they become known, typically engender *punitive attitudes*, like anger, condemnation, and blame, directed at the norm violator, and these attitudes sometimes lead to punitive behavior.

We believe that norms, as we've characterized them, are an important and theoretically useful subcategory of social rules, and that our characterization is broadly in line with other accounts, both historical and more recent (see Durkheim 1903/1953; McAdams 1997; Parsons 1952; Petit 1991). However, it is worth emphasizing that our account of norms is *not* intended as a *conceptual analysis* or an account of what the term "norm" means to ordinary speakers. Nor do we offer our characterization of norms as a formal definition. At best, it gives a rough-and-ready way to pick out what we believe is a theoretically interesting *natural kind* in the social sciences. If the framework for a psychological theory of norms set out in section 4 is on the right track, then a better account of the crucial features of norms can be expected to emerge as that theory is elaborated. One of **(p.287)** the components of our framework is a "norm database," and it is the theory's job to tell us what can and cannot end up in that database.

Though there are a substantial number of empirically well-supported generalizations about norms, those generalizations and the evidence for them are scattered in the literatures of a number of different disciplines. In the next two sections, we'll assemble some of these generalizations and say a bit about the evidence for each. We'll begin with social-level features of norms, and then turn to individual-level facts about the ways norms are acquired and how they influence behavior.

2. Some Social-Level Facts About Norms

Norms are a *cultural universal*. The ethnographic database strongly suggests that norms and sanctions for norm violations are universally present in all human societies (Brown 1991; Roberts 1979; Sober and Wilson 1998). Moreover, there is reason to think that the universal presence of norms is very *ancient*. There is no evidence that norms originated in some society and spread by contact to other societies in the relatively recent past. Rather, norms are reliably present and are highly elaborated in all human groups, including hunter-gatherer groups and groups that are culturally isolated. This is just what we would expect on the hypothesis that norms are very ancient. All of this, we think, suggests that there are innate psychological mechanisms specialized for the acquisition and implementation of norms, since the existence of these mechanisms would help explain the universal presence of norms in all human groups.

In addition to being present in all cultures, norms tend to be ubiquitous in the lives of people in those cultures. They govern a vast array of activities, ranging from worship to appropriate dress to disposing of the dead. And while some

norms deal with matters that seem to be of little importance, others regulate matters like status, mate choice, food, and sex that have a direct impact on people's welfare and their reproductive success.

Although norms are present in all human groups, one of the most striking facts about them is that the *contents* of the norms that prevail in different groups are quite variable. Moreover, these differences follow a characteristic pattern in which there is substantial homogeneity in the norms that prevail *within* groups and both commonalities and differences in the norms that prevail *across* groups. We believe that the distributional pattern of norms is an important source of evidence about the psychological mechanisms that underlie them. For this reason, we'll spend some time discussing the issue in more detail.

In assessing the distribution of norms across human groups, one question that immediately arises is: Are there any norms that are universally present in all human groups? The question must be handled with some care, since many candidate norm universals are problematic because they verge on being *analytic* —true in virtue of meaning alone. For example "Murder is wrong" or "Theft is wrong" don't count as legitimate universals since, roughly speaking, "murder" simply means killing someone else in a way that is not **(p.288)** permissible, and "theft" simply means taking something from another in a way that is not permissible. For this reason, it is important, wherever possible, to frame the contents of norms in a nonnormative vocabulary. While analytic principles like "Murder is wrong" and "Theft is wrong" may be universals, the *specific* rules that regulate the *circumstances* under which killing or taking an item in the possession of another person is permitted are not so nearly uniform across groups.

With this caveat in mind, we return to the question of the distributional pattern of norms across human groups. One important fact is that there *is* a pattern to be discerned; norms are not indefinitely variable or randomly distributed across human groups. Rather, there are certain kinds of norms one sees again and again in almost all human societies, though in order to discern these commonalities, one has to stay at a fairly high level of generality. For example, most societies have rules that prohibit killing, physical assault, and incest (or sexual activity with one's kin). In addition, most societies have rules promoting sharing, reciprocating, and helping, at least under some circumstances (Cashdan 1989). Most societies have rules regulating sexual behavior among various members of society, and especially among adolescents (though the content of these rules varies considerably) (Bourguignon and Greenbaum 1973). And most societies have at least some rules that promote egalitarianism and social equality. For example, in nearly all hunter-gatherer groups, attempts by individuals to garner a disproportionate share of resources, women, or power are disapproved of sharply (Boehrn 1999). Examples like these could be

multiplied easily in domains such as social justice, kinship, marriage, and many others.

While there is no doubt that there are certain high-level commonalities in the norms that prevail across groups, as one looks at norms in more detail, it is clear that there is tremendous variability in the *specific rules* one finds in different groups. Consider, for example, norms dealing with *harms*. While some kind of harm norm or other is found in virtually all human groups, the specific harm norms that prevail across groups are quite variable. In some simple societies, almost all harm-causing behaviors are strongly prohibited. Among the Semai, an aboriginal people of the Malaysian rain forest, for example, hitting and fighting, as well as more mundane behaviors such as insulting or slandering, are all impermissible, and Semai groups have among the lowest levels of violence of any human societies (Robarchek and Robarchek 1992). But other groups permit a much wider spectrum of harm-causing behaviors. In groups such as the Yanomamö of South America, the use of violence to settle conflicts is permitted (and indeed extremely common), and displays of fighting bravado are prized rather than condemned (Chagnon 1992). Among the Yanomamö, mortality due to intra- and intertribe conflict is extremely high, and some ethnographers have suggested that the level of mortality due to violence found among the Yanomamö is not at all uncommon in simple societies (Keeley 1996). In addition to variability in the kinds of harm and level of harm that are permitted, harm norms also differ with respect to the class of individuals a person is permitted to harm. Many groups draw a sharp distinction between harms committed against individuals within one's own community and individuals outside the group (though many groups do **(p.289)** *not* draw such a sharp distinction; LeVine and Campbell 1972). Moreover, some societies permit some kinds of violence directed against women, children, animals, and also certain marginalized subgroups or castes (Edgerton 1992). The variability in harm norms is also evidenced by the manner in which they change over time. The philosopher Shaun Nichols (2004, ch. 7) provides a fascinating description of the gradual change in harm norms in Western societies over the last 400 years.

Incest prohibitions are another case in which high-level commonalities are found in conjunction with variability at the level of specific rules. It appears that almost all societies have norms prohibiting sexual intercourse between members of the nuclear family (we'll call these nearly universal rules *core incest prohibitions*). But incest prohibitions almost always extend beyond this core. In particular, incest prohibitions almost always extend to *other kinds* of sexual activity, and they almost always extend beyond just the *nuclear family*; they prohibit sexual activity with at least some members of one's nonnuclear kin. But the details of how incest prohibitions extend beyond core incest prohibitions are, as numerous studies have revealed, tremendously variable (Murdock 1949). For example, at one extreme are *exogamous groups*, in which marriage with *anyone* within one's own tribal unit is considered incestuous, though the offense is

seldom seen as being of the same level of severity as intercourse within one's nuclear family.

Another feature of the distributional pattern of norms is that while most groups have some rule or other that falls under certain high-level themes, generalizations about commonalities in the norms found across groups typically have *exceptions*. For example, the incest prohibition is sometimes cited as the best example of a norm that is a universal feature of all human groups. And while it is true that core incest prohibitions can be found in virtually all groups, even this generalization may not be exceptionless. There is good evidence that brother-sister marriage (including sexual relations) occurred with some frequency in Egypt during the Roman period, and was practiced openly and unabashedly. In addition, brother-sister marriage is known to have occurred in a number of royal lineages, including those of Egypt, Hawaii, and the Inca empire (Durham 1991).

To sum up, we've identified three key features of the distributional pattern of norms. First, norms tend to cluster under certain general *themes*. Second, the specific rules that fall under these general themes are quite *variable*, though clearly thematically connected. And third, there are typically at least some *exceptions* that diverge from the general trend.

3. Some Individual-Level Facts About Norms

We turn now to some facts about how norms emerge within individuals, and how individuals are affected by the norms they acquire. There is excellent evidence indicating that norms exhibit a *reliable pattern of ontogenesis*. Regardless of their biological heritage, almost everyone (excepting those with serious psychological deficits) acquires the norms that prevail in the local cultural group in a highly reliable way. In no human group is it **(p.290)** the case that some individuals reliably acquire the prevailing norms while many others don't. It also appears that all individuals acquire at least some norms of their group relatively *early* in life. All normal children appear to have knowledge of rules of a distinctly normative type between three and five years of age, and can distinguish these normative rules from other social rules (Nucci 2001; Turiel 1983). In addition, some competences associated with norms, such as the ability to reason about normative rules and rule violations, appears very early. Denise Cummins has shown that children as young as three to four perform substantially better on deontic rule reasoning tasks than they do on similar indicative reasoning tasks (Cummins 1996).

Further evidence about the ontogenesis of norms comes from a major crosscultural study in which Henrich and his colleagues investigated norms of cooperation and fairness in 15 small-scale societies using standard experimental game paradigms. (We'll discuss these games more fully later.) While this study found considerable diversity in the norms of cooperation and fairness prevailing

in these societies, it also found that much of the crosscultural variation in norms among adults was already present by the time subjects reached the age of nine, and it persists thereafter (Henrich et al. 2001). In another crosscultural experimental study, Shweder and his colleagues examined moral norms in children and adults in Hyde Park, Illinois, and Bhubaneswar, India (Shweder, Mahapatra, and Miller 1987). As in Henrich and colleagues' study, there were lots of differences in the norms that prevailed in the two communities, and most of the differences were already established by the time subjects reached the age of seven.

Perhaps the most striking (and most overlooked) feature of norms is that they have powerful *motivation effects* on the people who hold them. Philosophers have long emphasized that from a subjective perspective, moral norms present themselves with a unique kind of subjective authority that differs from standard instrumental motivation. We believe that this philosophical intuition reflects a deep empirical truth about the psychology of norms, and we refer to the type of motivation associated with norms as *intrinsic motivation*. Our claim is that people are disposed to comply with norms even when there is little prospect for instrumental gain, future reciprocation, or enhanced reputation, and when the chance of being detected for failing to comply with the norm is very small. The claim we are making must be treated with care, however. At any given time, a person may be subject to multiple sources of motivation. So in some cases in which people are intrinsically motivated to comply with a norm, they may *also* be instrumentally motivated to comply with the norm. In other cases in which people are intrinsically motivated to comply with norms, they may nonetheless fail to comply for instrumental reasons. So our claim is not that people always *follow* norms or that when they follow norms they do so *only* because of intrinsic motivation. Rather, our claim is that humans display an independent intrinsic source of motivation for norm compliance, and thus that people are motivated to comply with norms *over and above* (and *to a substantial degree* over and above) what would be predicted from instrumental reasons alone.

(p.291) There is an implication of our claims about intrinsic motivation that is worth emphasizing. Many norms, though by no means all, direct individuals to behave *unselfishly*. More precisely, many norms direct individuals to behave in ways that are contrary to what would in fact maximize satisfaction of their selfish preferences. Thus, in saying that people are intrinsically motivated to comply with norms, we are committed to the claim that people are motivated to comply in a way that frequently leads them to behave genuinely unselfishly. While philosophers have taken the claim that people are intrinsically motivated to comply with norms to be obvious and platitudinous, economic theorists and evolutionary-minded scientists have often argued that such behavior is very implausible from the perspective of selfish rationality (see Barash 1979, 135, 167; Downs 1957). We believe the arguments used by these theorists are deeply flawed. But a full rebuttal would take us far from the current topic, and here we

instead emphasize that the claim that people are intrinsically motivated to follow norms has substantial *direct* empirical justification.

Some of this evidence comes from anthropology and sociology. A central principle of these disciplines is that people *internalize* the norms of their group. According to the internalization hypothesis, individuals exhibit a characteristic style of motivation in which the individual intrinsically values compliance with moral rules even when there is no possibility of sanction from an external source (Durkheim 1912/1968; Scott 1971). Internalization is invoked to explain a seemingly obvious and ubiquitous fact: having been taught to comply with the moral rules of their group, people exhibit a *lifelong* pattern of *highly reliable* compliance with the rule. Furthermore, this pattern of compliance does not seem to depend on overt coercion, or even the threat of coercion, at each particular instance in which compliance is displayed. Consistent with the internalization hypothesis, the ethnographic record routinely reports that people view norms as being distinctive because of their absoluteness, their authority, and the manner in which people regard them as deeply meaningful (see Edel and Edel 2000). These features of norms suggest that norm compliance is based on something over and above instrumental motivation.

Closer to home, the economist Robert Frank (1988) has pointed out a number of cases of norm compliance in day-to-day life that are not plausibly viewed as the product of instrumental rationality. His examples include tipping at a highway restaurant one will never revisit, jumping in a river to save a drowning person, refraining from littering on a lonely beach, returning a lost wallet containing a substantial amount of cash, and many others.

Though descriptive data of this sort is compelling enough, a problem for those who wish to defend the claim that people intrinsically comply with norms is that it is easy for skeptics to concoct a selfish instrumental motive for what superficially appears to be intrinsic compliance behavior. For this reason, experimental data that can distinguish the competing hypotheses is crucial. The social psychologist C. Daniel Batson has, over the course of a number of years, extensively studied the motivational structure of helping behavior using a number of ingenious experimental paradigms. Batson finds that helping behavior is best accounted for on the hypothesis that people promote the welfare of others as an ultimate end (especially when their empathy is engaged) and not on alternative **(p.292)** hypotheses that treat helping as instrumental toward ulterior benefits such as future reciprocation, or gaining social approval (Batson 1991). There is now a large literature in sociology and social psychology that reaches a similar conclusion. Reviewing this literature, Pilliavin and Charng note:

> There appears to be a paradigm shift away from the earlier position that behavior that appears to be altruistic must, under closer scrutiny, be revealed as reflecting egoistic motives. Rather, theory and data now being advanced are more compatible with the view that true altruism—acting with the goal of benefiting another—does exist and is part of human nature. (1990, 27)

But perhaps the most compelling data indicating that people follow norms as ultimate ends comes from experimental economics, where people's motivations to comply with norms of fairness and reciprocity can be precisely detected and quantified. There is now abundant evidence that in experimental games, subjects cooperate at levels *far* higher than instrumental rationality alone would predict. For example, subjects routinely cooperate in *one-time-only, anonymous* prisoner's dilemma games (Marwell and Ames 1981). In such games, choosing to cooperate is the "fair" thing to do, while choosing to defect will earn the subject a higher payoff, regardless of what the other person chooses. Furthermore, these results are obtained even when subjects are *explicitly told* that they will play the game only once, and their identity will remain anonymous. The fact that subjects still routinely choose to cooperate suggests that that they are complying with norms of fairness and reciprocity as an ultimate end, rather than pursuing what would satisfy their selfish preferences. There are a large number of other kinds of games, such as public goods games, the ultimatum game, the centipede game, and others in which similar results have been obtained (see Thaler 1992, especially chaps. 2 and 3, for a review).

In addition to emphasizing the intrinsic nature of motivations to comply with moral norms, philosophers have also recognized the intrinsic nature of motivation to *punish norm violations*. Kant, famously, was a retributivist who held that punishment for violations of moral norms is a moral duty and is intrinsically valuable, and a substantial number of other philosophers have endorsed the retributivist position (Kant 1887/1972, 102–7; see Ezorsky 1972, ch. 2, sec. 2). Other philosophers associated with distinct moral traditions have also recognized the important role of duties to punish in the moral domain. Mill, for example, maintains that moral violations are the ones that we feel that society *ought to punish* (Mill 1863/1979, ch. 5). And a number of other philosophers have advanced similar claims (Gibbard 1990, ch. 3; Moore 1987). Here, again, we believe that these philosophical intuitions reflect a deep descriptive truth.

Before discussing the empirical literature on intrinsic motivation to punish, it's worth reemphasizing some of the caveats made earlier. In claiming that people are intrinsically motivated to punish norm violations, we are not claiming that these motivations *always* translate into punitive behaviors. Human motivations are multifaceted and complex, and **(p.293)** people with intrinsic motivations to punish a norm violator may also have instrumental motivations not to punish.

Thus motivations to punish serve to raise the probability of punitive behaviors, though they needn't translate into punitive behaviors in every instance. Furthermore, we are not claiming that *every* norm violation generates intrinsic motivations to punish. Rather, our claim is that norm violations that have the appropriate salience and severity generate motivations to punish. So while there is a *reliable connection* between norm violations and motivations to punish, this connection need not be realized in every occurrence of a norm violation.

There is a large anthropological and sociological literature attesting to the fact that norm violations elicit both punitive emotions like anger and outrage—and punitive behaviors like criticism, condemnation, avoidance, exclusion, or even physical harm—from most people within a society, and that these attitudes and behaviors are directed at rule violators (Roberts 1979; Sober and Wilson 1998). Furthermore, many social scientists have explicitly noted that punishment for norm violation, of this informal type, is *universally present in all societies*. For example, ostracism is a human universal (Brown 1991); gossip and criticism are human universals (Dunbar 1996; Wilson et al. 2000); and in all human groups, systems of sanctions, which utilize ostracism and gossip, as well as other informal sanctions, are applied to those who violate moral norms (Black 1998; Boehm 1999).

But here, again, it might be argued that, though there is ample evidence that people are disposed to punish norm violators, they do so for strictly selfish instrumental reasons. For example, people may punish to send a message to the violator, which produces a selfish gain for the punisher because the violator is deterred from repeating the offense. However, there is good evidence that motivations to punish are often truly intrinsic, and that punishment is not inflicted for selfish instrumental reasons alone.

One particularly striking finding is reported in Haidt et al. (submitted). In this study, subjects were shown films in which a normative transgression occurs. Subjects were offered various alternative endings; they preferred endings in which the perpetrators of the transgression were made to suffer, knew the suffering was repayment for the transgression, and suffered in a way that involved public humiliation. More revealingly, though, subjects were also offered an alternative ending in which the perpetrator realized what he did was wrong, showed genuine remorse, and grew personally as a result. Subjects' *rejection* of this ending suggests that their motivation to punish is not based on selfish instrumental ends, such as avoiding being harmed by the perpetrator in the future. Rather, they appear to be motivated by intrinsic motivations to punish the violator.

The most powerful evidence for intrinsic motivation to punish norm violations comes from experimental economics. Since the early 1990s, there has been a surge of interest in experimental economics in studying people's motivations to

punish in controlled laboratory conditions. A large number of studies show that in various experimental situations and experimental games, people will punish others—*at substantial costs to themselves*—for violations of normative rules or a normative conception of fairness. This data is particularly **(p.294)** powerful because it permits quantitative measures of the extent to which motivations to punish are unselfish and instrumentally irrational.

To illustrate the pattern of results in the literature, we'll describe a study by Fehr and Gachter (2002). In this study, 240 subjects played a public goods game in groups of four. Each member of the group was given 20 monetary units (MUs) and could either invest in a group project or keep the money for himself. For each unit invested, each of the four group members received four-tenths of an MU back. If a subject chose not to invest, he kept the full one unit. Given these payoffs, if all the subjects invest fully, each receives 32 units. If all subjects choose not to invest, each receives 20 units. Of course, if one subject chooses not to invest but the others invest fully, the "free-riding" subject receives the highest payoff, 44 MUs. Thus, the public goods game sets up a conflict between collective benefit and selfish interest.

Fehr and Gachter studied behavior in the public goods game under two conditions—a "punishment" condition and a "no punishment" condition. In the punishment condition, after each period of the game (a period consisted of one round of investment), subjects were informed of others' contributions and given an opportunity to punish any other player. Punishment cost 1 MU for the punisher and subtracted 3 MUs from the punished person's payoffs. Thus punishment was a costly act, but it created an even more substantial harm for the person being punished. Fehr and Gachter changed the composition of the group after each period, and ran the game for a total of six periods. Subjects did not know the identity of the members of the group in which they were placed (and all participants knew this fact), so a person could not personally benefit from the act of punishing, nor could a person build a reputation for contributing or punishing. Thus, to the extent that punishment deterred free-riding, the deterrence benefit was enjoyed by others. In the no-punishment condition, subjects played an identical game except for the fact that there was no opportunity to punish (Fehr and Gachter 2002).

The results of this study are quite striking, because they seem to violate a number of canons of self-interested economic rationality. First of all, Fehr and Gachter found that subjects in the no-punishment condition invested at much higher levels than self-interested rationality predicts, consistent with our previous claim that people follow norms of fairness as ultimate ends. In addition, in the punishment condition, Fehr and Gachter found that subjects punished, punished reliably, and punished severely. In the six periods of the experiment, 84.3 percent of the subjects punished at least once, and 34.3 percent punished five or more times during the six periods. Since subjects knew that they

switched groups after every period and that their identity remained anonymous after every switch, their motivations to punish cannot be explained in terms of selfish rationality.

A number of more recent studies have shown an even more striking result. In various experimental situations and games, people will punish others at some cost to themselves even if they are *merely observers* of violations of normative rules or some normative conception of fairness, and they themselves are not directly affected by the norm violation (Fehr and Fischbacher 2004; Carpenter, Matthews, and Okomboli 2004). In a way, the **(p.295)** existence of "third-party punishment" of this sort is actually fairly obvious and unsurprising (though it is very surprising from the standpoint of selfish rationality). Our everyday experience with human beings in a social context reveals that norm violations elicit powerful feelings of outrage from third parties who aren't directly harmed by the violation. In our view, the existence of third-party punishment of this sort shows, rather decisively, that punishment is not performed for mere instrumentally selfish reasons but rather is performed for intrinsic reasons.

One final point to make about punitive motivation is that, while children are given instruction (or at least some kind of social input) with respect to the *contents* of the norms of their social group, they are seldom, if ever, given input about the need to punish violations of norms. Thus it is remarkable that children who acquire normative rules systematically exhibit punitive attitudes toward those who violate the rules *without having been taught to exhibit these punitive attitudes*. For example, children who learn that hitting babies is wrong do not need to be taught that one should exhibit anger, hostility, and other punitive attitudes toward those who hit babies (Edwards 1987).

4. The Psychological Architecture Subserving Norms

In this section, we briefly sketch a theory about the psychological mechanisms underlying the acquisition and implementation of norms. The theory posits two closely linked *innate mechanisms*, one responsible for norm acquisition, the other for norm implementation. The function of the acquisition mechanism is to identify behavioral cues indicating that a norm prevails in the local cultural environment, to infer the content of that norm, and to pass information about the content of the norm on to the implementation system, where it is stored and used. The acquisition mechanism, we maintain, begins to operate quite early in development, and its operation is both automatic and involuntary. People do not need to turn it on, and they cannot turn it off—though it *may* be the case that the acquisition mechanism gradually turns itself off starting at some point after adolescence. The implementation mechanism performs a suite of functions, including maintaining a database of normative rules acquired by the acquisition mechanism, generating *intrinsic* motivation to comply with those rules as ultimate ends, detecting violations of the rules, and generating intrinsic

motivation to punish rule violators. Figure 12.1 is a “boxological” rendition of the mechanisms we’re positing.

The cluster of mechanisms we’ve sketched provides what we think is a plausible first pass at explaining many of the facts assembled in the previous two sections. The innate component dedicated to norm acquisition explains the fact that norms are universally present, that people acquire the norms of their own group, and that norm acquisition follows a reliable pattern of ontogenesis that starts quite early in life. The innate execution component explains why people are intrinsically motivated to comply with norms and intrinsically motivated to punish norm violators; it also explains why children **(p.296)** manifest punitive attitudes toward norm violators without having been taught to do so. Of course, positing mechanisms that perform the functions we’ve described is only the first step in theory building. Nonetheless, for two quite different reasons, we think it is an important step. First, it makes substantive claims about innate mechanisms subserving the acquisition and implementation of norms, and it is hard to see how the facts we’ve assembled in sections 2 and 3 *could* be explained without positing innate psychological mechanisms that perform the functions we’ve sketched. Second, while our boxology raises more questions than it answers, it also provides a systematic framework in which those questions can be addressed. In the section that follows, we’ll discuss *some* of the questions we think our theoretical framework brings into sharper focus. But before getting on to that, we should emphasize that the psychological mechanisms we’ve described are only *part* of what will inevitably be a much more complicated account of the way the mind deals with normative rules. Some of those further complications will be noted in section 5.

5. Some Open Questions

Obviously, there are *lots* of questions that the theoretical framework sketched in section 4 leaves unanswered. In this section, we’ll only have space to discuss six of them.

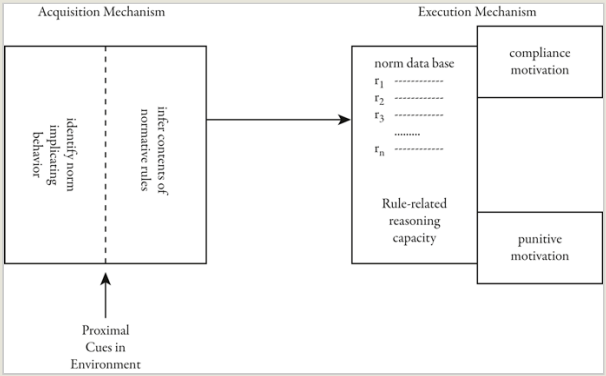

Figure 12.1 A first pass at a “boxological” sketch of the innate mechanisms underlying the acquisition and implementation of norms.

(p.297) 5.1 Norms Versus Moral Norms

In assembling our catalogue of social- and individual-level facts about norms, some of our claims were quite explicitly about *moral* norms, while others were about norms more generally. What is the relation between these two? As we noted in section 1, we think that norms, as we’ve

characterized them, are a theoretically important *natural kind* in the social sciences. It also strikes us as quite likely that the *intuitive* category of *moral* norms is not coextensive with the class of norms that can end up in the norm database posited by our theory. Perhaps the most obvious mismatch is that the norm database, for many people in many cultures, will include lots of rules governing what food can be eaten, how to dispose of the dead, how to show deference to high-ranking people, and a host of other matters that our commonsense intuition does not count as moral. So what *is* our commonsense intuition picking out? One possibility that might find encouragement in the influential literature on the "moral/conventional distinction" (Nucci 2001; Turiel 1983) is that moral rules or norms are *another natural kind*—either a subset of the norms in the norm database or a class of rules that includes some rules *that are in the* norm database and some that are not. Kelly and Stich (2007) have argued that experimental studies of the moral/conventional distinction do not support the claim that moral rules are a natural kind. But perhaps that conclusion can be reached by a different route. Another option is that our intuitions about which rules are moral are guided by a culturally local collection of prototypes or exemplars that have been heavily influenced by the Western religious and philosophical tradition, and that do not pick out a natural kind at all. A third possibility is that moral rules might turn out to constitute a natural kind that is identical with the norms characterized by our theory. On this view, our intuitions about which rules are moral are sometimes simply mistaken, in much the same way that the folk intuition that whales are a kind of fish was mistaken (Sripada, unpublished manuscript). Though empirical work on how people go about deciding that a rule is (or is not) a *moral* rule will surely be relevant to the debate among these three options, the debate also implicates contested issues on the border between semantics and metaphysics. And since progress in *those* areas is often hard to discern, we don't expect the matter to be settled anytime soon.

5.2 Proximal Cues

One of the jobs of the norm acquisition mechanism is to identify behavioral cues indicating that a norm prevails in the local cultural environment. What are those cues? Since norms, as we've characterized them, are rules whose violation is punished, it might be thought that the proximal cues for the acquisition processes must involve punishment. But we doubt that can be correct, because it is clear that some normative rules are acquired *before* the child observes a violation being punished, or even though the child *never* observes a rule violation at all. Another hypothesis about the proximal cues for norm acquisition comes from cognitive psychologist James Blair. Blair proposes that it is the **(p. 298)** display of *sad faces* by caretakers and others that, when paired with specific actions performed by the child, signals to the child that these actions count as normative transgressions. Evidence for this claim comes from the finding that psychopaths show abnormal emotional reaction to sad faces when

compared with normal subjects, and psychopaths also display specific deficits in moral reasoning, suggesting that they have failed to acquire normative rules appropriately (Blair 1995; Blair et al. 1997). However, in a convincing critique, Nichols (2004, ch. 1) argues that Blair's hypothesis is twice mistaken: sad faces are neither necessary nor sufficient to trigger norm acquisition.

There is intriguing evidence from the anthropological literature suggesting that the proximal cues facilitating norm acquisition at least partially consist of *explicit verbal instruction*. The psychologist Carolyn Pope Edwards analyzed records of day-to-day norm transgressions among children in a Luo-speaking community in Southern Kenya and in a toddler classroom in Poughkeepsie, New York. She found that children frequently receive repeated, *explicit verbal instruction* (and also verbal commands and threats) during the course of norm acquisition and development (Edwards 1987). However, the question of what proximal cues trigger the acquisition of norms is still very much open, and much further research is needed (see Nichols 2005 and Dwyer 2006 for further discussion).

5.3 Representational Format: How Are Norms Stored?

Many philosophers and psychologists who study norm-related reasoning assume that norms are stored in a sentence-like format regimented, perhaps, with the formalism of a deontic logic. However, we believe it is very much an open question whether this is the way norms are typically stored. The recent literature on the psychology of categorization suggests a number of plausible alternatives.

Exemplar theory (Murphy 2002; Smith and Medin 1981) offers a particularly intriguing option. On this account, norms might be stored as a cluster of exemplars, which can be thought of as representations of concrete, paradigmatic examples of actions that are required or prohibited by the norm. For example, people might store scenarios involving *hitting a defenseless child* and *stealing from the church collection plate* as exemplars of actions that are prohibited, and scenarios involving *keeping a deathbed promise* or *helping a stranger in distress* as exemplars of actions that are required. An exemplar-based theory of norm-guided judgment would propose that people judge novel actions in terms of their *similarity* to these stored exemplars—if an action is sufficiently similar to exemplars of prohibited actions, the action will be judged to be impermissible.[2] One way the exemplar-based account might work is that, in arriving at judgments of permissibility **(p.299)** or impermissibility, people search *exhaustively* through all of their stored exemplars, comparing each exemplar to the action being evaluated. On more complex (and in our view more plausible) versions of the exemplar-based account, it is not the case that *all* stored exemplars are accessed when making permissibility judgments. Rather, recent cognitive and emotional history serves to "prime," or activate, a subset of the relevant exemplars, and it is only this subset that is utilized in generating the

judgments. On this version of the exemplar-based account, a person may make different judgments about the same case on different occasions, because recent circumstances have primed different subsets of her stored exemplars. Stich (1993) has speculated that the exemplar-based account provides a plausible explanation for many aspects of moral judgment. For example, the account helps explain the importance of myths and parables in moral pedagogy, since these stories can help build a rich stock of exemplars of morally praiseworthy and morally blameworthy conduct. The exemplar-based account also provides a ready explanation of the fact that moral judgment seems so sensitive to factors (such as the emotional "spin" used in describing a case) that might prime one or another exemplar.

In addition to exemplar-based approaches, the literature on the psychology of categorization suggests a number of other ways of understanding the processes that underlie judgments of permissibility and impermissibility. The representational structures invoked might include prototypes, stereotypes, theories, and narratives among others (see Murphy 2002 for a comprehensive review). In addition, theorists have proposed connectionist-inspired theories of permissibility judgment (Casebeer 2003). An intriguing possibility is that different kinds of processes underlie permissibility judgments in different contexts, in much the same way that different exemplars might be activated in different contexts. For example, people might utilize an exemplar-based process for forming permissibility judgments in the context of day-to-day norm-related cognition, especially when such judgments are made rapidly and "on the fly." However, when there is ample time for reflection, they may seek to form permissibility judgments by carefully and deliberately assessing actions in terms of their relationship with stored general rules and principles. But all of this, we hasten to add, is no more than speculation. The empirical study of the representational format of norms has barely begun.

5.4 The Role of the Emotions

There is a long tradition in philosophy suggesting that emotions play a central role in the processes underlying moral judgment and moral behavior (Gibbard 1990; Hume 1739/1964). While there are many different ways that emotions might interact with the norm psychology we've sketched, we are inclined to think that the evidence is clearest for the involvement of emotions in the generation of punitive motivation directed at those who violate norms. Indeed, there is a substantial body of data suggesting that humans have universal, species-typical emotional structures that mediate motivations to punish. This evidence indicates that three phenomena are closely linked: normative rule **(p. 300)** violations, the experience of certain emotions—including disgust and contempt, but in particular anger—and the experience of strong motivations to punish the elicitor of the emotion (see Haidt 2003 for a review). Though the

relevant literature is enormous, it is not very cohesive. We'll give just a few illustrative examples.

Klaus Sherer and his colleagues undertook a large crosscultural study of emotions using a questionnaire method, and they found that subjects rate unfairness and immorality most highly as elicitors for the emotion of anger (Sherer 1997). David Sloan Wilson and Rick O'Gorman used a fictional scenario method and found that subjects invited to take the perspective of someone who is "wronged" experience anger, and that the strength of their anger is dependent on the importance of the fairness norm being violated (Wilson and O'Gorman 2003). In another study, Lawrence and his colleagues found that low doses of the dopamine receptor antagonist sulpiride produce selective deficits in a number of measures of anger, and also produce selective deficits in motivations to punish, as measured by subjects' willingness to punish others for violations of fairness norms (Lawrence et al. 2002; Lawrence, personal communication). We believe that these studies demonstrate a tight relationship between norm violations, emotional reactions, and motivations to punish, which in turn suggests that intrinsic motivations to punish norm violations are mediated by emotions.

In a particularly ingenious recent experiment, Wheatley and Haidt (2005) showed that emotion also seems to play a role in the production of moral *judgment*. The subjects in this experiment were hypnotized and told to feel disgust when they encountered the emotionally neutral words "take" or "often." Subjects were then asked to judge scenarios in which people behaved in morally problematic ways or in entirely unproblematic ways. Half of the subjects were given versions of the scenarios with the hypnotic cue word included, while the other half received nearly identical versions of the scenarios with the hypnotic cue word omitted. The presence of the hypnotic cue word in morally problematic scenarios led the subjects to asses the transgressions more harshly, while in the unproblematic scenarios, the presence of the word led a significant number of subjects to judge that the agent's actions were morally questionable. Findings like these suggest that emotions may play a role in producing moral judgments that subjects are aware of and can report. However, it is far from clear whether emotions *always* play a role in the generation of moral judgments. On the basis of neural imaging studies, Greene (2004) has suggested that there may be a second pathway leading to moral judgments—perhaps one in which explicit reasoning plays a role—that may not involve the emotions at all.

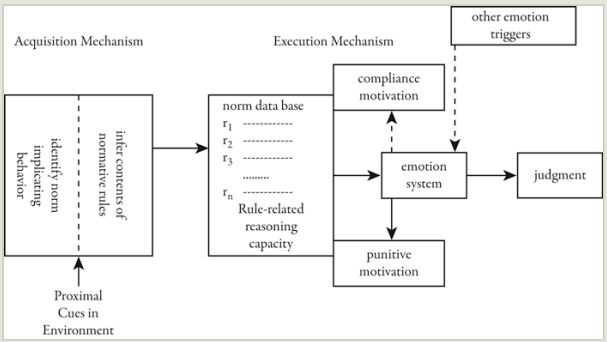

Figure 12.2 A more detailed sketch of the mechanisms underlying the acquisition and implementation of norms that includes the role of the emotion system. Solid lines indicate links that we take to be well supported by evidence; dotted lines indicate more speculative links.

We are heartened by the fact that serious empirical work on these issues has blossomed in recent years, though clearly there is still a great deal we do not know. It is tempting to speculate that, in addition to playing a role in generating punitive motivation, emotions also play a role in *compliance* motivation, though we have been unable to find any very persuasive evidence in support of this conjecture. In addition, since the emotion systems that are involved in the generation of moral judgments can be triggered by components **(p. 301)** of the mind other than the norm system, it would be very interesting, indeed, to know more about how that process works and how it influences moral judgment. In figure 12.2, we've added some components of the emotion system to the bare-bones boxology of figure 12.1.

5.5 The Role of Explicit Reasoning

Some of the most interesting and important questions about the psychology of norms focus on the role of explicit reasoning in shaping and justifying people's judgments and their behavior. Historically, philosophers, especially those in the Kantian tradition, and psychologists, especially those in the Kohlbergian tradition, have emphasized the role of explicit moral reasoning in the identification and acceptance of new normative rules and principles (Kohlberg, Levine, and Hewer 1983). Kohlbergians maintain that people pass through a sequence of *moral stages*. Earlier stages are characterized by egoistic kinds of thinking, while later stages are characterized by more objective and detached thought. According to Kohlberg, it is through a process of reasoning and reflection that people move away from earlier egoistic stages and come to adopt more objective perspectives that are supposed to be more acceptable from the standpoint of rationality.

The Kohlbergian picture seems to imply that reasoning or rationality can play a role in discovering *genuinely novel* moral principles, though we're inclined to be skeptical of this claim, since it is hard to see how pure rationality might discover novel moral principles *ex nihilo*. But there is another, more plausible way to interpret Kohlberg. Kohlberg frequently emphasizes the importance of "ideal perspective-taking" in moral reasoning **(p.302)** (Kohlberg 1981). The idea is

that people strive to find principles for resolving moral dilemmas that are *reversible*, in the sense that the principles apply irrespective of the particular role in the dilemma occupied by the subject. Kohlberg seems to suggest that it is a brute fact about human psychology that irreversible principles are seen as unsatisfactory, and are progressively replaced during the course of moral development by principles that are more fully reversible. So one way of understanding Kohlberg is that he is proposing that people hold a tacit moral "metaprinciple": Accept moral principles that pass the test of reversibility in preference to competing principles that are less fully reversible. On this interpretation, the metaprinciple isn't prescribed by pure reason alone, but it is nevertheless an important, and perhaps universal, principle that governs the operation of high-level reasoning in the moral domain. Another role for explicit moral reasoning in norm psychology is in *identifying inconsistencies* in one's preexisting moral beliefs, which in turn can lead to revisions in these beliefs. Moral philosophers often call this basic procedure of identifying inconsistencies in one's normative beliefs and making revisions and adjustments that enhance their overall consistency "the method of reflective equilibrium."

In the last two paragraphs, we've referred rather loosely to people's *moral beliefs* and the *moral principles* they accept. But how are these beliefs and principles related to the norms stored in the norm database our theory posits? One possibility is that they are *identical*—that moral beliefs and principles just are the entries (or perhaps a subset of the entries) in the norm database. If that is the case, and if explicit reasoning can modify moral beliefs in the ways we've described, then this sort of reasoning can modify the contents of the norm database. But, as we noted in section 4, we suspect that the norm psychology we've been elaborating is only one part of the complex system the mind exploits when dealing with normative rules. Thus it is entirely possible that the moral beliefs and principles that Kohlberg and others are concerned with are stored somewhere else in the mind. They might, for example, be stored in the "belief box," along with factual beliefs, or they might reside in a dedicated system that is distinct from the norm system. These two options, both of which are versions of what we call *the two sets of books hypothesis*, are broadly consonant with "dual attitude" and "dual processing" theories that have been proposed for a number of other psychological capacities (Chaiken and Trope 1999; Stanovich 1999; Wilson, Lindsey, and Schooler 2000). We suspect that some version of the two sets of books hypothesis is correct, though we would be the first to admit that evidence for the hypothesis is not thick on the ground. If the hypothesis is true, it would go a long way toward 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.

Wherever moral beliefs are stored, both the Kohlbergian and reflective equilibrium accounts of moral reasoning allow explicit moral reasoning and explicit moral beliefs to play an important causal role in determining the contents of people's moral judgments. For this reason, we can call both theories *rationalist* accounts of moral judgment. **(p.303)** Recently, however, the rationalist view has been challenged by the social psychologist Jonathan Haidt. According to Haidt, the casual relationship is often the reverse of that proposed in rationalist theories—rather than moral reasoning contributing to the formation of moral judgments, much moral reasoning is actually *post hoc justification*. Haidt argues that people's moral judgments are typically determined by their affective reactions to the case at hand, and they then use explicit reasoning processes to justify these antecedently arrived-at emotionally driven judgments.

In defending this "emotional dog and rational tail" picture, Haidt demonstrates the phenomenon he calls "moral dumbfounding" (Haidt 2001). Subjects are confronted with scenarios describing actions that most people consider to be unacceptable, but the scenarios are carefully contrived so that the typical reasons one might offer when asked why the action is wrong are not available. For example, one scenario is as follows.

> 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? (814)

Subjects immediately say that it was wrong for the siblings to make love. However, the typical reasons one might offer for this judgment—the danger of inbreeding, long-term emotional harm—don't apply in this case. Subjects nevertheless persist in their judgment that what the siblings did was wrong, saying something like "I don't know why, I can't explain it, I just know it's wrong" (Haidt 2001). According to Haidt, the phenomenon of moral dumbfounding suggests that quick emotion-driven systems play the primary role in generating at least some moral judgments. Explicit moral reasoning, by contrast, may often play the role of merely identifying socially acceptable justifications for these emotion-driven judgments.

In figure 12.3, we've supplemented figure 12.2 with various proposals about the role of explicit reasoning in moral judgment and moral belief formation.

5.6 Innate Constraints and Biases

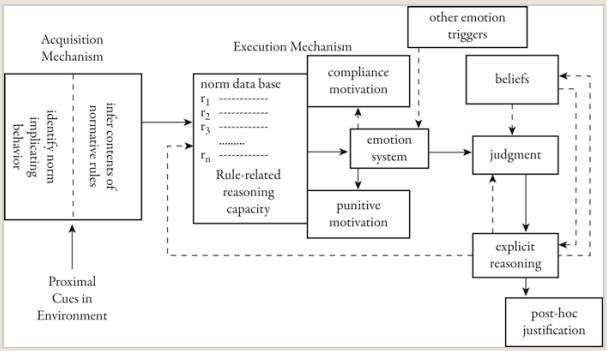

Figure 12.3 A sketch of the mechanisms underlying the acquisition and implementation of norms that includes various proposals about the role of explicit reasoning in moral judgment and moral belief formation. Solid lines indicate links that we take to be well supported by evidence; dotted lines indicate more speculative links with little empirical support.

On the theory we've sketched, the function of the norm acquisition mechanism is to identify norms in the surrounding cultural environment, infer their content, and pass that information along to the implementation component. One way to gain a deeper understanding of the norm acquisition process—and of the pattern of distribution of norms across cultures—is to explore the ways the acquisition system may be innately **(p.304)** constrained or biased. As a backdrop for thinking about these matters, we've found it useful to consider a *null hypothesis* that claims that the acquisition system exhibits no constraints or biases, and that it will acquire *all* and *only* those norms that are present in the child's cultural environment.[3] We've dubbed this "the Pac-Man thesis," inspired by the video game character that gobbles up everything it gets close to. If the Pac-Man thesis is true, then the norm acquisition system is equally unselective and unconstrained. There are, however, at least four ways in which the Pac-Man thesis *might* turn out to be false, and each of these corresponds to a distinct type of constraint or bias on norm acquisition.

Perhaps the most obvious way for the Pac-Man thesis to be mistaken is for some normative rules to be *innate*. Though there is a large philosophical literature debating the best interpretation of innateness claims in psychology (Cowie 1999; Griffiths 2002; Samuels 2002), for our purposes, we can consider a normative rule to be innate if various genetic and developmental factors make it the case that the rule would emerge in the norm database in a wide range of environmental conditions, even if (as a result of some extraordinary set of circumstances) the child's "cultural parents"—the people she encounters during the norm acquisition process—do *not* have the norm in *their* norm database. If there were innate norms of this sort, then they would almost certainly be cultural universals. Barring extraordinary circumstances, we should expect to find them in all human **(p.305)** groups. However, as we noted in section 2, the ethnographic and historical evidence does not support the existence of such exceptionless universals. So, while there is still much to be learned, we're

inclined to think that the available evidence does not support the existence of innate norms.

Another way the Pac-Man thesis might be false is that there might be an innately restricted set of possibilities from which all norms must be drawn during the course of acquisition. One way to unpack the idea of an innately restricted space of possibilities is by analogy with Noam Chomsky's *principles and parameters* approach to language learning (Chomsky 1988). According to Chomsky, the language faculty is associated with a set of parameters that can be set in various permissible ways. The child's linguistic experience serves to "toggle" the parameters associated with the language faculty, thus accounting for important aspects of the child's mature language competence. The parameters implicitly define the class of humanly learnable languages, so if a child were to be confronted with a language outside this class, the child would not learn it. A number of theorists have proposed that a broadly Chomskian principles and parameters model might provide a useful way to understand moral norm acquisition, and also serves to explain how norm variability is compatible with the existence of universal innate constraints (Harman 1999; Mikhail, Sorentino, and Spelke 1998; Nichols 2005; Stich 1993; Dwyer 2006), and recent experimental work by Marc Hauser and his colleagues suggests that there might indeed be universal constraints of a broadly Chomskian sort in the domain of harm norms (Hauser et al. 2007).

But there are other ways to understand the idea that norm acquisition is constrained by an innately restricted set of possibilities, ways that appear to be importantly distinct from the Chomskian principles and parameters model. For example, Alan Page Fiske has proposed that there are *four relational models* that structure all human social exchanges: communal sharing, equality matching, authority ranking, and market pricing (Fiske 1991). Fiske argues that the diversity of social arrangements and relationships found across human groups can ultimately be understood in terms of the operation of these four relational models. In addition, Richard Shweder and his colleagues have maintained that moral systems in all human societies are structured under one of the so-called big three families: community, authority, and divinity (Shweder et al. 1998). Paul Rozin and his colleagues expand on this idea with the proposal that each member of the big three family of moralities has an associated emotion that plays a primary role in mediating people's moral reactions—these emotions being contempt, anger, and disgust, respectively (Rozin et al. 1999). Though the ideas proposed by Fiske, Shweder and colleagues, and Rozin and colleagues are intriguing, it is not clear whether they are best understood as positing innate structures that serve to constrain or otherwise limit the space of moral norms that can be acquired, or whether they are positing some other kinds of psychological structures.

A third way for the Pac-Man thesis to be false would be as a result of the operation of what we call “Sperberian biases,” which we name after anthropologist Dan Sperber, who **(p.306)** has probably done more than anyone else to emphasize their importance (Sperber 1996). The Pac-Man thesis maintains that a child will always end up with an accurate copy of the norms of her cultural parents. But since no transmission process is error free, this sort of flawless copying is at best an idealization. Sometimes copying errors are random, but there are a variety of ways in which copying processes can give rise to *systematic* errors. For example, some sorts of normative rules may be more or less “attractive,” due to the way they interact with one’s preferences, aversions, emotions, and other elements of one’s psychology. For the same reasons, or for other reasons, some normative rules might be easier to detect (i.e., they may be more salient), easier to infer, or easier to remember, store, or recall. The transmission process will be influenced systematically by all these factors. When copying errors change less attractive rules into more attractive ones, the new rules will be more likely to be retained and transmitted, but when copying errors change more attractive rules into less attractive ones, the new rules will be more likely to be eliminated. It is these systematic processes affecting norm transmission that we call “Sperberian biases.” Sperberian biases are typically *weak*. They need not play a role in every instance of transmission from a cultural parent to a child, and often they will affect very few. Nevertheless, when their effects are summated over populations and over time, they generate a fairly strong population-level force that can have the effect of changing the distribution of norms in the direction favored by the Sperberian bias.

We can illustrate the operation of Sperberian biases by considering an example. Shaun Nichols has proposed that disgust acts as a Sperberian bias in the cultural transmission of etiquette norms (Nichols 2002). According to Nichols, *disgust* generates this bias by making certain kinds of etiquette rules more salient and more easily stored and recalled, and he marshals some intriguing evidence for these claims. Using data from sixteenth-century etiquette manuals from northern Europe, Nichols shows that etiquette rules whose violation engenders disgust are more likely to be part of today’s etiquette codes than rules whose violation fails to do so. This finding suggests that the cumulative operation of disgust as a bias on the transmission of etiquette rules has had the long-term effect of shifting the distribution of etiquette rules over time in the direction favored by the bias. In the same way that disgust might engender a Sperberian bias in the case of etiquette norms, it’s plausible that other cognitive structures, including various beliefs, preferences, aversions, and emotions, might engender Sperberian biases in the cultural transmission of other sorts of norms. We are inclined to think that the crosscultural distribution pattern of norms described in section 2 suggests that Sperberian biases have played a very powerful role in the transmission and evolution of norms. But making the case

for this conjecture is a substantial project that will have to wait for another occasion (see Sripada 2008).

A final way the Pac-Man thesis might be mistaken turns on the operation of biases of a very different sort. Thus far, we have been tacitly assuming that the cultural parents to whom a child is exposed all share the same norms. But obviously this is not always the case. Often a child will be exposed to cultural parents who have themselves internalized **(p.307)** significantly different norms. When this happens, the norm acquisition mechanism may utilize various selection principles, or *model selection biases*, in order to determine which cultural parent to copy. Various selection principles have been described in the literature (Boyd and Richerson 1985). These include a prestige bias leading the acquisition system to focus on a high-prestige person as a model, and age and gender biases that might, for example, focus the system on a model of the same sex who is slightly older. Alternatively, the acquisition system might rely on a conformity bias, adopting the cultural variant that is the most common. There is some evidence for age and gender biases in the transmission of norms (Harris 1998), and lots of evidence for prestige and conformity biases in the transmission of other cultural variants (Henrich and Boyd 1998; Henrich and Gil-White 2001). But how, exactly, this aspect of norm acquisition works is very much an open question.

6. Conclusion

Norms exert a powerful and pervasive influence on human behavior and human culture. Thus, the psychology of norms deserves to be a central topic of investigation in cognitive science. Our goal in this essay has been to provide a systematic framework for this endeavor. We've sketched the broad contours of a cluster of psychological mechanisms that can, we think, begin to explain some of the important facts about norms that have been recounted in various disciplines. Against the backdrop of the psychological architecture we've proposed, we've assembled a collection of open questions that the cognitive science of norms will have to address in the future. Clearly, in the study of the psychological processes that subserve norms, there is *lots* of work still to do. We will be very well satisfied indeed if our efforts provide a useful framework for organizing and integrating this work.

References

Bibliography references:

Barash, D. 1979. *The Whisperings Within*. Harper and Row.

Batson, C. D. 1991. *The Altruism Question*. Lawrence Erlbaum Associates.

Black, D. 1998. *The Social Structure of Right and Wrong*. Academic Press.

Blair, J. 1995. A Cognitive Developmental Approach to Morality. *Cognition* 57.

Blair, R., L. Jones, F. Clark, and M. Smith. 1997. The Psychopathic Individual: A Lack of Responsiveness to Distress Cues? *Psychophysiology* 34.

Boehm, C. 1999. *Hierarchy in the Forest*. Harvard University Press.

Bourguignon, E., and L. Greenbaum. 1973. *Diversity and Homogeneity in World Societies*. HRAF Press.

Boyd, R., and P. Richerson. 1985. *Culture and the Evolutionary Process*. University of Chicago Press.

(p.308) Brown, D. 1991. *Human Universals*. McGraw-Hill.

Carpenter, J., P. Matthews, and O. Okomboli. 2004. Why Punish? Social Reciprocity and the Enforcement of Prosocial Norms. *Journal of Evolutionary Economics* 14, 4.

Casebeer, W. 2003. *Natural Ethical Facts*. MIT Press.

Cashdan, E. 1989. Hunters and Gatherers: Economic Behavior in Bands. In S. Plattner, ed., *Economic Anthropology*. Stanford University Press.

Chagnon, N. 1992. *Yanomamö*. 4th ed. Harcourt Brace Jovanovich.

Chaiken, S., and Y. Trope. 1999. *Dual Process Theories in Social Science*. Guilford Press.

Chomsky, N. 1988. *Language and Problems of Knowledge*. MIT Press.

Cowie, F. 1999. *What's Within? Nativism Reconsidered*. Oxford University Press.

Cummins, D. 1996. Evidence for Deontic Reasoning in 3- and 4-year olds. *Memory and Cognition* 24.

Downs, A. 1957. *An Economic Theory of Democracy*. Harper Collins Publishers.

Dunbar, R. 1996. *Grooming, Gossip and the Evolution of Language*. Harvard University Press.

Durham, W. 1991. *Coevolution*. Stanford University Press.

Durkheim, E. 1903/1953. *Sociology and Philosophy*. Free Press.

Durkheim, E. 1912/1968. *The Elementary Forms of the Religious Life*. Allen and Unwin.

Dwyer, S. 2006. How Good Is the Linguistic Analogy? In P. Carruthers, S. Laurence, and S. Stich, eds., *The Innate Mind: Culture and Cognition*. Oxford University Press.

Edel, M., and A. Edel. 2000. *Anthropology and Ethics*. Transaction Publishers.

Edgerton, R. B. 1992. *Sick Societies*. Free Press.

Edwards, C. P. 1987. Culture and the Construction of Moral Values: A Comparative Ethnography of Moral Encounters in Two Cultural Settings. In J. Kagan and S. Lamb, eds., *The Emergence of Morality in Young Children*. University of Chicago Press.

Ezorsky, G., ed. 1972. *Philosophical Perspectives on Punishment*. State University of New York Press.

Fehr, E., and U. Fischbacher. 2004. Third Party Punishment and Social Norms. *Evolution and Human Behavior* 25, 2.

Fehr, E., and S. Gachter. 2002. Altruistic Punishment in Humans. *Nature* 415.

Fiske, A. P. 1991. *Structures of Social Life*. Free Press.

Frank, R. 1988. *Passion Within Reason*. W. W. Norton and Company.

Gibbard, A. 1990. *Wise Choices, Apt Feelings*. Harvard University Press.

Greene, G. 2004. fMRI Studies of Moral Judgment. Unpublished lecture given at the Dartmouth College Conference on the Psychology and Biology of Morality, Hanover, New Hampshire.

Griffiths, P. E. 2002. What Is Innateness? *Monist* 85, 1.

Haidt, J. 2001. The Emotional Dog and Its Rational Tail. *Psychological Review* 108, 4.

Haidt, J. 2003. The Moral Emotions. In R. J. Davidson, K. Scherer, and H. H. Goldsmith, eds., *Handbook of Affective Sciences*. Oxford University Press.

Haidt, J., J. Sabini, D. Gromet, and J. Darley. Submitted. What Exactly Makes Revenge Sweet?

Harman, G. 1999. Moral Philosophy and Linguistics. In K. Brinkmann, ed., *Proceedings of the 20th World Conference of Philosophy*, vol. 1: *Ethics*. Reprinted in G. Harman, *Explaining Value and Other Essays in Moral Philosophy*. Clarendon Press, 2000.

Harris, J. R. 1998. *The Nurture Assumption: Why Children Turn Out the Way They Do*. Free Press.

Hauser, M., F. Cushman, L. Young, R. Kang-Xing Jin, and J. Mikhail. 2007. A Dissociation Between Moral Judgments and Justifications. *Mind and Language* 22: 1–21.

(p.309) Henrich, J., and R. Boyd. 1998. The Evolution of Conformist Transmission and the Emergence of Between Group Differences. *Evolution and Human Behavior* 19.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, and H. Gintis. 2001. *Foundations of Human Sociality*. Oxford University Press.

Henrich, J., and F. Gil-White. 2001. The Evolution of Prestige: Freely Conferred Deference as a Mechanism for Enhancing the Benefits of Cultural Transmission. *Evolution and Human Behavior* 22.

Hume, D. 1739/1964. *A Treatise of Human Nature*. Clarendon Press.

Kant, I. 1887/1972. Justice and Punishment (from *Critique of Practical Reason*). In G. Ezorsky, ed., *Philosophical Perspectives on Punishment*. State University of New York Press.

Keeley, L. 1996. *War Before Civilization: The Myth of the Peaceful Savage*. Oxford University Press.

Kelly, D., and S. Stich. 2007. Two Theories About the Cognitive Architecture Underlying Morality. In P. Carruthers, S. Laurence, and S. Stich, eds., *The Innate Mind: Foundations and the Future*. Oxford University Press.

Kohlberg, L. 1981. Justice and Reversibility. In L. Kohlberg, *Essays on Moral Development*, vol. 1. Harper and Row.

Kohlberg, L., C. Levine, and A. Hewer. 1983. *Moral Stages: A Current Formulation and a Response to Critics*. Karger.

Lawrence, A. D., A. J. Calder, S. M. McGowan, and P. M. Grasby. 2002. Selective Disruption of the Recognition of Facial Expressions of Anger. *NeuroReport* 13, 6.

LeVine, R. A., and D. Campbell. 1972. *Ethnocentrism: Theories of Conflict, Ethnic Attitudes and Group Behavior*. John Wiley.

Marwell, G., and R. E. Ames. 1981. Economists Free Ride: Does Anyone Else? *Journal of Public Economics*.

McAdams, R. H. 1997. The Origin, Development, and Regulation of Social Norms. *Michigan Law Review* 96.

Mikhail, J., C. Sorentino, and E. Spelke. 1998. Towards a Universal Moral Grammar. In M. Gernsbacher and S. Derry, eds., *Proceedings, Twentieth Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum and Associates.

Mill, J. S. 1863/1979. *Utilitarianism*. Hackett.

Moore, M. S. 1987. The Moral Worth of Retribution. In F. Schoeman, ed., *Responsibility, Character and the Emotions*. Cambridge University Press.

Murdock, G. P. 1949. *Social Structure*. Free Press.

Murphy, G. L. 2002. *The Big Book of Concepts*. MIT Press.

Nichols, S. 2002. On the Genealogy of Norms: A Case for the Role of Emotion in Cultural Evolution. *Philosophy of Science* 69.

Nichols, S. 2004. *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford University Press.

Nichols, S. 2005. Innateness and Moral Psychology. In P. Carruthers, S. Laurence, and S. Stich, eds., *The Innate Mind: Structure and Contents*. Oxford University Press.

Nucci, L. P. 2001. *Education in the Moral Domain*. Cambridge University Press.

Parsons, T. 1952. *The Social System*. Free Press.

Petit, P. 1991. Virtus Normativa: Rational Choice Perspectives. *Ethics* 100, 4.

(p.310) Pilliavin, J. A., and H. W. Charng. 1990. Altruism: A Review of Recent Theory and Research. *American Sociological Review* 16.

Robarchek, C. A., and C. J. Robarchek. 1992. Cultures of War and Peace: A Comparative Study of Waorani and Semai. In J. Silverberg and P. Gray, eds., *Aggression and Peacefulness in Humans and Other Primates*. Oxford University Press.

Roberts, S. 1979. *Order and Dispute: An Introduction to Legal Anthropology*. St. Martin’s Press.

Rozin, P., L. Lowery, S. Imada, and J. Haidt. 1999. The CAD Triad Hypothesis: A Mapping Between Three Moral Emotions (Contempt, Anger, Disgust) and Three Moral Codes (Community, Autonomy, Divinity). *Journal of Personality and Social Psychology* 76.

Samuels, R. 2002. Nativism in Cognitive Science. *Mind and Language* 17.

Scott, J. F. 1971. *The Internalization of Norms*. Prentice-Hall.

Sherer, K. 1997. The Role of Culture in Emotion-Antecedent Appraisal. *Journal of Personality and Social Psychology* 73.

Shweder, R., M. Mahapatra, and J. Miller. 1987. Culture and Moral Development. In J. Kagan and S. Lamb, eds., *The Emergence of Morality in Young Children*. University of Chicago Press.

Shweder, R., N. Much, M. Mahapatra, and L. Park. 1998. The “Big Three” of Morality (Autonomy, Community, And Divinity), and the “Big Three” Explanations of Suffering. In A. Brandt and P. Rozin, eds., *Morality and Health*. Routledge.

Smith, E., and D. Medin. 1981. *Categories and Concepts*. Harvard University Press.

Sober, E., and D. S. Wilson. 1998. *Unto Others*. Harvard University Press.

Sperber, D. 1996. *Explaining Culture*. Blackwell Publishers.

Sripada, C. S. 2008. Nativism and Moral Psychology. In W. Sinnott-Armstrong, ed., *Moral Psychology*, vol. 1: *The Evolution of Morality: Adaptations and Innateness*. MIT Press.

Sripada, C. S. Unpublished manuscript. Carving the Social World at Its Joints: Conventions and Moral Norms as Natural Kinds.

Stanovich, K. 1999. *Who Is Rational?* Lawrence Erlbaum.

Stich, S. P. 1993. Moral Philosophy and Mental Representation. In M. Hechter, L. Nadel, and R. Michod, eds., *The Origin of Values*. De Gruyter.

Thaler, R. H. 1992. *The Winners’ Curse: Paradoxes and Anomalies in Economic Life*. Free Press.

Turiel, E. 1983. *The Development of Social Knowledge: Morality and Convention*. Cambridge University Press.

Wheatley, T., and J. Haidt. 2005. Hypnotically Induced Disgust Makes Moral Judgments More Severe. *Psychological Science* 16: 780–84.

Wilson, D. S., and R. O’Gorman. 2003. Emotions and Actions Associated with Norm Breaking Events. *Human Nature* 14, 3.

Wilson, D. S., C. Wilczynski, A. Wells, and L. Weiser. 2000. Gossip and Other Aspects of Language as Group-Level Adaptations. In C. Heyes and L. Huber, eds., *The Evolution of Cognition*. MIT Press.

Wilson, T. D., S. Lindsey, and T. Schooler. 2000. A Model of Dual Attitudes. *Psychological Review* 107.

Notes:

(1) . One issue we won’t consider is how the psychological mechanisms we’ll posit might have evolved. We believe that one of the advantages of the account we’ll offer is that there is a plausible account of the evolution of these mechanisms. But assembling this evolutionary story is a substantial project which we won’t attempt to undertake here.

(2) . The notion of “similarity” used in an exemplar-based account can be made precise in a number of different ways (see Murphy 2002 for a review). For our purposes, an intuitive notion of similarity will suffice.

(3) . Though we’ll usually describe the norm acquirer as a “child,” this is just a stylistic convenience—“norm acquirer” is a singularly awkward term. Whether and when the norm acquisition system shuts down, or slows down, as people mature, are open questions.