[Forthcoming in W. Sinnott-Armstrong (ed.), *Moral Psychology*. Oxford University Press.]

Is Morality Innate?

Jesse J. Prinz
jesse@subcortex.com

0. Introduction

> "[T]he Author of Nature has determin'd us to receive… a Moral Sense, to direct our Actions, and to give us still nobler Pleasures." (Hutcheson, 1725: 75)

Thus declares Francis Hutcheson, expressing a view widespread during the Enlightenment, and throughout the history of philosophy. According to this tradition, we are by nature moral, and ourS concern for good and evil is as natural to us as our capacity to feel pleasure and pain. The link between morality and human nature has been a common theme since ancient times, and, with the rise of modern empirical moral psychology, it remains equally popular today. Evolutionary ethicists, ethologists, developmental psychologists, social neuroscientists, and even some cultural anthropologists tend to agree that morality is part of the bioprogram (e.g., Cosmides & Tooby, 1992; de Waal, 1996; Haidt & Joseph, 2004; Hauser, 2006; Ruse, 1991; Sober & Wilson, 1998; Turiel, 2002). Recently, researchers have begun to look for moral modules in the brain, and they have been increasingly tempted to speculate about the moral acquisition device, and innate faculty for norm acquisition akin to celebrated language acquisition device, promulgated by Chomsky (Dwyer, 1999; Mikhail, 2000; Hauser, this volume). All this talk of modules and mechanism may make some shudder, especially if they recall that eugenics emerged out of an effort to find the biological sources of evil. Yet the tendency to postulate an innate moral faculty is almost irresistible. For one thing, it makes us appear nobler as a species, and for another, it offers an explanation of the fact that people in every corner of the globe seem to have moral rules. Moral nativism is, in this respect, an optimistic doctrine—one that makes our great big world seem comfortingly smaller.

I want to combat this alluring idea. I do not deny that morality is ecumenical, but I think it is not innate—at least that the current state of evidence is unpersuasive. Morality, like all human capacities, depends on having particular biological predispositions, but none of these, I submit, deserves to be called a moral faculty. Morality is a byproduct—accidental or invented—of faculties that evolved for other purposes. As such, morality is considerably more variable than the nativism program might lead us to think, and also more versatile. It is exciting to see cognitive scientists taking such an interest in morality, but that trend carries the risk of reification. I want to argue for a messier story, but I hope that the story leads to better understanding of how we come to care about the good. I think the nativist story oversells human decency, and undersells human potential.

I will survey a lot of ground here, and, of necessity all too quickly. With luck, a fast sweep of the landscape will suffice to sew seeds of doubt (see also Prinz, 1997; forthcoming; Nichols, 2005).

1. What is Morality?

Much in our life is governed by norms: the direction we face in the elevator, the way we walk, what we wear, the volume we speak in, the way we great people, and pretty much everything else that we do in our waking hours. Not all of these norms count as moral. My concern here is with moral norms. This is certainly not the place to fully defend an account of what distinguishes moral norms from other norms, but I will at least indicate where I stand. I think moral norms are sentimental norms: they are underwritten by various sentiments. In particular, moral norms are grounded in the moral emotions. Versions of this approach has a number of supporters today (Gibbard, Blackburn, Nichols, Haidt, D'Arms & Jacobson), and, importantly, it was also the preferred view of the British moralists who did more than any other group of philosophers to promote the view that morality is innate.

The British moralists claimed that moral norms are based on approbation and disapprobation. An action is right if, under idealized conditions, we would feel approbation towards those who do it; and an action is wrong if, under idealized conditions, we feel disapprobation. Idealized conditions include things like full knowledge and freedom from bias. The terms "approbation" and "disapprobation" are a bit antiquated, but I think they can be treated as umbrella terms for two classes ofS moral emotions: emotions of moral praise and emotions of moral blame. Emotions of praise include gratitude, esteem, and righteousness. Emotions of blame include other-directed-emotions, such as anger, contempt, disgust, resentment, and indignation as well as self-directed emotions, such as guilt and shame.

To count as a moral norm, these emotions must behave in particular ways. First, a norm does not count as moral simply in virtue of eliciting *one* of the aforementioned emotions in isolation. At a minimum, moral rules involve both self-directed emotions and other-directed emotions. You might feel disgusted when you see a friend cut her finger open accidentally, but you would not feel ashamed or guilty about doing that yourself. Second, our emotions must be directed at third parties if they are to ground moral norms. Moral concern arises when we are not directly involved. Third, mature moral judgments are enforced by meta-emotions. If you do something wrong and don't feel guilty, I will be angry at you for your conduct and for your lack of remorse. I am inclined to think that meta-emotions are necessary for moralizing, but I won't rest anything on that claim here. However, I do want to suggest that these three conditions are jointly sufficient for having a moral attitude, and that the first condition is necessary. To have a moral attitude towards ϕ-ing, one must have a moral sentiment that disposes a one to feel a self-directed emotion of blame for ϕ-ing, and an emotion of other-directed blame when some else ϕs.

I cannot adequately support the claim that moral norms are sentimental norms here, but I offer three brief lines of evidence.

First, psychologists have shown that moral judgments can be altered by eliciting emotions. For example, Weatley and Haidt (2005) hypnotized subjects to feel a pang of disgust whenever the heard an arbitrary neutral word, such as "often." They gave these subjects stories describing various individuals and asked them to make moral assessments. Compared to a control group, the hypnotized subjects gave significantly more negative moral appraisals when the key word was in the story, and they even morally condemned individuals whom control subjects described in positive terms.

Second, emotional deficits result in moral blindness. Psychopaths suffer from a profound deficit in negative emotions, including moral emotions (Cleckley, Hare, Patrick, Blair, Kiehl). They also have a profound deficit in their understanding more moral rules. When they talk about morality, Cleckley (1941) says they simple "say the words, but they cannot understand." Blair (1995) has shown that psychopaths fail to draw a distinction between moral and conventional rules; he argues that they regard moral rules as if they were merely conventional. The prevailing interpretation of these data is that psychopaths cannot form moral judgments because they lack the emotions on which those judgments ordinarily depend.

Third, there seems to be a conceptual link between emotions and moral judgments. Someone who was fully convinced that an action would maximize happiness can still believe that that action isn't morally good. Someone who believes that a action would lead to a practical contradiction when universalized does not necessarily believe that the action is morally bad. Likewise, I submit, for every other non-sentimentalist analysis of moral judgments. (This is the kernel of truth in Moore's open question argument.) But consider a person who feels rage at someone for performing an action and would feel guilty if she herself had performed that action. On my intuitions, such a person does thereby believe that the action is morally bad. She might revise her belief or question her belief, but one could correctly accuse her of moralizing. It is an open question whether the things we emotionally condemn are really wrong, but it is not an open question whether emotionally condemnation constitutes a moral attitude.

I will presuppose the sentimentalist account of moral judgment throughout this discussion, but not every argument will depend on it. In looking for innate morality, I will be looking for evidence that we are innately disposed to make moral judgments, and I will assume that those are based on moral emotions. The hypothesis that morality is innate is not simply the hypothesis that we are innately disposed to behave in ways that are morally praiseworthy. Bees are altruistic (they die to defend their colonies), but we do not say they have a moral sense. Innate good behavior can be used as evidence for an innate moral sense, but, as we will see, the evidence is far from decisive.

2. What is Innateness?

Defining innateness is itself a thorny philosophical problem. Some have argued that innateness is an incoherent concept. Others think that we can talk about genes innately coding for proteins, but not for phenotypes. I will assume that innateness is a defensible construct. Conceptually, the main idea is that a psychological phenotype P is innate if it is acquired by means of psychological mechanisms that are dedicated to P, as opposed to psychological mechanisms that evolved for some other purpose of for no purpose at all

(Compare, Cowie, 1999). This is only a rough characterization; I will identify cases of innateness using operational definitions.

Some innate traits are very rigid. They manifest themselves in a fixed way, and they are relatively impervious to change. A lot of insect behaviors are like this. They operate as if driven by simple, unchanging programs. I will refer to such traits as *buggy*. Humans may have some psychologically buggy traits. We feel fear when objects loom towards us, and we experience red when presented with ripe strawberries.

Some innate traits have a range of environmentally sensitive manifestations, but the range is highly circumscribed. Considered the blue-headed wrasse. These fish change genders under certain circumstances. When the male leader of a harem dies, the largest female become male. There are two settings (male and female), and environment selects between them. I will call such traits *wrassey*. If Chomsky is right, the human language faculty contains wrassey rules, which he calls parameters. For example, universal grammar says that prepositions can either proceed nouns or follow them, and primary linguistic data then sets the toggle on this switch.

Some innate traits allow much greater flexibility. Consider bird songs. Good mimics, like the starling, can imitate the songs of many other birds. In this respect, their songs have an open-ended range. But their capacity to imitate is not a general purpose learning system. It is evidently evolved for the function of learning songs. The actual songs are not innate, but they are the result of an innate song acquisition device. Some human traits are *starlingy*. Consider our capacity to abstract perceptual prototypes from experience, and use them for subsequent categorization—open-ended but evolved for that purpose.

When we encounter traits that are buggy, wrassey, or starlingy, we can say that they have an innate basis. In contrast there are traits that seem to be byproducts of other capacities. Skinner conditioned pigeons to play toy pianos. Playing pianos is not an innate capacity in pigeons; instead it is acquired using a general-purpose learning system (operant conditioning). We too learn much using general-purpose learning systems, as when we learn how to drive. In contrast, consider the lowly flea. Fleas do remarkable things in a flea circuses. For example, a group of fleas can play a game of toss using a tiny ball. The trick is achieved by coating the ball in a chemical that the fleas instinctively dislike, and when one comes into contact with the ball, it propels it away towards another flea, who then propels it away, and so on. Fleas are not using general purpose learning systems to achieve this behavior; they are using systems evolved for the specific purpose of avoiding noxious chemicals, but they are using those systems to do something other than what they were designed for. Some human capacities may be like flea toss. Projecting spitballs, for example, is learned using some general-purpose mechanism, but it also capitalizes on our capacity for spitting, which probably evolved for a specific function, not for deployment in prankish games.

Pigeon piano and flea toss give two models for traits that are not innate: general-purpose learning, and moonlighting mechanisms. We will see if morality can be understood on either model. If morality is innate, we should expect it to be buggy, wrassey, or starlingy. If morality were buggy, we might expect to find robust moral universals. If morality were wrassey, we might expect to see a fixed number of different variants on the same schematic moral rules. If morality were starlingy, we would expect that moral norm acquisition could not be explained by appeal to nonmoral learning

mechanisms. If any of these nativist views were right, we might expect morality to be modular and developmentally predictable. We might also expect to see precursors of morality in other species. I will consider attempts to defend nativism in each of these ways, and I will return to bugs, birds, and fish in the concluding section.

3. Is Morality Innate?

3.1. Universal Rules

One might begin developing a case for moral nativism by identifying universal moral rules. If certain rules can be found in just about every culture, that could be explained by the supposition that those rules are innate. This strategy parallels arguments that have been used to support linguistic nativism. If aspects of grammar are found in all languages, then those aspects of grammar may be innate. There are two risks in arguing from universality to innateness. First, some universals (or near universals) are not innate. Most cultures have fire, weapons, religion, clothing, art and marriage. Many also have taxes, vehicles, and schools. It is unlikely that any of these things are innate. Humans the world over face many of the same challenges, and they have the same cognitive resources. If these two are put together, the same solutions to challenges will often arise. Second, there are cases where universal traits are biologically based, but not domain specific. The sounds that are used in spoken languages are innate, in so far as our capacity to make those sounds and to hear perceive them categorically is biologically prepared. But these capacities probably weren't evolved for language. Chinchillas and birds can categorically perceive human speech sounds and rats can use acoustic information to distinguish between specific languages, such as Dutch and Japanese. To use universals to argue for innateness, a moral nativist should show (a) that there are moral universals, (b) that there are no plausible non-nativist explanation of these; and (c) that required innate machinery is specific to the domain of morality. Each of these points is difficult to establish. In this section, I will focus on a strong form of (a), according to which there are universal moral rules *with specific content*, as opposed to universal rule schema with variable content. For further discussion of this question, see Prinz (forthcoming).

What are some candidates for universal moral rules? One natural candidate is a general prohibition against harm, or at least against harming the innocent. Most people seem to have an aversive response to witnessing the suffering of others, and most people seem to avoid gratuitous acts of violence. This universal is believed by some to be innately based and to be a basic building block of human morality (Turiel, 2002; Blair, 1995).

Is there a universal prohibition against harm? The evidence is depressingly weak. Torture, war, spousal abuse, corporal punishment, belligerent games, painful initiations, and fighting are all extremely widespread. Tolerated harm is as common as its prohibition. There is also massive cultural variation in whom can be harmed and when. Within our own geographic boundaries, subcultures disagree about whether capital punishment, spanking, and violent sports are permissible. Globally, every extreme can be found. In the Amazon, Yanamomo warriors engage in an endless cycle of raiding and revenge (Chagnon, 1968). Among the Ilongot of Luzon, a boy was not considered a man

until he took the head of an innocent person in the next village; when he returned home, women would greet him with a chorus of cheers (Rosaldo, 1980). In the New Guinea highlands, there are many groups that engage in perpetual warfare; between 20 and 35 percent of recorded male deaths in these groups are due to homicide (the numbers for women are much lower) (Wrangham, 2004). Throughout geography and history, cannibalism has been a common practice, most indulgently pursued by the Aztecs who sometimes consumed tens of thousands in a single festival (Harris, 1986). Brutality is also commonplace in large-scale industrial societies. As a U.S. General recently said, "Actually, it's a lot of fun to fight. You know, it's a hell of a hoot. It's fun to shoot some people" (Schmitt, 2005).

Of course most cultures prohibit *some* harms, but there are nonnativist explanations for that. Such prohibitions are a precondition for social stability. Moreover, harming people often serves no end. We rarely have anything to gain from doing so. How would your life be improved if you could punch your neighbors regularly? Harming people only brings gains in special circumstances. Harming children who misbehave can keep them in line, harming criminals can serve as a deterrent, and harming enemies can allow us to obtain their goods. These, of course, are exactly the kinds of harms that have been tolerated by most cultures. So our universal prohibition against harm amounts to the platitude: "Harm when and only when the pros outweigh the cons." This is an empty mandate. It is just an instance of a general mandate to avoid gratuitous acts: "For any action A, do A when and only when pros outweigh the cons." There is no universal prohibition against harm, as such, just a prohibition against *l'acte gratuit*.

The triviality objection also counts against the tempting idea that we have a pro tanto reason to avoid harm. On this construal, the harm norm says, "Avoid harm unless something else there is an overriding reason to harm." This injunction is empty unless we can come up with a principled list of overriding reasons. If cultures can "overrule" harm norms more or less arbitrarily, then the pro tanto rule is equivalent to: "Avoid harm except in those cases where it's okay not to avoid harm." One can see that this is empty by noting that one can replace the word "harm" with absolutely any verb and get a true rule. The anthropological record suggests that the range of overriding factors is open-ended. We can harm people for punishment, for beauty, for conquest, and for fun. There is little reason to think these are principled exceptions to a rule that weighs on us heavily under all other circumstances. I suspect that harm avoidance is not even a universal impulse, much less a universal moral imperative.

This cynical response to the universality claim does not do justice to the fact that we don't like to see others in distress. Doesn't vicarious distress show that we have an innate predisposition to oppose harm? Perhaps, but it's not a moral predisposition. Consider the communicative value of a conspecific's scream. The distress of others alerts us to danger. Seeing someone suffer is like seeing a snake or a bear. It's an indication that trouble is near. It's totally unsurprising, then, that we find it stressful.

In response, nativists might reply that vicarious distress promotes prosocial behavior. Blair (1995) argues that vicarious distress is part of a violence inhibition mechanism. When fighting, an aggressor will withdraw when a sparing partner shows a sign of submission, including an expression of distress. Here, vicarious distress directly curbs violence. Doesn't this show that vicarious distress is part of a hard-wired moral capacity?

Probably not. Withdrawal of force is not a moral act. Submission triggers withdrawal because conspecific aggression probably evolved for dominance, not murder. Moreover, Blair's violence inhibition mechanism is highly speculative. His main evidence is that we experience vicarious distress when looking at pictures of people in pain; this, I just argued, may just be a danger-avoidance response. Admittedly, we inhibit serious violence when *play* fighting, but, by definition, play fighting is fighting between friendly parties. If you like the person you are roughhousing with, you are not going to draw blood or deliver a deathblow. No special violence inhibition mechanism is needed to explain that. This raises a further point. We are innately gregarious: we socialize, form attachments, and value company. Rather than presuming that we are innately disposed to avoid harm, we might say we are innately disposed to take pleasure in other people's company. Gregariousness is not, in and of itself, a moral disposition ("make friends" is not a moral injunction), but it may have implications for morality. We dislike it when our loved-ones are harmed. Human friendship promotes caring, which, in turn promotes the formation of rules that prohibit harm. Prohibitions against harm may be byproducts of the general positive regard we have for each other.

I am not persuaded, therefore, that we have a violence inhibition mechanism or a biologically programmed prohibition against harm. This conclusion may sound deeply unflattering to our species, but that is not my point. As I just indicated, I think we may be biologically prone to care about each other, and I also think there are universal constraints on stable societies, which tend to promote the construction of rules against harm. More generally, it must be noted that other species (e.g., squirrels, birds, and deer) don't go around killing each other constantly, but we are not tempted to say that they have moral rules against harm. They don't need such rules, because they have no biological predispositions to aggression against conspecifics. Likewise, we may have no such predispositions, so the lack of a biologically based prohibition against violence does not mean that we are nasty and vicious. I would surmise that our default tendencies are to be pretty pleasant to each other. The difficulty is that humans, unlike squirrels, can recognize through rational reflection, that violence can have positive payoffs. With that, there is considerable risk for nastiness, and that risk, not biology, drives the construction of harm norms. All this is armchair speculation, but it is enough to block any facile inference from pan-cultural harm norms to an innate moral rule. Harm prohibitions are not universal in form; they can be explained without innateness, through societal needs for stability; and the innate resources that contribute to harm prohibitions may not be moral in nature. In particular, harm avoidance may not be underwritten by moral sentiments.

I want to turn now to another pair of alleged moral universals: sharing and reciprocity. Human beings all over the world tend to share goods. Individuals don't hoard everything they obtain; they give it away to others. We tend to regard this as a morally commendable behavior, and failure to share is morally wrong. We also tend to reciprocate. If someone does us a good turn, we do something nice for them later. This is also moralized, and it is closely related to sharing. When we share, our acts of charity are often reciprocated, and we expect reciprocation, when possible. We condemn free riders, who accept offerings from others, but refuse to share.

Sharing and reciprocation are nearly universal, but they vary in significant ways across cultural boundaries. In some cultures, men eat meals before women and children,

and they are not expected to share to the same degree. In most cultures, there are people who do more than their fair share, and do not get adequately compensated. Among the Tasmanians, women apparently did the overwhelming majority of food collection, while men idled (Edgerton, 1992). In our own culture, the wealthy are expected to pay significant taxes, but they are certainly not expected to divide their profits. There are even apparently cultures where sharing is very rare. The Sirionó of Eastern Bolivia "constantly quarreled about food, accused one another of hoarding it, refused to share it with others, ate alone at night or in the forest and hid food from family members by, on the part of women, secreting it in their vaginas" (Edgerton, 1992: 13).

To assess cross-cultural differences in conceptions of fairness, a group of anthropologists recently conducted a series of studies in fifteen small-scale societies (Henrich et al. 2004). They asked members of these societies to play ultimatum games. The rules are simple. One player is given a certain amount of money and then told that she can keep some of it for herself, and offer some to a second player (who is not related to the first player); if the second player does not accept the offer, neither player gets anything. The ultimatum game tests for ideals of fairness, because the player making the offer is motivated to make offers that the other player will accept. If the other player considers the initial offer unfair, she will reject it. When done in the West, players tend to offer 45% on average. I am offered $100 dollars, I will offer you $45, taking more than half, but being generous. If I offered you considerably less, say $1 dollar, you would refuse to accept it out of spite, and I would lose my profit. When members of small-scale societies play the ultimatum game, everyone offers considerably more that 1%. Apparently, massively inequitable offers are rejected by most people everywhere, even when that 1% is a considerable sum. But there are still remarkable cultural differences. We tend to offer 45%. In some cultures, people offer more, and in some they offer less. Among the Machiguenga of Peru, the average sum offered was 26% and the most frequent offer was 15%, which is far below what most American subjects would consider fair. If I offered you $15 and took $85 for myself, you'd be sorely tempted to turn down the offer, and you would probably harbor a grudge. The Machiguenga have different standards, apparently. They may value sharing, but they clearly don't expect each other to share as much as we do in ultimatum games. Such findings suggest that there is not a fixed biological rule that drives us to share; the amount we share is variable.

Even so, the fact that people do share to some degree in must cultures suggests that there is a biological predisposition towards sharing—or so the nativist would argue. I am not fully convinced. Sharing also has non-nativist explanations. A person who has obtained a valued resource has strong incentives to share. Sharing helps avoid theft and it helps win friends. Sharing is a kind of insurance policy. If I give you something, you will be nice to me, and you may offer me something in the future. The nativist will be quick to respond that this reasoning presupposes reciprocity. I have no reason to think that you will share with me in the future unless people in general reciprocate acts of kindness. Mustn't reciprocity be innate? Perhaps not. There may be cultural explanations for why people reciprocate. Reciprocity promotes cooperation. If a culture has an economy that depends on cooperation, it will succeed only if reciprocity is promoted. This is true in the case of cultures that engage in heavy trade, cultures that have large-scale farms, and cultures that hunt very large game. Members of such cultures tend to offer equitable splits on the ultimatum game. The Machiguenga are foragers and

horticulturalists. Their farms are small, family-run, and temporary. So the Machiguenga do not depend heavily on non-kin, and it is unsurprising, then, that they do not make equitable offers. Thus, there is reason to think reciprocity emerges to serve cultural subsistence practices, and the evidence from variation in ultimatum games supports that hypothesis.

Indeed, some evolutionary game theorists argue that non-kin *reciprocity* must be a cultural construction. Biologically, our behavior is driven by genes, and our genes promote only those behaviors that increase their chances of being replicated. If generosity were genetically determined, and our genes led some of us to be generous, then free riders with stingy genes would take advantage, and the generous genes would die out. If, however, generosity were driven by cultural inculcation, then all normal members of a cultural group would equally likely to reciprocate, and the free rider problem would be reduced. Genes alone can't make us self-scarifying, but culture can. If this is right, then fairness and reciprocity are neither universal in form nor biologically based. I will return to this issue when I discuss animal behavior below.

Let me consider with one more example of a putatively universal moral norm: the incest taboo. Few of us feel especially inclined to have sexual relations with close kin. We are morally outraged when we hear about cases of incest, and just about every culture on record condemns incest in one form or another. There is even a genetic explanation for this universal. Inbreeding can lead to the spread of harmful recessive traits, so families that inbreed are likely to die out. Genes that promote exogamous behavior have a biological advantage. When combined with the apparent universality of incest prohibitions, we seem to have a pretty good case for moral nativism.

The evidence, however, is less secure on close examination. First of all there is massive cultural variation in which relationships count as incestuous. In some cultures, incest is restricted to the immediate family; in others it includes cousins; in some, only blood relatives are off limits; in others sex with affinal kin is equally taboo. The case of cousins is especially illustrative. In the contemporary Judeo-Christian West, sex with a first cousin is considered revolting; sex with a second cousin is more permissible but it still causes some people to snicker or look down their noses. The latter tendency may be a residue of the fact that the medieval Church prohibited marriage with cousins *up to seven degrees*. This was unprecedented. In the ancient world, cousin marriage was commonplace. For example, the Hebrew Bible tells us that Isaac was married to his cousin Rebecca, and Jacob was married to his two cousins, Rachel and Leah. In many contemporary cultures, cousin marriage is strongly encouraged. In one study, it was found that 57% of Pakistani couples were first cousins (Modell and Darr, 2002), and about the same rate of consanguineous marriages can be found in Saudi Arabia (El-Hamzi, et a1. 1995). There are also cultures that tolerate sexual relations between closer relatives. The Hebrew Bible contains explicit prohibitions against immediate family incest, but it also tells us that Abraham was married to his half-sister, and that Lot's daughters seduced their father and bore his children. Greco-Roman citizens in Ptolemaic Egypt married their full siblings at very high rates, Thonga hippopotamus hunters used to have sex with their daughters, and the ancient Zoroastrians allegedly encouraged all forms of immediate-family incest (see Prinz, 2007, for review).

Hauser (this volume) has rightfully argued that we should exercise great caution in drawing conclusions from exotic cases. In his defense of moral nativism, he warns

that rare exceptions cannot be used to refute the hypothesis that a particular rule is innate. I fully agree, and I don't want to place too much weight on Zorastrian sexual proclivities. So let me divide my critique of nativism about incest taboos into two parts. The first concerns sex outside the immediate family, such as the prohibition against first cousin incest. Violations of this rule are not exotic or unusual. 20% of the world's couples are estimated to be married to cousins (Bittles, 1990). There is no reason to think there is an innate taboo against sex outside of the immediate family. Now consider immediate family incest. Here, statistics tell in favor of a universal norm. Most cultures avoid sex with immediate kin. The exceptions show that these norms can be overridden by culture, not that the norms are learned. But the case against moral nativism about immediate family incest can be fought on other grounds. If procreating with immediate family members can cause a genetic depression, then there is reason to think we would evolve a tendency to avoid incest. This concession may look like it supports the case for moral nativism, but I think it actually does the opposite. Incest avoidance may be phylogenetically ancient. It may long predate the emergence of our species and the emergence of morality. If so, we may have an innate tendency to avoid incest, but not an innate moral rule against incest. If we naturally avoid something, we don't need a moral rule against it. If we are disgusted by rotting food, we don't need to have a moral rule to prevent us from eating it; we do not feel ashamed when we accidentally eat bread with mould on it, and we would not condemn another person for doing so. To turn incest avoidance into an incest taboo, a culture must punish those who engage in incest and condition perpetrators to feel ashamed. The transition from incest avoidance to incest taboos takes cultural effort. If this story is right, then there should be a large number of societies with no explicit moral prohibition against immediate family incest. This is exactly what the anthropological record seems to show. Thornhill (1991) found that only 44% of the world's cultures, in a large diverse sample, have immediate-family incest taboos. This casts doubt on the conjecture that there is an innate *moral* rule.

It would be hasty to draw any extravagant antinativist conclusions from the discussion so far. I have not demonstrated that there are no innate universal moral rules. Instead, I have argued that some of the most obvious candidates for innate universal moral rules are either not innate, or not universal, or not essentially moral. I think the considerations raised here suggest that it will be difficult to make a strong case for nativism by identifying universal moral rules. Moral nativists must rely on other evidence.

3.2 Universal Domains

I have been arguing that it is difficult to find examples of moral universals. The rules by which people abide vary across cultural boundaries. In response, the nativist might complain that I was looking for universals at too fine a grain. Perhaps specific moral precepts are variable, but broad moral categories are universal. By analogy, even if one did not find linguistic universals at the level of words, but one might expect to find universals at the level of syntactic categories. If we ascend to a more abstract level of moral competence, we might find that there are moral universals after all. This is an increasingly popular view among moral nativists. It is a version of what Hauser calls "temperate nativism," and defenders include Fiske (1991), Haidt and Joseph (2004), and

Shweder et al. (1997). One way to think about this approach is that we have several innate moral domains, which determine the kinds of situations that are amenable to moralization. The moral domains may even contain rule schema, whose variables get filled in by culture. For example, there might be an innate rule of the form (*x*)[Don't harm *x*, unless P]. Culture determines the scope of the quantifier (family, neighbors, all people, cows, fetuses, etc.), and the exceptions (initiation rights, revenge, deterrence, sports, etc.).

Rather than surveying all of the innate domain theories, I will focus on one recent example, which attempts to synthesize many of the others. Haidt and Joseph (2004) find that certain moral domains are mentioned more frequently than others when authors try to classify moral rules across cultures (and even across species). There are four domains that enjoy significant consensus. The first is the domain of suffering; all societies seem to have rules pertaining to the well being of others. The schematic harm prohibition might fall into this domain, along with rules that compel us to help the needy. The second domain concerns hierarchy; here we find rules of dominance and submission, which determine the distribution of power in a society. Next comes reciprocity; this is domain containing rules of exchange and fairness, like those discussed in the previous section. Finally, there is a domain of purity; these rules are especially prevalent in non-secular societies, but purity rules also include some dietary taboos and sexual mores. Haidt and Joseph believe that each domain corresponds to an innate mental module, and each kind of rule is regulated by a different family of emotions. Suffering elicits sympathy and compassion; hierarchies are enforced by resentment and respect; reciprocity violations provoke anger and guilt; and purity violations instill disgust. These domains are universal, but culture can determine the specific content of rules in each. What counts as an impermissible harm in one culture, may be morally compulsory in another. Thus, the moral domains do not furnish us with a universal morality, but rather with a universal menu of categories for moral construal. If we morally condemn some action it is in virtue of construing it as a violation in one of these domains.

The innate moral domains theory is a significant departure from the view that we have innate moral rules. It allows for considerable moral variation. In this regard, it is a significant departure from the Enlightenment moral sense theories, according to which human beings are naturally able to perceive objective moral truths. Nevertheless, it is a form of moral nativism, and it is my task here to assess its plausibility. As with the innate rule theories, there are three questions to ask: (a) are moral domains universal? (b) Can they be learned? And (c) are they essentially moral?

Let's begin with the question of universality. As Haidt and Joseph admit, the four moral domains are emphasized to a greater or lesser degree in different cultures. Our culture is especially preoccupied with suffering and reciprocity, whereas hierarchy and purity are more important in some parts of India. This affects how we construe moral transgression. Here is a simplified example. If someone is raped in the West, people sympathize with the victim and feel rage at the rapist. In India, there will be rage at the rapist, but there is also a tendency to think the victim has become adulterated, and that shame has been brought on her household potentially lowering their social status. This does not refute the hypothesis that the four domains are universal, but it does suggest that they do not play the same roles across cultures. And this raises the possibility that some of the domains are not construed as *morally* significant in every culture. In our culture,

we tend to resist moralizing impurities. In other research, Haidt et al. (1993) shows that American college students are disgusted when they hear about a man who masturbates into a chicken carcass, but they do not consider him immoral. In low socioeconomic status populations in Brazil, the same individual is morally condemned. Perhaps the purity domain has a moral status in those populations and not ours. On the sentimentalist theory that I am endorsing, this might be explained by saying that low SES Brazilians have a moral sentiment towards masturbating with a chicken carcass: they find it both disgusting and shameful, and they would be inclined to blame or punish offenders. Bourgeois Americans simply find such behavior yucky. To take another case, consider that Gahuku Gama headhunters in Papua New Guinea. According to Read (1955), they do not consider it immoral to cause harm, unless that harm comes to a member of their social group. On one interpretation, they do not moralize suffering, as such, but only hierarchy and reciprocity; they Gahuku Gama think the have responsibilities to the people who depend on them and on whom they depend. The upshot is that if the four domains are universal, it does not follow that they are universally moral.

Now consider the question of learning. Are the four domains necessarily innate, or is there an alternative explanation of how they emerge? One alternative is suggested by the fact that the domains are associated with different emotions. Let's grant that the emotions mentioned by Haidt and Joseph are innate. We are innately endowed with sympathy, respect, anger, disgust, and so on. These innate emotions may be sufficient to explain how moral domains emerge over development. To illustrate, consider purity. Suppose people are naturally disgusted by a variety of things, such as pollution, rotting meat, bodily fluids, disfigurement, and certain animals. This hodgepodge is unified by the fact that they all cause disgust. The disgust response has natural elicitors, but it can be extended to other things if those things can be construed as similar to the natural elicitors. For example, we can, through construal, view spitting, oral sex, eating insects, defecation, and body modification as disgusting. Now suppose, for whatever reason, that a particular society chooses to condemn some of these behaviors. That society will draw attention to the similarity between these behaviors and natural disgust elicitors, and it will inculcate feeling of both self- and other-directed blame for those who engage in them under certain circumstances. Once a society uses disgust to moralize certain behaviors, its members can be said to have a purity domain in their moral psychology. But, if this story is right, then the domain is a learned extension of a nonmoral emotion.

These remarks on universality and learning have both ended up in the same place. The four domains that Haidt and Joseph postulate may not be essentially moral. They may be outgrowths of universal emotions that evolved for something other than moral judgment. Each of the emotions they mention has nonmoral applications. We feel sympathy for the sick, but we do not make moral judgments about them; we feel respect for great musicians, but we do feel morally obligated to submit to their authority; we feel angry at those who frustrate our goals, but we do not necessarily think they are morally blameworthy for doing so; and we feel disgusted by rotting food, but we would not denounce the person who finds pleasure in it. The four moral domains may be byproducts of basic emotions. Negative emotions play a moral role only when they are transformed into full-blown moral sentiments. In particular, negative emotions become moral in significance only when we become disposed to feel corresponding emotions of blame towards self and others. Anger and disgust towards others take on a moral cast

only when we would feel blameworthy ourselves for behaving in a similar way. In sum, I think Haidt and Joseph's four domains may be universal, but I am not convinced that they are unlearned or essentially moral. Research has not shown that all people have have full-fledged moral sentiments towards behaviors in each of the four domains.

In response, the moral nativist might opt for a different strategy. Rather than looking for a family of different universal domains, nativists might look for a more fundamental divide; they might postulate a single domain of moral rules and distinguish these from nonmoral rules. On universal rule theories, some specific moral rules are universal; on universal domain theories, some general categories of moral rules are universal; on the strategy I want to consider now, the only thing that is universal is the divide between moral and nonmoral rules—it is universal that we have a morality, though the content of morality can vary in open-ended ways.

Here I am tempted to respond by pointing out that, at this level of abstraction, the postulation of moral universals does little explanatory work. Comapre Geertz: "That everywhere people mate and produce children, have some sense of mine and thine, and protect themselves in one fashion or another from rain and sun are neither false nor, from some points of view, unimportant; but they are hardly very much help in drawing a portrait of man" (1973: 40). If the only moral universal is the existence of morality itself, then an adequate account of human moral psychology will have to have to focus on culturally learned rules to gain any purchase on how we actually conduct our lives.

I do not want to let things rest with this dismissal. The claim that we have a universal disposition to create moral rules is not entirely empty. The nativist might even compare this disposition to bird songs. Starlings may not have any specific song in common, but their tendency to sing and to acquire songs by imitation is the consequence of an innate, domain-specific faculty. Surely the fact that all cultures have moral rules is an indication of an innate moral faculty, albeit a very flexible one. Mustn't the anti-nativist concede this much? I think not.

To make this case, I want to consider a popular version of the proposal that morality is a human universal. Turiel (2002), Song et al. (1987), Smetana (1995), and Nucci (2001) argue that, in all cultures, people distinguish between moral rules and rules that are merely conventional. For example, it's morally wrong to kick random strangers, but it is only conventionally wrong to wear pajamas to the office. Proponents of this view think that content of moral rules might be innately fixed; in particular, they think they might all be rules involving harms. In this sense, they might be regarded as defending a version of the innate domain theory according to which there is a single innate domain based on sympathy. I will not be concerned with that feature of the approach here. My main interest is the question of whether all cultures distinguish moral and conventional rules, whatever the content of those rules may be.

Defenders of the universal moral/conventional distinction test their hypothesis by operationalizing the difference between moral and conventional rules. Moral rules are said to have three defining characteristics: they are considered more serious than conventional rules; they are justified by appeal to their harmful effects on a victim; and they are regarded as objectively true, independent of what anyone happens to believe about them. Kicking a stranger is a serious offense; it is wrong because it causes pain; and it would be wrong even if the local authorities announced that it was acceptable to kick strangers. Wearing pajamas to the office is not very serious; it causes no pain; and it

would be acceptable if the authorities permitted it (imagine an office slumber party or a new fashion trend).

Smetana (1995) and Turiel (1998) survey evidence that this basic division is drawn across cultures, economic classes, religions, and age groups. It seems to be universal. They think that learning may play an important role in fostering sensitivity to this distinction, but the distinction itself is unlearned. Learning awakens innate understanding of the moral domain. I will return to the issue of learning below. For now, I also want to grant for now, that people can universally distinguish rules using the three criteria: some transgressions are serious, intrinsically harmful, and authority independent. What I want to question is whether these criteria really carve out a domain that deserves to be called morality. I want to argue that the criteria are neither necessary nor sufficient for accommodating rules that are pretheoretically regarded as moral. My discussion is heavily influence by Kelly and Stich (forthcoming), who offer a trenchant critique.

First consider seriousness. Some violations of pretheoretically moral rules are serious, but others are not. It is morally wrong to eat the last cookie in the house without offering to share, but not extremely wrong. Conversely, it seriously wrong to go to work naked, even though wearing clothing is just a societal convention. Next consider intrinsic harm. One might justify one's distaste for scarification by pointing out that it is intrinsically harmful, but this distaste reflects a personal preference, not a moral denunciation; many of us would say scarification is morally acceptable but intrinsically harmful. Conversely, some people regard certain actions as morally unacceptable, but not intrinsically harmful. For example, Haidt et al. (1993) found that some people regard it as morally wrong to wash a toilet with the national flag. Finally, consider authority independence. In many cultures people morally condemn behavior that is regarded as authority dependent. Jews, for example, sometimes say that certain dietary laws (e.g., combining dairy and meat) hold in virtue of divine command, and these laws would not hold if God had commanded different, and they do not hold for non-Jews. Smetana, Turiel, and Nucci can accommodate this case only by saying that Jews imagine God is harmed by diet violations, or by saying that Jews actually regard such rules as merely conventional. Both replies are flagrantly *ad hoc*. Conversely, there are many rules that are authority independent but not necessarily moral: we should cultivate our talents, we should avoid eating rotten meat, and we should take advice from those who are wiser than us. There are even cases of actions that are serious, instrinsically harmful, authority independent and, nevertheless, not immoral. Gratuitously sawing off one's own foot is an example.

In sum, the operational criteria used by Turiel, Nucci, and Smetana do not coincide perfectly with the pretheoretical understanding of the moral domain. If people are universally sensitive to these criteria, it does not follow that they universally comprehend the moral/conventional distinction. Indeed, there may be cultures where moral and conventional rules are inextricably bound. For many traditional societies, contingent social practices, including rules of diet and ornament, are construed morally. People who violate these rules are chastised and made to feel guilt or shame. The distinction is often blurry at best.

Indeed, I suspect that moral and conventional are two orthogonal dimensions. Consider a rule like, "don't harm a member of your in-group." Stated abstractly, this rule may not have any identifiable conventional component, but things change as soon as we

begin to make the rule specific enough to apply in practice. We can harm in-group members in initiation rights, for example, or in sporting events. Cultural conventions determine the scope of harm prohibitions. So we cannot fully specify such norms without appeal to some contingent features of culture. Consequently, we will have harm norms that are authority contingent: It is morally wrong to scar a teenager's face with a stone tool in this culture, but morally acceptable in cultures where the practice of scarification is embraced. Correlatively, rules that seem to be patently conventional have a moral dimension. It's conventionally wrong to wear shoes inside in Japan, but failure to comply with this rule is a form of disrespect, and the precept that we should respect others is moral. These examples suggest that rules of conduct generally have both moral and conventional components. The very same act can count as a moral violation or as a conventional violation depending on how it is described.

This last point is not intended as a rejection of the moral/conventional distinction. I think the distinction is real, but it is not a distinction between kinds of rules, but rather a distinction between components of rules. But how are we to distinguish those components? I think the moral dimensions of rules (Don't harm! Show respect!) are themselves culturally constructed. So, the distinction between moral dimensions of rules and conventional dimensions cannot be a distinction between absolute dimensions and culturally relative dimensions. Instead, I think the difference is psychological. There are dimensions of rules that we *regard* as moral, and dimensions rules that we *regard* as merely conventional. In keeping with the account of moral judgment that I offered earlier, I would say that the moral dimensions of rules are the dimensions that are psychologically grounded in moral sentiments. On my criteria, any dimension of a rule enforced by emotions of self-blame and other-blame and directed at third parties qualifies as a moral rule. When we say that a specific requirement is merely *conventional*, we express our belief that we would not blame (or at least we would try not to blame) someone who failed to conform to that rule in another culture. We do not blame Westerners for wearing shoes at home when they are in the West. When we say that it is *morally* wrong to disrespect others, we express our belief that we would blame someone for disrespecting others. Of course, the disposition to blame people for behaving in some way may itself be a culturally inculcated value.

I have been arguing that the moral/conventional distinction is more complicated than it initially appears, but I have not rejected that distinction completely. I have admitted that certain aspects of our rules are based on emotional patterns of blame, and others are not grounded in emotion. This gives the nativist a foothold. I have admitted that there is a way to distinguish the moral and the conventional, and the nativist is now in a position to propose that the distinction that I have just been presenting is universal. Nativists can say that all cultures have rules that are grounded in moral sentiments. I certainly don't know of any exceptions to this claim, but I am unwilling to infer that this is evidence for nativism.

In responding to Haidt and Joseph, I suggested that moral rules may emerge as byproducts of nonmoral emotions. If all cultures have rules grounded in moral sentiments, it does not follow that we have an innate moral domain. In all cultures, people have realized that behavior can be shaped by conditioning emotional responses. Parents penalize their children the world over to get them to behave in desirable ways. Some penalties have negative emotional consequences, as they thereby serve to foster

associations between behavior and negative emotions. This may be an important first step in the emergence of moral rules. Other steps will be required as well, and I will consider them in the concluding section of this chapter. The present point is that the universality of emotionally-grounded rules should not be altogether surprising given the fact that we shape behavior through penalizing the young. Neither the tendency to penalize, nor the resultant emotionally-grounded rules qualify as evidence for an innate moral code. But education through penalization could help to explain why emotionally-grounded rules (the building-blocks of morality) are found in all cultures.

3.3 Modularity

Thus far I have expressed skepticism about moral universals. If there are substantive universal moral rules or moral domains, they have yet to be identified. Moral nativists will have to look for other forms of evidence. One option is to look for moral modules in the brain. Innate faculties are often presumed to be both functionally and anatomically modular. To be functionally modular is, roughly, to process information specific to a particular domain. Modules are also sometimes said to be informationally encapsulated: they do not have access to information in other modules (Fodor, 1983; see Prinz, 2006, for a critique of Fodorian modules). To be anatomically modular is to be located within proprietary circuits of the brain. The language faculty is often presumed to be modular in both of these senses. We process language using language-specific rules and representations, and those rules and representations are implemented in specific regions of the brain with are vulnerable to selective deficits. Functional modularity provides some support for nativist claims, because capacities acquired using general cognitive resources often make use of rules and representations that are available to other domains. Anatomical modularity provides support for nativity claims because some of the best candidates for innate modules (e.g., the sensory systems and, perhaps, language) are anatomically localizable. If moral capacities could be shown to be functionally and anatomically modular that would help the case for moral nativism.

To explore this strategy, I will begin with some work by Cosmides and Tooby and their colleagues (1992). Cosmides and Tooby do not try to prove that there is a single coherent morality module. Rather, they argue that one specific aspect of moral reasoning is modular. (Presumably they think that future research will reveal that other aspects of moral reasoning are modular as well.) In particular, they say we have a module dedicated to reasoning about social exchanges, and this module contains inference rules that allow us to catch cheaters: individuals who receive benefits from others without paying the appropriate costs. To argue for functional modularity, they present subjects with a class of conditional reasoning problems called the Wason Selection Task. When presented with conditionals outside the moral domain, subjects perform very poorly on this task. For example, subjects might be told that, according to women's magazine, "If a woman eats salad, then she drinks diet soda." They are then asked about which women they would need to check to confirm whether this conditional is true. Subjects realize that they need to check women who eat salads, but they often don't realize that they must also check women who don't drink diet soda. In contrast, subject perform extremely well they are presented with conditionals that involve cheater-detection. For example, some subjects are told they need to check for violators of the rules, "If you watch TV, your

room has to be clean." Subjects immediately recognize that they must check people with dirty rooms and make sure that they are not watching TV. Cosmides and Tooby (1992) argue that, if people perform well on the cheater-detection task and poorly on a structurally analogous reasoning task, then cheater-detection probably recruits a specialized modular reasoning system. If there is a cheater-detection module, then that provides *prima facie* evidence for moral nativism.

As several critics have pointed out, there is a flaw is the argument for a cheater-detection module. To show that proprietary rules are being used for cheater-detection, Cosmides and Tooby must make sure that the control task is structurally analogous. The salad/soda case must be exactly like the TV/room case, except that one involves the moral domain and the other does not. But these cases are extremely different. In the salad/soda example, subjects are asked to determine whether a conditional is true, and in the TV/room case, they are asked to assume that the conditional is true, and find violators. Put differently, one task concerns a strict regularity, and the other concerns a rule. Regularities and rules are fundamentally different. If there are violators of an alleged strict regularity, the regularity must be false; if there are violators of a rule, the rule can be true. Moreover, rule violations elicit emotions, whereas regularity violations usually do not; and we are motivated to find violators of rules, because there are negative consequences if we do not. Thus, reasoning about rules and regularities should, on any account, recruit different resources, and we should be unsurprised to find that people are better at one than the other. To show that there is module dedicated to the moral task of detecting cheaters, Cosmides and Tooby cannot pit a regularity against a rule. They should pit non-moral rules against moral rules. There are many rules outside the moral domain. For example, there are prudential rules, such as "If you keep your guns in the house, then unload them." Like the moral rule, this one remains true, even if people don't conform to it. If subjects performed badly on prudential rules, but well on moral rules, that would be evidence for a moral module. But this is not what Cosmides and Tooby have found. Subjects perform well on both moral and prudential conditionals. This suggests that we have a general-purpose capacity for reasoning about rules, rather than a module restricted to the moral task of cheater-detection.

In response, Cosmides and Tooby proliferate modules. Rather than taking the striking similarities in moral and prudential reasoning as evidence for shared cognitive resources, they argue that there are two modules at work: one for prudential rules and one for cheater-detection. To support their case, they look for dissociations in performance on these two tasks. In healthy subjects, no dissociations have been found, but Stone et al. (2002) have identified an individual with a brain injury who can no longer perform well on cheater-detection conditionals, even though he continues to perform well on prudential conditionals. The fact that one capacity can be impaired without impairing the other suggests that cheater detection is both functionally modular and anatomically module—or so Stone et al. argue. But this conclusion does not follow. The patient in question does not have a selective deficit in cheater-detection, nor in moral reasoning. Instead, he has a large lesion compromising both his orbitofrontal cortex and his anterior temporal cortex (including the amygdala) in both hemisphere. The result is a range of deficits in social cognition. Stone et al. do not present a full neuropsychological profile, but they mention impairments in faux pas recognition and in comprehension of psychological vocabulary. Orbitofrontal regions are also implicated in the elicitation of social emotions and in the

assignment of emotional significance to social events. Given the size and location of the lesion, it is reasonably to presume that that the patient in this study has a range of general deficits in conceptualizing the social domain. These deficits are not restricted to cheater detection or moral cognition. To respond successfully on a cheater detection task, one may have to be able to respond emotionally to social stimuli. If this patient is unable to do that, then it is unsurprising that he performs poorly on moral conditionals. The problem is not that he has a broken moral module, but that he can't think well about the social domain.

The patient's general social cognition deficit could disrupt performance on the Wason task in two ways: first, he may not be able to recognize that the cheater-detection conditionals express rules, because that requires thinking about social obligations; second, even if he does comprehend that a rule is being expressed in these cases, he may not be able to elicit emotional concern about violations of that rule. In either case, the patient's abnormal conceptualization of the social domain may prevent him from inputing the cheater-detection conditionals into his general-purpose system for reasoning about rules. In other words, the patient does not provide evidence that cheater-detection involves any moral modules. His behavior can be explained by postulating general systems for thinking about the social domain and general systems for thinking about rules. I conclude that Stone et al. have not adequately supported the modularity hypothesis, and, thus, their data cannot be used to support moral nativism.

Before leaving this topic let me consider two more lines of research that might be taken as evidence for the modularity of morality. First, consider psychopaths. Psychopaths have IQ scores within the normal range, and they perform relatively well on most standard aptitude tests, but they are profoundly impaired in moral competence. As noted above, psychopaths do not distinguish between moral and conventional rules (Blair, 1995). Blair concludes that psychopaths have a selective deficit in moral competence, and he associated this deficit with abnormalities in the their central nervous systems. In particular, some psychopaths have reduced cell volumes in parts of frontal cortex and the amygdala. This suggests that there is a moral module in the brain.

My response to this argument is already implicit in my discussion of psychopaths in section 1. Psychopaths do not have a selective deficit. They have profound deficiencies in *all negative emotions*. This is a diagnostic symptom of psychopathy, which is easy to observe, and it has been confirmed in numerous laboratory tests. Psychopaths are less amenable than control subjects to normal fear conditioning (Birbaumer et al., 2005), they have diminished startle potentiation (Patrick, 1994), little depression (Lovelace and Gannon, 1999), high pain thresholds when compared to non-criminals (1993), and difficulties in recognizing facial expressions of sadness, anger, and disgust (Stevens et al. 2001; Kosson et al. 2002). Without negative emotions, psychopaths cannot undergo the kind of conditioning process that allows us to build up moral rules from basic emotions. Psychopathy is not a moral deficit, but an emotional deficit with moral consequences.

The final line of research that I will consider is based on neuroimaging of healthy individuals when they engage in moral perception. Moll et al. (2002) tried to identify moral circuits in the brain by comparing neuronal response to pictures of moral scenes and neuronal responses to unpleasant pictures that lack moral significance. For example, in the moral condition, subjects view pictures of physical assaults, abandoned children,

and war. In the nonmoral condition, they see body lesions, dangerous animals, and body products. Moll et al. report that, when compared to the nonmoral condition, moral photographs cause increased activation in orbital frontal cortex and medial frontal gyrus. The authors conclude that these areas play a critical role in moral appraisals. It is tempting to say that this study has identified a moral module—an area of the brain dedicated to moral cognition.

That interpretation is unwarranted. First of all, the brain structures in question are implicated in many social cognition tasks, so we do not have reason to think they are specialized for moral appraisals. Second, there is considerable overlap between the moral picture condition and the unpleasant picture condition. Both cause increased activation limbic areas like the amygdala and insular cortex, as well as visual areas (due presumably to increased attention to the photographs). It is reasonable to infer that negative pictures, whether moral or nonmoral result in activation of similar emotions, as indicated by the overlapping limbic response. The main difference between the two kinds of pictures is that the moral pictures also elicit activity in brain centers associated with social cognition. This is unsurprising: seeing a child is more likely to induce a social response than seeing a body product. So, Moll et al. have not proven that there is a moral module. Their results support the opposite conclusion: moral stimuli recruit domain-general emotion regions and regions associated with all manner of social reasoning (as we saw in discussion of the Wason task). The study does not reveal any regions that are distinctively moral (for a similar assessment, see Greene and Haidt, 2002).

I conclude that there is no strong evidence for a functional or anatomical module in the moral domain. Nativists must look elsewhere to support their view.

3.4 Poverty of the Stimulus

In linguistics, the best arguments for innateness take the following form: children at age *n* have linguistic rule R; children at age *n* have not had exposure to enough linguistic data to select rule R from many other rules using domain-general learning capacities; therefore, the space of possible rules from which the select must be innately constrained by domain-specific learning capacity. Arguments of this form are called arguments from the poverty of the stimulus. The case for moral nativism would be very strong if nativists could identify poverty of stimulus arguments in the moral domain. I will consider two attempts to defend moral nativism along these lines (see also Nichols for further discussion, 2005).

The first argument owes to Dwyer (1999). She has been one of the most foreceful and articulate defenders of the analogy between language and morality (two other important proponents are Mikhail, 2000; and Hauser, this volume). Dwyer focuses on the moral/conventional distinction. She notes that children begin to show sensitivity to this distinction at a very young age (between 2 and 3), yet they are not given explicit instruction. Parents do not verbally articulate the distinction between the two kinds of rules, and the penalize children for transgressions of both. In addition, there are considerable variations in parenting styles, yet children all over the world seem to end up understanding the distinction. Dwyer takes this as evidence for an innate capacity.

I think this is exactly the kind of argument that moral nativists should be constructing, but I don't think this particular instance of it succeeds. Above, I raised

some general worries about the moral/conventional distinction, but I want to put those to the side. Let's assume that the distinction is real and that the operationalization offered by people like Turiel, Nucci, and Smetana captures it successfully. The question before us is whether this distinction can be acquired without an innate moral capacity. I think it can.

To begin with, we are assuming along with researchers in this tradition that moral and conventional rules are associated with different patterns of reasoning. Moral transgressions are regarded as more serious, more harmful, and less contingent on authorities. These reasoning patterns are exhibited by both children and adults. Therefore, children are presumably exposed to these different reasoning styles. The stimuli to which they are exposed are not impoverished. They can learn how to differentiate moral and conventional rules by imitating and internalizing the different reasoning patterns in the moral educators.

This is not idle speculation. There is ample evidence that parents adapt their styles of disciplinary intervention to the type of rule that a child violates (see Smetana, 1989; and Grusec and Goodnow, 1994, for a review). Moral rule violations are likely to be enforced using power assertion and appeals to rights, and conventional rules are likely to be enforced by reasoning and appeals to social order. Differential rule enforcement has been observed among parents of 3-year-olds, and is presumably operative before that age as well (Nucci and Weber, 1995). This is consistent with anecdotal evidence. I was recently at a party with four 1.5-year-olds, and I made three casual observations: these children did not show remorse when they harmed each other; at such moments parents intervened with angry chastisement, social ostracism ("sit in the corner"), and reparative demands ("say you're sorry"); and parents never exhibited anger or punitive responses when children violated conventional norms, such as rules of etiquette. Grusec and Goodnow (1994) cite evidence that differential disciplinary styles are also used cross-culturally in Japan and India. In addition, children get socialized into moral competence by observation of adults outside of the household and from social interactions with peers. A child who violates a conventional rule may be ridiculed by peers, but she is unlikely to incur worse than that (imagine a child who wears pajamas to school one day). A child who violates a moral rule, however, is likely to incur her peers' wrath (imagine a child who starts fights). In short, different kinds of misdeeds have different ramifications, and a child is surely cognizant of this.

Dwyer might respond by conceding that children get enough feedback to know which of their own misdeeds are moral transgressions, but she might insist that they don't get enough feedback to generalize from those misdeeds to other actions that they have never experienced. Children do some bad things, but they do not commit every possible moral transgression. A child may learn from experience that it is bad to be a bully, but a child cannot learn from experience that it is bad to be an axe murderer or an embezzler. In other words, the child faces a challenging induction problem: how to generalize from a few examples of juvenile misconduct to whole class of moral wrongs. Mustn't a child have innate moral rules to extend the category? I don't think so. Adults explicitly tell children not to harm others, and this formula can generalize to novel cases. In some moral domains, generalization from familiar cases to novel cases may be harder (hierarchy norms and sexual mores come to mind), and here, I would predict that children do a bad job at predicting adult moral values. Nativists need to come up with an example

of an inductive inference that children make in spite of insufficient instruction. I am not aware of any such case.

Let me turn from the moral/conventional distinction to another argument from the poverty of the moral stimulus. To prove that morality is innate, we might look for signs of moral sensitivity in individuals who have had no moral training. There is virtually no data available on this question because it would be unethical to raise children without moral guidance. There is, however, one anecdote worth reporting. In 1799, a boy estimated to be twelve-years-old emerged from a forest in Saint Sernin sur Rance in France. He had apparently grown up alone in the woods without any adult supervision. The boy was given the name "Victor," after a character in a popular play, and he was placed in the care of Jean-Marc-Gaspard Itard, a young physician working in Paris. Itard tried to civilize Victor, and he wrote a book about his efforts. In one poignant episode, Itard attempted to discover with Victor had a sense of justice. I quote at length:

> [A]fter keeping Victor occupied for over two hours with our instructional procedure I was satisfied both with his obedience and his intelligence, and had only praises and rewards to lavish upon him. He doubtless expected them, to judge from the air of pleasure which spread over his whole face and bodily attitude. But what was his astonishment, instead of receiving the accustomed rewards … to see me suddenly … scatter his books and cards into all corners of the room and finally sieze upon him by the arm and drag him violently towards a dark closet which had sometimes been used as his prison at the beginning of his stay in Paris. He allowed himself to be taken along quietly until he almost reached the threshold of the door. There suddenly abandoning his usual attitude of obedience, he arched himself by his feet and hands against the door posts, and set up a most vigorous resistance against me, which delighted me…because, always ready to submit to punishment when it was merited, he had never before … refused for a single moment to submit…[U]sing all my force I tried to lift him from the ground in order to drag him into the room. This last attempt excited all his fury. Outraged with indignation and red anger, he struggled in my arms with a violence which for some moments rendered my efforts fruitless; but finally, feeling himself giving way to the power of might, he fell back upon the last resource of the weak, and flew at my hand, leaving there a deep thrash of his teeth. (Itard, 1801: 94-5)

Itard calls this "incontestable proof that [Victor had] the feeling of justice and injustice, that eternal basis of social order" (95). For our purposes, it is relevant as a possible case of a poverty of the stimulus argument. Victor shows sensitivity to injustice despite having been raised (as it were) by wolves. This is an apparent example of moral competence without moral education. Poverty of the stimulus.

Or is it? Itard himself might deny this interpretation. He was eager to take credit for Victor's moral education. Itard says, "On leaving the forest, our savage was so little susceptible to this sense [of justice] that for a long time it was necessary to watch him carefully in order to prevent him from indulging in his insatiable rapacity" (p. 93). Itard subjects Victor to increasingly sever punishments to prevent him from thieving. Victor's subsequent sense of justice may have been implanted through this process. Alternatively,

Itard may have misdescribed his pupil's mindset in the preceding episode. Victor had always been rewarded for doing his lessons well, and he had come to expect the reward of Itard's praises. On the occasion of this experiment, Itard replaced praises with wrath, and Victor reacted violently. This is hardly surprising. Victor was known for erratic tantrums, and he was accustomed to routine. In this case, his tantrum might have been set off by Itard's unanticipated assault. Even rats can react violently when expected reward in suddenly replaced by punishment. Rats do not need a moral sense to have a strong response under conditions of radically reversed reinforcement. For all we know, Victor had no more moral competence than a pestilent rodent.

Poverty of the stimulus arguments are powerful tools in making a case for nativism. Perhaps such arguments will ultimately be found in the moral domain, but current evidence is consistent with the conclusion that children acquire moral competence through experience.

3.5 Fixed Developmental Order

Linguistic nativists sometimes argue for their cause by pointing out that language emerged in predictable ways. Children pass through similar stages in linguistic development, at similar ages, and arrive at linguistic competence around the same time. This is taken as evidence for the conclusion that language unfolds on an endogenously controlled schedule of maturation, like the secondary sex characteristics. If language were learned using general learning mechanisms, we would expect to see greater individual differences. People, after all, have different learning styles, bring different amounts of knowledge to bear, and are exposed to different experiences. One could argue for moral nativism by showing that moral development unfolds in a predictable way. One could argue that there is a fixed schedule of moral stages. Moral nativists who want to pursue this strategy might be inclined to call on the most famous theory of moral development: the theory of Lawrence Kohlberg (1984).

According to Kohlberg, there are six stages of moral development. The first two stages are "preconventional." At stage 1, children behave well out of fear of punishment, and, at stage 2, children chose behaviors that the view to be in their own best interest. The next two stages are called "conventional" because children become sensitive to the fact that certain behaviors are expected by members of their society. Stage 3 ushers in a "good boy, good girl" orientation, in which children want to be well regarded by others. At stage 4, we become preoccupied with law and order, choosing actions that conform to social norms and promote social stability. The final two stages in Kohlberg's framework are post-conventional. At stage 5, people justify the actions that they feel obligated to perform by appealing to a social contract, and, at stage 6, people attain a principled conscience; they selection actions on the basis of universal principles, rather than local customs. Kohlberg assumes that these stages have a kind of ontogenetic inflexibility: we pass through them in a fixed sequence. Nativists should like this picture because it parallels the stage-like progression of language, which is widely believed to be innate.

Ironically, Kohlberg does not argue that morality unfolds through a maturation process. He does not think his stages are biologically programmed. As a student of Piaget, he argues instead that social experiences cause children to reason about their current views, and this process of reflections prompts progressive improvements. Each

stage is a rational successor to its predecessor. It would be a mistake to call this an *anti*nativist view, but neither is it a nativist view. Nativists cannot find a true ally in Kohlberg. Still, they might abandon his Piagetian orientation and give his levels a nativist spin.

This strategy is unpromising, because Kohlberg's theory is deeply flawed. First of all, it is a theory of how we morally reason, not a theory of how we form moral opinions. One might think our opinions are based on reasoning, but this probably isn't the case. There is evidence that moral reasoning is a posthoc process that we use to justify moral opinions that are acquired in some non-rational way (Haidt, 2001). If that's right, stage-like advances in moral reasoning may reflect a domain-general advance in rational capacities, not a change in our moral faculty. The moral opinions we have may be acquired on mother's (bended) knee, and rationalized through progressively sophisticated arguments.

Second, empirical evidence has not confirmed a linear progression through Kohlberg's levels. Critics point out that people often reason at multiple levels at once, they occasionally skip stages, and they sometime move backwards through Kohlberg's sequence (Krebs et al. 1991; Puka, 1994). Evidence has also failed to establish that people advance to the highest stages in Kohlberg's scale. There was so little support for reasoning at stage 6, that Kohlberg regarded it as merely theoretical (Colby et al., 1983). It turns out that most adults only reliably attain stage 4 competence, and they make it to this point in their late teens or twenties; even graduate students have been shown to be stage 4 moral reasoners (Mwamwenda, 1991). This is somewhat embarrassing for the moral nativist, because cognitive capacities that are widely believed to be innate tend to emerge earlier in life. It is also noteworthy that the moral stage attained correlates with the degree of education, suggesting that moral reasoning skills are the result of training rather than maturation (Dawson, 2002).

There is also cross-cultural variation (Snarey, 1985). In small-scale village and tribal societies, people reason only at Kohlberg's third stage of development (Edwards, 1980). This does not undermine Kohlberg's theory, because he claims that environmental stimulation contributes to moral development, but it is a devastating blow to the nativist who wants to argue that Kohlberg's stages reflect the unfolding of a biogrogram. Puberty does not have radically different onset times in small-scale societies, and nor does language acquisition.

There are other objections to Kohlberg (including Gilligan's, 1982, feminist critique), but this brief survey should suffice. The evidence for a fixed sequence of moral stages is underwhelming; and, to the extent such stages exist, there is little reason to think they are the result of biological maturation. Moral nativists cannot find in Kohlberg the requisite parallel to the stage-like progression in language acquisition.

3.6 Animal Precursors

I will discuss just one more class of arguments for moral nativism. Nativists often use cross-species comparisons to support their views. These comparisons can work in two ways. First, nativists can establish that other species lack some human trait, despite having similar perceptual and associative reasoning capacities. Such contrasts can show that the trait in question is not acquired through perception or conditioning, and this can

be used to support the conclusion that the trait requires domain-specific learning mechanisms. Alternatively, nativists can establish that other species have a rudimentary version of some human trait, and this can be used to support the conclusion that the trait emerged through incremental biological evolution. In moral psychology, it has become increasingly popular to pursue this latter strategy. Researchers look for animal homologues of human moral traits. Most resist the temptation to say that non-human animals have a moral sense, but it is widely believed that there are precursors to human morality in the animal kingdom. I will not review the evidence here, but I want to consider a few examples that might be used in support of moral nativism (for more discussion, see de Waal, 1996; Hauser, 2001).

Let's begin, as is the custom, with rats. One might assume that these lowly creatures are oblivious to each other's welfare, but there is some reason to think that is not the case. Decades ago, Church (1959) discovered that rats would stop pressing on a lever to release food if, while doing do, they saw another rat in an adjacent chamber being shocked. Similar behaviors were subsequently observed in pigeons (Watanabe & Ono, 1986) and rhesus monkeys (Masserman, et al, 1964). Rats and pigeons resume eating after a short while, but some monkeys will endure sustained starvation to avoid seeing a conspecific in agony. Of course, vicarious distress is not necessarily altruistic. It could be, as mentioned above, that animals use the distress of others as a sign for danger. If an animal is peacefully foraging for food and it hears a conspecific cry, it will probably stop foraging and seek shelter, because the cry indicates the presence of a threat. Failure to respond in this way would be profoundly maladaptive. Consequently, these experiments do not reveal much about animals' concerns for the fellows. Even the monkeys who starved themselves may have done so because they were mortally afraid of being shocked. With vicarious distress, witnessing another creature's pain is literally painful, so the experiments essentially show that monkeys (like rats) will avoid food when they fear pain. Rats just overcome this tendency more easily.

Following up on the Church rat study, Rice and Gainer (1962) wanted to see if rat engage in helping behavior. They hoisted one rat high up in the air causing it to squeal and writhe. They discovered that rats on the ground would lower the suspended rat by depressing a lever, rather than watching him suffer. This is an interesting result, because it shows that rats will work to avoid seeing other rats in pain. But this behavior may be a byproduct of the vicarious stress mechanisms. If rats suffer when they see the distress of their conspecifics, then it is unsurprising to find that they will work to help others. This tendency may be among the ingredients that evolved into genuinely prosocial tendencies, but there is no reason to attribute a moral sense to rats. We don't even need to suppose that rats have *concern* for each other. They just have vicarious distress.

I have already granted that humans experience vicarious distress, and I have intimated that it may play a limited role in the construction of moral rules (such as harm prohibitions). Vicarious distress may help us infer which actions are morally suspect. Blair (1995) has argued, not implausibly, that vicarious distress is a necessary precondition to the development of normal moral responses. But it is certainly not a sufficient condition. Vicarious distress is not itself a moral attitude, and it does not prevent us from conducting and condoning acts of incredible brutality.

Let's turn from vicarious distress to fairness. Humans have a keen sense of fairness and we resent it when we are not adequately compensated for our work. Brosnan

and de Waal (2003) have argued that essentially the same tendencies exist in capuchin monkeys. They trained monkeys to exchange disks with experimenters for food reward. Nome monkeys received cucumbers as the reward, and others received grapes—a much more desirable food. Some of the monkeys performing the task could see what another monkey was receiving. The crucial finding is that monkeys who received cucumbers were perfectly willing to perform the task when they did not see what other monkeys were getting, but they were significantly more likely to reject the food reward when they witnessed another monkey getting a grape for equal work. Brosnan and de Waal suggest that this reflects a nascent sense of fairness.

This interpretation has been subjected to convincing critiques. For example, Henrich (2004) argues that monkeys cannot be responding to inequity, because, by refusing to take the cucumber, they are actually increasing inequity, not reducing it. He cites evidence that humans will accept inequitable pay if they have no reason to think that rejecting that pay will have any impact on those who are receiving more. Wynne (2004) notes that cucumber-receiving monkeys also refuse rewards in a control condition, in which they see grapes being placed in a pile nearby rather than seeing grapes being given to another monkey. The natural interpretation is not that monkeys have a sense of equity, but rather that they will turn down mediocre rewards when something better is in view. This is simply an instance of the famous Tinklepaugh effect. Tinklepaugh (1928) showed that monkeys will turn down an otherwise desirable food reward (lettuce), when a more desirable reward has been observed (bananas). Compare a child who stops playing with a feeble toy when she spots a more exciting toy across the room. Brosnan and de Waal (2004) reply to this objection by arguing that there is a crucial difference in him their capuchins behave in the control condition with the pile of grapes and the unfairness condition where they observe another monkey receiving grapes. In the unfairness condition, the capuchins are increasingly likely to refuse cucumber compensation with each trial, whereas, in the control condition, they initially refuse cucumber compensation, but they then begin accept cucumbers again after several trials. The authors conclude that the capuchins must be morally indignant in the unfairness condition. But this trend can be explained without assuming the monkeys have a moral sense. As an inanimate object, the heap of grapes may become less interesting over time, and hence easier to ignore (monkey attention systems, like ours, inhibit return of attention to a previously attended location). Watching another monkey receive grapes is a more exciting stimulus; a moving conspecific is harder to ignore. In addition, while watching another monkey eat grapes, the capuchin with the cucumbers might become increasingly aware of the fact that she could be enjoying those grapes as well. She may not be thinking, "This is unfair! I'm getting less reward for the same work," but rather, "Yum! I could be eating grapes right now." The study does not distinguish fairness from envy, or mere desire.

The Brosnan and de Waal study has an ambitious aim. The authors attempt to show that monkeys make judgments about injustice. Perhaps one could find more plausible evidence for protomorality if one lowers the bar. Hauser et al. (2003) presents experimental evidence in support of a more modest hypothesis: monkeys reciprocate differentially. More specifically, Hauser et al. attempt to show three things: (a) monkeys will not give food to other monkeys who do not give them food; (b) monkeys will give food to other monkeys from whom they have received food; but (c) monkeys will reciprocate only if the monkeys who gave them good did not do so as a byproduct of

selfish actions. This last point is crucial. If monkeys merely gave food to every monkey that had given them food, the behavior might be explained as a conditioned response. If a monkey is, by chance, given food by another monkey, and then by chance, gives food in return, the first monkey's generous behavior will be positively reinforced, and that monkey will be likely to give food the next time around. In this way, patterns of reciprocal giving can emerge through conditioning. But, if monkeys reciprocate only with other monkeys who have given selflessly, then the conditioning story will lose plausibility. Selective reciprocation would be evidence that monkeys distinguish altruism from selfishness—a rudimentary moral judgment.

Hauser et al. (2003) tried to establish this with cotton-top tamarins. Two tamarins were placed in adjacent cages (Player 1 and Player 2), and they alternated trials in a reciprocation game. First, the authors established that tamarins would give food to those who gave them food selflessly. If Player 1 pulled a bar that gave Player 2 a piece of food but gave nothing to Player 1, then, on subsequent trials, Player 2 would reciprocate, by doing the same. If Player 1 never pulled the bar to give Player 2 food, then Player 2 would not reciprocate. This much might be explained by conditioning. In the crucial test, Hauser et al. set things up so that when Player 1 pulled the bar, she would get one piece of food and Player 2 would get 3. When Player 2's turn came up, she would have an opportunity to reciprocate, by pulling a bar that would give her nothing in reward, but it would give Player 1 a piece of food. In these trials Player 2 rarely reciprocated. Hauser et al. reason as follows. Player 2 can see that Player 1 is getting food each time that Player 1 gives food to Player 2; so Player 1 is not giving away that food selflessly; and since it is a selfish act, there is no reason for Player 2 to reciprocate. Tamarins appreciate altruism. Or do they?

I think these results can be explained without assuming that monkeys have an nascent moral sense. On the conditioning account, monkeys will give food to each other if doing so has been positively reinforced in the past. In order for positive reinforcement to take place, a monkey who gives food must receive food afterwards. But reinforcement can work only if the monkey thinks the reward is *a consequence* of her behavior. If a monkey receives a reward that would have come about no matter what, the monkey will not interpret that reward as contingent on her own behavior. This explains the experimental results. Player 2 sees that Player 1 is receiving food every time that Player 1 gives food to Player 2. So Player 2 should predict that Player 1 will pull the bar no matter what. As a result, Player 2 has no reason to think that Player 1's generosity is contingent on Player 2's response. Therefore, Player 2 will never have perceive a reward to be contingent upon her own bar pulling behavior. So she will not pull the bar when it is her turn, and no reciprocation will take place. There is no need fro proto-morality here. The psychological mechanism underlying these results are not much more sophisticated than the mechanisms that Skinner postulated in his behaviorist theory of animal learning.

I don't mean to suggest that non-human primates have no prosocial predispostions. It is well established that both monkeys and apes exchange goods, and that the amount that they give to others (or allow others to take) is dependent on the amount that they have received or are likely to receive from others. Monkey and apes reciprocate (de Waal, 1996). Are we to infer from this that they have a proto-morality?

I think that inference would be a mistake. First, notice an ambiguity in "proto-morality." The term might refer to disposition to behave in ways that we regard as

worthy of moral praise. Any creature that engages in self-sacrificing actions might be credited with having a proto-morality in this sense. Bees engage in altruistic behavior. But "proto-morality" might also mean a nascent understanding of right and wrong. On this interpretation, non-human animals can be said to have a proto-morality only if they have psychological motives or appraisals that can qualify as homologues of the evolutionary precursors to our own moral motives and appraisals. A moral motive is a desire to do something because it's the right thing to do, and a moral appraisal is the belief that something is morally right or morally wrong. I don't think there is any reason to attribute either of these to monkeys and apes when they engage in acts of reciprocation. Like the self-sacrificing bees, some of this behavior may be thoughtless and automatic, and some of it may be driven by non-moral motives. For example, there is now very good experimental evidence that primate food sharing correlates with harassment (Stevens, 2004). That suggests that many case of primate "altruism" may really reflect primate fear of intimidation. This is not to deny that some primates have nobler motives. Primates share with close companions more than strangers. But such behavior does not entail that primates make moral judgments. We give things to our friends because we like them, not because we deem that action morally praiseworthy.

If monkeys and apes were forming moral appraisals, we would expect to find two things that have not been well-demonstrated in other species. First, we would expect to find self-directed emotions of blame; apes who do not share should feel guilty about that. Second, we would expect to find third-party concern; apes would become outraged when they see two unrelated apes engage in an inequitable exchange. Apes may have both tendencies, but the evidence is scant (de Waal, 1996).

Until further evidence is in, we should resist the conclusion that apes make moral appraisal or act from moral motives. But, the nativist might object, that does not rule out the hypothesis that they make *proto*-moral appraisals. I'm not exactly sure what these would be. One possibility is that a proto-moral appraisal is an appraisal comprising an other-directed emotion of blame, with no disposition to form self-blame emotions or to blame others when they mistreat unrelated third-parties. I think it is misleading to call such appraisals proto-moral, but that's a terminological quibble. I do concede that human moral appraisals may utilize psychological mechanism that are homologous with the mechanisms that cause an ape to respond negatively when a conspecific, say, refuses to share. I would even concede that our biological predisposition to reciprocate fortifies us with expectations that form the foundation of our moral attitudes towards exchange. Our biological predispositions to reciprocate are fortified by culturally inculcated moral attitudes, that promote self-blame and third party concern. These concessions do suggest that we can learn something about human morality by studying other creatures. But I emphatically deny that the psychological mechanisms in other creatures overlap sufficiently with our own to support any kind of nativist claims about morality. Moralizing does build on innate resources that we share with apes, but those resources do not qualify as moral. Hence, ape altruism does not establish that morality is innate.

4. Where Do Morals Come From?

The foregoing survey suggests that there is no solid evidence for an innate moral faculty. The conclusion can be summarized by revisiting the models of innateness introduced at the beginning of this discussion.

Some innate traits are buggy: they manifest themselves in the same rigid way across the species. If morality were buggy, we would expect to find universal moral rules. One might also expect to find a species typical maturation pattern, with fixed developmental stages. We find neither. Moral rules show amazing variation across cultures, and developmental stages vary in sequence, time course, and end point.

Some innate traits are wrassey: they vary across the species and are sensitive to environmental inputs. Wrassey traits are not open-endedly varied, however. They have a few possible settings that get triggered under different circumstances. This suggest that we have innate moral domains comes close to being a wrassey proposal. On this view, everyone cares about suffering, hierarchy, reciprocity, and purity, but the specific content of these domains, and their importance, varies across cultural environments. Notice, however, that the variation is quite open ended. There are countless different rules pertaining to suffering, for example, and countless ways of arranging a social hierarchy. So these domains do not look like the toggle switches we find in blue-headed wrasses, or in the principles and parameters account of the language faculty. Moreover, there is reason to think these domains may be learned.

Some innate traits are like bird songs: there is an open-ended variety of ways that they might be expressed, but each may depends on a domain-specific learning mechanism. Arguments for the modularity of the moral domain and arguments from the poverty of the stimulus are both designed to demonstrate that there are domain-specific resources in the moral domain. I found both of these arguments wanting.

In sum, I think the evidence for moral nativism is incomplete, at best. We have, as yet, no strong reason to think that morality is innate. This conclusion is surprising because morality seems to crop up in every society, no matter how isolated and how advanced. Massive variation in religion, physical environment, and means of subsistence have no impact on the existence of morality, even if the content of morality varies widely. Capacities are canalized in this way are often innate. Often, but not always. Take the case of religion. Some people think there is an innate religion module in the human brain, but this is a minority opinion. The dominant view of religion is that is a byproduct of other human capacities: theory of mind systems, a thirst for explanation, a good memory for the exotic, and emotional response to intense sensory pageantry, and so on (Boyer, 2001; Whitehouse, 1999). Religion, like morality, appears everywhere, but not because it is innate. It appears everywhere because it is a nearly inevitable consequence of other capacities.

I think the same is true for morality. In criticizing arguments for moral nativism, I indicated some of the capacities that may underlie morality. Let me mention four important psychological capacities here:

(1) Nonmoral emotions. Emotional conditioning (the main method used in moral education) may allow us to construct behavioral norms from our innate stock of emotions. If caregivers punish their children for misdeeds, by physical threat or withdrawal of love, children will feel badly about doing those things in the future. Herein lie the seeds or remorse and guilt. Vicarious distress may also be important here.

If we have a nonmoral but negative emotional response to the suffering of others, moral educators can tap into this, and use it to construct harm norms.

(2) Metaemotions. In addition to our first-order emotions, we can have emotions about emotions. We can feel guilty about feeling mad, or guilty about not feeling anything at all. This is double important for the emergence of morality. First, we sometimes judge that our first-order moral emotions are inappropriate. Consider sexual norms. A person raised to oppose homosexuality may have an inculcated, negative emotional response to homosexuals, but she may not like having that response, and she may feel guilty about it. Her second-order guilt about blaming homosexuals can play a causal role in re-shaping her first-order emotions. Second, we have norms about how people should feel. A happy victimizer, who causes harm without remorse is judged morally worse than a remorseful victimizer (Arsenio and Lover, 1995). By adopting rules about what people should feel, not just how they should behave, we can have greater influence on behavior.

(3) Perspective taking (theory of mind). Nonhuman animals can be emotionally conditioned to behave in conformity with rules, but they usually do not respond negatively when third-party conspecifics violate those rules. A monkey may punish another monkey for stealing, and the punished monkey may feel bad (e.g., scared or sad or submissive) as a result, but both monkeys may be indifferent when then see another monkey stealing from a third party. Human beings tend to show third-party concern, and this may be a consequence of the fact that we are good at taking the perspective of others. When we see the victim of a transgression, we imagine being that victim, and we experience anger on her behalf.

(4) Nonmoral preferences and behavioral dispositions. In addition to our innate stock of emotions, there may be some innate social behaviors that lend themselves to moralization. Reciprocity and incest avoidance are two examples. These behaviors are not moral to begin with, I argued, because they are not innately underwritten by self-blame emotions and third party concerns. When coupled with human emotional capacities, these behavioral tendencies take on a more moralistic character, and, perhaps, theory of mind mechanism allow us to identify with unrelated victims of misdeeds, and acquire a concern for third parties.

In addition to these four psychological mechanisms, there will also be situational factors that drive the formation of moral rules. There are some social pressures that all human beings face. In living together, we need to devise rules of conduct, and we need to transmit those rules in ways that are readily internalized. Nonhuman animals are often violent, but their potential for bad behavior may be lower than ours. Because we can reason, there is a great risk that human beings will recognize countless opportunities to take advantage of our fellows. We can recognize the value of stealing, for example, and come up with successful ways to get away with it. High intelligence may be the ultimate consequence of an evolutionary arms race, and, with it, the capacity for bad behavior greatly increases. Intelligence is the greatest asset of a free rider. To mitigate this increased risk, cultures need to develop systems of punishment and inculcate prosocial values. Cultures need to make sure that people feel badly about harming members of the

in-group and taking properties from their neighbors. Without that, there is a potential collapse in social stability. This is a universal problem, and given our psychological capacities (for emotion, reciprocation, mental state attribution, etc.), there is also a universal solution. All cultures construct moralities. Elsewhere, I have described at length the ways in which specific cultural circumstances can shape specific moralities (Prinz, 2007). One can explain why, in certain circumstances, cannibalism, incest, polyandry, and raiding have had adaptive value. The moral systems we inherit from our communities often contain rules that are vestiges of problems that our ancestors faced. The rules are as varied as the problems, but the universal need to achieve social stability guarantees that *some* system of moral rules will be devised.

These suggestions are sketchy and speculative, but I don't mean to be presenting a model of moral development here. Rather, I want to suggest that there is an exciting research program waiting to be explored. Just as cognitive science has looked into the psychological mechanisms that lead to the emergence of religion, we can discover the mechanisms that make us moral. In both cases, those mechanisms may not be specific to the resulting domain. If I am right, then morality is not buggy, wrassey, or starlingy. It is more like pigeon piano (the result of general purpose conditioning mechanism) and flea toss (a new use for systems that evolved to serve other functions). Morality is a byproduct of other capacities.

Sometimes cognitive scientists use such arguments to support skeptical solutions. If religion is just an accidental byproduct of over-senstitive theory of mind mechanisms, then perhaps we should try to get rid of it. Likewise, one might argue, if morality is just a byproduct of emotional systems, then maybe we should get rid of it. This conclusion doesn't follow. As I just suggested, morality may be a solution to a social coordination problem, and, without it, we would be much worse off.

That said, the antinativist thesis does have an important ramification. The old-school moral sense theorists, like Hutcheson, often assumed there was a single human morality. There is one set of moral rules, and those are the rules we are innately designed to appreciate. Modern moral nativists are less prone to seeing morality as completely fixed, but the nativist program certainly gives rise to the impression that morality is very highly constrained. Nativists about language point out that certain grammars just don't exist, and could not. Perhaps certain moralities could not exist either, but the range of possible moralities vastly exceeds the range of possible grammars. And with that discovery, we can recognize that morality is an extremely flexible tool. If morality is something we construct, then, like other tools, it is also something we can change and improve upon. We can try to reshape moral systems to better serve our current needs, and achieve greater degrees of social cohesion. The cognitive science of morality will, I think, play an important role in learning the boundaries and optimal techniques for moral change.[1]

[1] I am indebted to Walter Sinnott-Armstrong for comments on an earlier version, to Stefan Linquist for a discussion of innateness, and to Valerie Tiberius for a commentary. All three were very helpful. I have also benefited, perhaps too late, from discussions with audience members at Dartmouth, Columbia, Rutgers, Oxford, Leeds and at the Society for Philosophy and Psychology meeting at Wake Forest University.

References

Arsenio, W. F., and Lover, A. (1995). Children's conceptions of sociomoral affect: Happy victimizers, mixed emotions and other expectancies. In M. Killen and D. Hart (Eds.) *Morality in everyday life: Developmental perspectives* (pp. 87-128). Cambridge: Cambridge University Press.

Birbaumer, N., Veit, R., Lotze, M., Erb, M., Hermann, C., Grodd, W., and Flor, H. (2005). Deficient fear conditioning in psychopathy: A functional magnetic resonance imaging study. *Archives of General Psychiatry*, 62, 799-805.

Bittles, A. H. (1990). Consanguineous marriage: Current global incidence and its relevance to demographic research. Research report no. 90-186, Population Studies Center, University of Michigan.

Blair, R. J. R. (1995). A cognitive developmental approach to morality: Investigating the psychopath. *Cognition*, 57, 1-29.

Boyer, P. (2001). *Religion explained*. New York: Basic Books.

Brosnan, S. F. and de Waal, F. B. M. (2003). Monkeys reject unequal pay. *Nature*, 425, 297-299.

Brosnan, S. F., and de Waal, F. B. M. (2004). Brosnan and de Waal reply. *Nature*, 428, 140.

Chagnon, N. A. (1968). *Yanomamö: The fierce people*. New York, NY: Holt, Rinehart and Winston.

Church, R. M. (1959). Emotional reactions of rats to the pain of others. *Journal of Comparative and Physiological Psychology*, 52, 132-134.

Cleckley, H. M. (1941). *The mask of sanity: An attempt to reinterpret the so-called psychopathic personality*. St Louis, MO: The C. V. Mosby Company.

Colby, A., Kohlberg, L., Gibbs, J., and Lieberman, M. (1983). A longitudinal study of moral judgment. *Monographs of the Society for Research in Child Development*, 48, Nos. 1-2.

Cosmides, L. and Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides, and J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 163-228). New York, NY: Oxford University Press.

Cowie, F. (1999). *What's within? Nativism reconsidered*. Oxford: Oxford University Press.

Dawson, T. L. (2002). New tools, new insights: Kohlberg's moral reasoning stages revisited. *International Journal of Behavioral Development*, 26, 154-166.

de Waal, F. B. E. (1996). *Good natured: The origins of right and wrong in humans and other animals*. Cambridge, MA: Harvard University Press.

Dwyer, S. J (1999). Moral competence. In K. Murasugi and R. Stainton (Eds.), *Philosophy and linguistics* (pp. 169-190). Boulder, CO: Westview Press.

Edgerton, R. B. (1992). *Sick societies: Challenging the myth of primitive harmony*. New York, NY: The Free Press.

Edwards, C. P. (1980). The development of moral reasoning in cross-cultural perspective. In R. H. Munroe and B. B. Whiting (Eds.), *Handbook of cross-cultural human development*, New York: Garland Press.

El-Hamzi, M., Al-Swailem, A., Warsy, A., Al-Swailem, A., and Sulaimani, R. (1995).

Consanguinity among the Saudi Arabian Population. *American Journal of Medical Genetics*, 32, 623-626.

Fedora, O., and Reddon, J. R. (1993). Psychopathic and nonpsychopathic inmates differ from normal controls in tolerance to electrical stimulation. *Journal of Clinical Psychology*, 49, 326-331.

Fiske, A. P. (1991). *Structures of social life: The four elementary forms of human relations*. New York: Free Press.

Geertz, C. J. (1973). *The interpretation of cultures*. New York: Harper.

Gilligan, C. (1982). *In a different voice*. Cambridge, MA: Harvard University Press

Greene, J. D. and Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, 6, 517-523.

Grusec, J. E. and Goodnow, J. J. (1994). Impact of parental discipline methods on the child's internalization of values: A reconceptualization of current points of view. *Developmental Psychology*, 30, 4-19.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108, 814-834.

Haidt, J. and Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133, 55-66.

Haidt, J., Koller, S., & Dias, M. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, 65, 613-628.

Harris, M. (1986). *Good to eat: Riddles of food and culture*. New York, NY: Simon and Schuster.

Hauser, M. D. (2001). *Wild minds: What animals really think*. New York, NY: Henry Holt.

Hauser, M. D. (2006). *Moral minds: How nature designed our sense of right and wrong.* New York: Ecco Press.

Hauser, M. D., Chen, M.K., Chen, F., and Chuang, E., (2003). Give unto others: Genetically unrelated cotton-top tamarin monkeys preferentially give food to those who altruistically give food back. *Proceedings of the Royal Society of London, Series B*, 270, 2363-2370.

Henrich, J. (2004). Inequity aversion in Capuchins? *Nature*, 428, 139.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., and Gintis, H. (2004). *Foundations of human sociality*. New York: Oxford University Press.

Hutcheson, F. (1725/1994) *An inquiry into the original of our ideas of beauty and virtue* In R. S. Downie (Ed.) *Philosophical writings*. London: J. M. Dent.

Itard, J. M. G. (1801/1962). *The wild boy of Aveyron*. Trans. By G. Humphrey and M. Humphrey. New York: Appleton-Century-Crofts.

Kelly, D., and Stich, S. (forthcoming). Two theories about the cognitive architecture underlying morality. In P. Carruthers, S. Laurence, and S. Stich (Eds.), *The innate mind, Vol. III: foundations and the future*. Oxford: Oxford University Press.

Kohlberg, L. (1984). *The psychology of moral development: Moral stages and the life cycle*. San Francisco, CA: Harper & Row.

Kosson, D. S., Suchy, Y., Mayer, A. R., and Libby, J. (2002). Facial affect recognition in criminal psychopaths. *Emotion*, 2, 398-411.

Krebs, D., Denton, K., Vermeulen, S.C., Carpendale, J.I., and Bush, A. (1991). The

structural flexibility of moral judgment. *Journal of Personality and Social Psychology: Personality*, 61, 1012-1023.

Lovelace, L., & Gannon, L. (1999). Psychopathy and depression: Mutually exclusive constructs? *Journal of Behavior Therapy and Experimental Psychiatry*, 30, 169-176.

Masserman, J. H., Wechkin, S., and Terris, W. (1964). "Altruistic" behavior in rhesus monkeys. *American Journal of Psychiatry*, 121, 584-585.

Mikhail, J. (2000). Rawls' linguistic analogy: A study of the 'generative grammar' model of moral theory described by John Rawls in 'A Theory of Just ice.' Doctoral Dissertation, Department of Philosophy, Cornell University.

Modell, B., and Darr, A. (2002). Genetic counselling and customary consanguineous marriage. *Nature Reviews Genetics*, 3, 225-9.

Moll, J., de Oliveira-Souza, R., Bramati, I. and Grafman, J. (2002). Functional networks in emotional moral and nonmoral social judgments. *NeuroImage,* 16, 696-703.

Mwamwenda, T. S. (1991). Graduate students' moral reasoning. *Psychological Reports*, 68, 1368-1370.

Nichols, S. (2005). Innateness and moral psychology. In P. Carruthers, S. Laurence, and S. Stich (Eds.), *The innate mind: Structure and content*. New York, NY: Oxford University Press.

Nucci, L. P. (2001). *Education in the moral domain*. Cambridge: Cambridge University Press.

Nucci, L. P. and Weber, E. (1995). Social interactions in the home and the development of young children's conceptions of the personal. *Child Development*, 66, 1438-1452.

Patrick, C. J. (1994). Emotion and psychopathy: Startling new insights. *Psychophysiology*, 31, 319-330.

Prinz, J. J. (2006). Is the mind really modular? In R. Stainton (Ed.), *Contemporary debates in cognitive science* (pp.). Oxford: Blackwell.

Prinz, J. J. (2007). *The emotional construction of morals*. Oxford: Oxford University Press.

Prinz, J. J. (forthcoming). Against moral nativism. In M. Bishop and D. Murphy (Eds.) *Stich and his critics*. Oxford: Blackwell.

Puka, B. (Ed.) (1994). *Moral development. A compendium. Vol. 4 The great justice debate: Kohlberg criticism*. New York: Garland Publishing.

Read, K. E. (1955). Morality and the concept of the person among the Gahuku-Gama. *Oceania*, 25, 233-282.

Rice, G. E., Jr., and Gainer, P. (1962). "Altruism" in the albino rat. *Journal of Comparative and Physiological Psychology*, 55, 123-125.

Rosaldo, M. Z. (1980). Knowledge and passion: Ilongot notions of self and social life. Cambridge: Cambridge University Press.

Ruse, M. (1991). The significance of evolution. In P. Singer (Ed.), *A companion to ethics* (pp. 500-510). Oxford: Blackwell.

Schmitt, E. (2005, February 4). General Is scolded for saying, "It's fun to shoot some people," *New York Times*.

Shweder, R. A., Much, N. C., Mahapatra, M., and Park, L. (1997). The "Big Three" of morality (autonomy, community, divinity), and the "Big Three" explanations of

suffering. In P. Rozin and A. Brandt (Eds.), *Morality and health*. New York, NY: Routledge.
Smetana, J. G. (1981). Preschool children's conceptions of moral and social rules. *Child Development*, 52, 1333-1336.
Smetana, J. G. (1989). Toddlers' social interactions in the context of moral and conventional transgressions in the home. *Developmental Psychology*, 25, 499-508.
Smetana, J. G. (1995). Morality in context: Abstractions, ambiguities and applications. In R. Vasta (Ed.), *Annals of child development, Vol. 10* (pp. 83-130). London: Jessica Kingsley.
Snarey, J. R. (1985). Cross-cultural universality of social-moral development: A critical review of Kohlbergian research. *Psychological Bulletin*, 97, 202-232.
Sober, E. and Wilson, D. S. (1998). *Unto others: The evolution and psychology of unselfish behavior*. Cambridge, MA: Harvard University Press.
Song, M., Smetana, J., and Kim, S. Y. (1987). Korean children's conceptions of moral and conventional transgressions. *Developmental Psychology*, 23, 577–582.
Stevens J. R. (2004). The selfish nature of generosity: Harassment and food sharing in primates. *Proceedings of the Royal Society of London, Series B*, 271, 451-456.
Stevens, D., Charman, T., & Blair, R. J. R. (2001). Recognition of emotion in facial expressions and vocal tones in children with psychopathic tendencies. The Journal of Genetic Psychology, 762(2), 201-211.
Stone, V., Cosmides, L., Tooby, J., Kroll, N. and Knight, R. (2002). Selective impairment of reasoning about social exchange in a patient with bilateral limbic system damage. *Proceedings of the National Academy of Sciences*, 99, 11531-11536.
Thornhill, N. W. (1991). An evolutionary analysis of rules regulating human inbreeding and marriage. *Behavioral and Brain Sciences*, 14, 247-293.
Tinklepaugh, O. L. (1928). The self-mutilation of a male Macacus rhesus monkey. *Journal of Mammalogy*, 9, 293-300.
Turiel, E. (1998). Moral development. In N. Eisenberg (Ed.), *Social, emotional, and personality development* (pp. 863-932). New York: John Wiley.
Turiel, E. (2002). *The culture of morality: Social development, context, and conflict*. Cambridge: Cambridge University Press.
Watanabe, S., and Ono, K. (1986). An experimental analysis of "empathic" response: Effects of pain reactions of pigeon upon other Ppgeon's operant behavior, *Behavioural Processes*, 13, 269-277.
Wheatley, T. and Haidt, J. (2005). Hypnotically induced disgust makes moral judgments more severe. *Psychological Science*, 16, 780-784.
Whitehouse, H. (1999) *Arguments and icons*. Oxford: Oxford University Press.
Wrangham, R. (2004). Killer species. *Daedalus*, 133, 25-35.
Wynne, C. D. L. (2004) Fair refusal by capuchin monkeys. *Nature*, 428, 140.