20 Innateness and Moral Psychology

Shaun Nichols

Although linguistic nativism has received the bulk of attention in contemporary innateness debates, moral nativism has perhaps an even deeper ancestry. If linguistic nativism is Cartesian, moral nativism is Platonic. Moral nativism has taken a backseat to linguistic nativism in contemporary discussions largely because Chomsky made a case for linguistic nativism characterized by unprecedented rigor. Hence it is not surprising that recent attempts to revive the thesis that we have innate moral knowledge have drawn on Chomsky's framework. I'll argue, however, that the recent attempts to use Chomsky-style arguments in support of innate moral knowledge are uniformly unconvincing.

The central argument in the Chomskian arsenal, of course, is the Poverty of the Stimulus (POS) argument. In section 1, I will set out the basic form of the POS argument and the conclusions about domain specificity and innate propositional knowledge that are supposed to follow. In section 2, I'll distinguish 3 hypotheses about innateness and morality: rule nativism, moral principle nativism, and moral judgment nativism. In sections 3-5 I'll then consider each of these hypotheses in turn. I'll argue that while there is some reason to favor rule nativism, the arguments that moral principles and moral judgment derive from innate moral knowledge don't work. The capacity for moral judgment is better explained by appeal to innate affective systems rather than innate moral knowledge. In the final section, I'll suggest that the role of such affective mechanisms in structuring the mind complicates the standard picture about poverty of the stimulus arguments and nativism. For the affective mechanisms that influence cognitive structures can make contributions that are neither domain general nor domain specific.

1 Poverty, Innateness, and Domain Specificity

Like most toweringly influential arguments in philosophy, Chomsky's POS argument is at its core quite simple. We can suppose that empiricist learning proceeds by applying domain general learning mechanisms (e.g., hypothesis testing) to environmental input. The idea behind the POS argument is that the environment doesn't contain enough information to enable an empiricist learner to acquire the linguistic competence that children exhibit (Laurence & Margolis 2001; see also Botterill & Carruthers 1999, Cowie 1999). This shows that children are not merely empiricist learners when it comes to language. This argument is only strengthened if it turns out that children acquire the capacity early in development (Samuels 2002, 238).

It's important to distinguish two inferences drawn from the POS argument, a negative and a positive conclusion (see, e.g., Laurence & Margolis 2001, 248). If the POS argument works at all, it delivers the *negative conclusion* that the acquisition of language can't be explained by the empiricist proposal. This anti-empiricist conclusion is of course of signal importance. But the anti-empiricist conclusion would not be very sticky without a positive proposal as well. One standard interpretation of Chomsky's POS arguments is that they are supposed to lead to a positive conclusion. Fodor puts it thus: "The bottom line of Poverty of Stimulus Arguments, as Chomsky uses them, is that innate domain specific information is normally recruited in first language acquisition" (Fodor 2001). Hence, the positive conclusion is that first language acquisition involves "innate, domain specific information". The connection between these elements is fairly clear. The body of information is restricted to the domain of language, so the domain is specified by the information itself. And the body of information is innate.[1]

In the recent literature in developmental psychology and evolutionary psychology, there are a number of somewhat different notions of domain specificity (see, e.g., Carruthers forthcoming, Samuels et al. 1999, Karmiloff-Smith 1992). However, since the focus here will be on POS-style arguments, our interests will be in the notion of domain specificity that plays the central role in POS arguments. As Cowie puts it in her discussion of POS arguments, nativists invoke domain-specific mechanisms to explain the "gap between the information provided by experience about some domain… and the ideas or beliefs we acquire concerning that domain" (Cowie 1999, 37; see also Laurence and Margolis 2001). Hence, for our purposes, domain-specific mechanisms will be mechanisms that are not part of the stock of empiricist mechanisms but that are devoted to special functions or special tasks. The standard examples are mechanisms that are devoted to the domains of language, mindreading, and folk physics. Domain-specific databases constitute one kind of domain-specific mechanism. In addition, some cognitive mechanisms are thought to be domain-specific *processors*. Perhaps the best known species of this genus is the Fodorean module, a

[1] There is much discussion about how to define innateness (see, e.g., Cowie 1999; Samuels 2002), but I am happy enough to rely on exemplars of innate traits (e.g., ears) and non-innate traits (e.g. scars) as a rough guide to whether a cognitive trait is innate (see Laurence & Margolis 2001, 219-20).

mechanism which processes only certain kinds of information, viz., information restricted to a particular domain. A nativist might invoke either domain-specific databases or processors (or both) to explain the acquisition of a capacity that outstrips the resources of the empiricist learner.

2 Three kinds of nativism about norms

Now that we have the general background on nativism in place, we can turn to our focus on the status of nativism in the moral domain. There are a number of psychological joints at which the normative domain can be cut. I will distinguish three kinds of nativist claims about moral capacities: rule nativism, moral principle nativism, and moral judgment nativism.

2.1 Rule Nativism

People obviously have a capacity to recognize and reason about rules, and the basic capacity for rule comprehension is a natural candidate for a nativist proposal. To frame the nativist proposal, it will be useful to draw on Kant's distinction between hypothetical and nonhypothetical imperatives. Hypothetical imperatives are rules that serve one's interests like "Put oil in your car". This imperative applies to us because we desire to prevent our engine from seizing up. If for some reason we *want* our engine to seize up (say, because we're conducting an engine test) then the imperative no longer applies. Some imperatives, however, apply to us even if they don't serve our interests. Kant's examples here were moral imperatives, like "Don't lie"; this moral imperative applies to us even when lying is obviously in our best interests. However, in a widely influential essay, Philippa Foot argues that moral imperatives aren't the only cases of nonhypothetical imperatives. Foot begins by noting that on Kant's characterization, hypothetical imperatives are "those telling a man what he ought to do because … he wants something and those telling him what he ought to do on grounds of self-interest" (Foot 1972, 306). She then proceeds to give two examples of nonmoral norms that are not hypothetical in this self-interested sense. First, Foot offers an example from etiquette, the norm that invitations addressed in the third person should be answered in the third person, and she claims that "the rule does not *fail to apply* to someone who has his own good reasons for ignoring this piece of nonsense, or who simply does not care about what, from the point of view of etiquette, he should do" (Foot 1972, 308). Even though I may have no interest in following the rule of etiquette, it still applies to me. Foot's second example invokes a club rule: "The club secretary who has told a member that he should not bring ladies into the smoking-room does not say, 'Sorry, I was mistaken' when informed that this member is resigning tomorrow and cares nothing about his reputation in the club" (Foot 1972, 308-9). Here again, it is not in the member's interests to obey the rule, but it is still the case that he is breaking the rule – he is doing something that he is not supposed to do.

There are a number of further distinctions to draw between different kinds of imperatives.[2] But for our purposes, the class of nonhypothetical imperatives is central.[3] The capacity to recognize and reason about these nonhypothetical imperatives is plausibly a fundamental capacity implicated in moral judgment, and one might well maintain that we have innate mechanisms dedicated to this basic capacity. The precise label for this view would be "nonhypothetical-imperative comprehension nativism", but I'll abbreviate this to "rule nativism".

2.2 Moral principle nativism

In addition to a capacity for rule comprehension, people exhibit knowledge of distinctively moral principles. One might claim that certain of these moral principles are innately specified. Obvious candidates here are principles that seem to be universal. For instance, some claim that in every culture there are prohibitions against rape, violence, and murder (Pinker 1994, 414; see Brown 138-9). These might be regarded as public expressions of innate moral principles. As we will see, the analogy with grammatical principles leads some theorists to propose a counterpart to Chomsky's Universal Grammar, a "Universal Moral Grammar" (Harman 1999, 225; Mikhail 2002, 1088). We can call this kind of view "moral principle nativism."

2.3 Moral judgment nativism

In the psychological literature, the capacity for moral judgment has perhaps been most directly and extensively approached empirically by exploring the basic capacity to distinguish moral violations from conventional violations (for reviews see Smetana 1993 and Tisak

[2] For instance there is the additional Kantian notion of the categorical imperative, which allegedly present an action as "objectively necessary". Etiquette norms and school rules are clearly not categorical even if they are nonhypothetical.

[3] The focus on nonhypothetical imperatives was suggested to me by recent work by Chandra Sripada and Stephen Stich.

1995). Rather than attempt to define the moral and conventional domains, the easiest way to see the import of the data on moral judgment is to consider how subjects distinguish canonical examples of moral violations (e.g., hitting, pulling hair) from canonical examples of conventional violations (e.g., talking during storytime). From a young age, children distinguish canonical moral violations from canonical conventional violations on a number of dimensions. For instance, children tend to think that moral transgressions are generally less permissible and more serious than conventional transgressions. Children are also more likely to maintain that the moral violations are "generalizably" wrong, e.g., that pulling hair is wrong in other countries too. And the explanations for why moral transgressions are wrong are given in terms of fairness and harm to victims. For example, children will say that pulling hair is wrong because it hurts the person. By contrast, the explanation for why conventional transgressions are wrong is given in terms of social acceptability – talking out of turn is wrong because it's rude or impolite, or because "you're not supposed to." Further, conventional rules, unlike moral rules, are viewed as dependent on authority. For instance, if at another school the teacher has no rule against talking during storytime, children will judge that it's not wrong to talk during storytime at that school; but even if the teacher at another school has no rule against hitting, children claim that it's still wrong to hit.

These findings on the moral/conventional distinction are neither fragile nor superficial. On the contrary, the findings are quite robust. They have been replicated numerous times using a wide variety of stimuli. Furthermore, the research apparently plumbs a fairly deep feature of moral judgment. For, as recounted above, moral violations are treated as distinctive along several quite different dimensions. Finally, this turns out to be a persistent feature of moral judgment. It's found in young and old alike. Thus, we might think of this as reflecting a kind of *core moral judgment.* Accordingly, one might maintain that some innate moral knowledge guides the child in developing such an early appreciation of the distinctive status of morality. Call this view "moral judgment nativism."[4]

3 The case for innateness of rule comprehension

[4] These three nativist proposals might be teased apart in various ways. Rule nativism does not entail moral principle nativism -- the capacity for rule comprehension need not carry with it any particular principles. Neither does rule nativism entail moral judgment nativism. For the recognition of nonhypothetical imperatives like etiquette rules does not deliver the moral/conventional distinction.

A number of recent theorists have proposed something like rule nativism (e.g., Cummins 1996; Sripada & Stich forthcoming). Recall that the focal capacity is the ability to recognize and reason over nonhypothetical rules. Is it plausible that this ability derives from empiricist learning mechanisms? I'll sketch a kind of POS argument that might support rule nativism; this argument is enhanced by evidence on young children's facility with rules.

As empiricism was described above, the empiricist learner has a set of domain general capacities (e.g., hypothesis testing) for processing input from the environment. In addition, in the present context it will be important to allow the empiricist learner general purpose means-ends reasoning. The enthusiast for rule nativism might argue as follows. It's easy to see how the empiricist learner might come to hold *hypothetical imperatives.* For the empiricist learner just determines that certain actions get better results for him than other actions. Following certain rules helps him to get what he wants. However, there is no obvious story about how the empiricist learner might come to acknowledge *nonhypothetical imperatives.* When confronted with the environmental information concerning etiquette, for example, the empiricist learner might think that it's in his best interests to reply in the third person to invitations addressed in the third person. However, it's not at all clear how empiricist learning mechanisms would lead him to acknowledge that *even if it is not in his best interests*, he should reply in the third person. People clearly have this capacity to acknowledge nonhypothetical imperatives that apply even when they run against one's desires and interests. As a result, people's capacity for this kind of rule comprehension must depend on some innate contribution beyond what empiricists allow. The mind is apparently prewired to have a cognitive slot for nonhypothetical rules.[5]

As noted in section 1, a POS argument is only strengthened if we find that the capacity in question emerges early in development. And there is indeed evidence for the early emergence of rule comprehension. By the age of four, children are adept at detecting transgressions of both familiar precautionary rules and arbitrary novel rules (Cummins 1996, Harris & Núñez 1996). This evidence shows a strikingly early capacity for rule comprehension. In particular, it shows that young children are quite capable of assimilating

[5] An obvious empiricist response is to maintain that the 'nonhypothetical imperatives' are really just heuristics that the empiricist learner recognizes as being in his interests in the long run. But this seems to distort the facts about normative judgment. When children learn norms like the rules of etiquette, they often have no idea whether following the rule will benefit them or not.

information about which sorts of actions are prohibited and then using this information appropriately to judge whether a given action is a transgression.

The foregoing scarcely provides a knockdown argument for rule nativism. One salient fact is that the child is exposed to *lots* of admonitions and instruction in the normative domain. Parents and teachers are constantly telling kids what shouldn’t be done, and perhaps the empiricist can concoct some story about how the cognitive slot for nonhypothetical imperatives emerges through general reasoning. Nonetheless, the arguments for rule nativism seem sufficiently promising to make rule nativism a contender. The case for rule nativism is also appreciably better than the other nativist arguments to be considered below.

4 The case for the innateness of moral principles

Theorists arguing for distinctively moral nativism, as opposed to the broader kind of *rule* nativism considered above, have found the analogy with linguistics irresistible. In recent work, Gilbert Harman and John Mikhail suggest that just as we have an innate set of grammatical principles guiding our language acquisition, we also have an innate set of moral principles, a “universal moral grammar” (Harman 1999; Mikhail 2002; see also Stich 1993). Harman and Mikhail advert to two key points to support the case for the existence of a Universal Moral Grammar. People seem to be committed to a set of subtle, untaught moral principles, and this might be explained by positing a Universal Moral Grammar; positing such a grammar would also explain the existence of cross-culturally universal moral principles. I’ll consider the merits of these arguments in turn.

4.1 Unlearned moral principles

According to the Chomskian POS argument, the child has knowledge of grammatical principles which could not possibly have been learned from the available evidence (using empiricist learning mechanisms); hence, these principles must be part of the innate Universal Grammar. Harman maintains that a parallel argument might be made for moral principles. Just as there are unlearned syntactic principles, Harman suggests that there are “unlearned moral principles” which are part of a universal moral grammar (Harman 1999, 224-5).

Harman draws on the large literature devoted to the “Trolley Problem” to make the case for unlearned moral principles. Philosophical research in this area resembles linguistic research insofar as the project is to consider a wide range of test cases against our intuitions and to determine a set of principles that will capture our intuitions about the cases (Harman 1999, 224). Here’s Harman’s gloss of the standard trolley case:

> You are driving a trolley and the brakes fail. Ahead five people are working on the track with their backs turned. Fortunately you can switch to a side track, if you act at once. Unfortunately there is also someone on that track with his back turned. If you switch your trolley to the side track, you will kill one person. If you do not switch your trolley, you will kill five people
>
> (Harman 1977, 57).

Most people think that it is permissible to switch to the side track, killing one person but saving five. After all, the choice is between one person dying and five persons dying. However, this is hardly the end to it. For consider the variant in which you have to throw a person onto the tracks to stop the train from hitting the five people. In this case, most people regard the action as impermissible.[6]

Since the origin of the trolley literature (Foot 1967), the Doctrine of Double Effect (DDE) has been a prevailing candidate for capturing intuitions about a range of trolley cases. According to the DDE it can be permissible to perform an act that has an unintended but foreseen side effect that one is forbidden from intending. Hence, in the initial trolley case, it is permissible to switch the trolley even though it has an effect (the killing of an innocent) which it would be impermissible to intend. It will serve us better to have a fuller characterization of the principle:

> The principle holds that under strict conditions it is permissible foreseeably to bring about an effect of a type that it is never permissible to intend. These conditions are: that the act itself…be morally good or indifferent; that the bad effect…be an unavoidable, unintended effect of the act which also achieves the good effect…; and that the good effect be sufficiently weighty to warrant causing the bad effect
>
> (Uniacke 1998, 120).

Obviously, the DDE is subtle and sophisticated. And few people are explicitly taught this doctrine. As a result, Harman argues, if the DDE is "adequate to an ordinary person's I-

[6] These sorts of cases are discussed at length in Thomson (1986). For empirical confirmation of the pattern of intuitions described above, see Greene et al. (2001) and Mikhail (forthcoming).

morality [the moral idiolect of an individual]" this would provide reason to think that the principle is part of universal moral grammar:

> An ordinary person was never taught the principle of Double-Effect…, and it is unclear how such a principle might have been acquired from the examples available to the ordinary person. This suggests that the relevant principle is built into I-morality ahead of time, in which case we should expect it to occur in all I-moralities (or be a default case, or something of the sort). In other words, the principles should be part of universal moral grammar
>
> (Harman 225).[7]

It is, as Harman notes, thoroughly implausible that people are taught the DDE. So, if this principle is adequate to people's moral views, then Harman suggests, the principle must be a built-in element of a "universal moral grammar".

The suggestion that we have a universal moral grammar is enticing, but the above argument for unlearned moral principles fails to support any such innate moral knowledge. To begin, the claim that the DDE might turn out to be "adequate to an ordinary person's I-morality" (225) is ambiguous on a crucial dimension that loomed important in philosophical discussions of linguistics. Grammatical intuitions, it was agreed by all sides, play a vital role in linguistics. However, it is important to distinguish between an *external* and an *internal* approach to linguistics (see Stich & Ravenscroft 1994). On the external approach, the linguists' job is precisely to come up with a grammar that is *externally adequate* to the linguistic intuitions. That is, the goal is to assemble a set of principles that captures most of these intuitions. It's possible that there are a number of quite different grammars that will satisfy this goal, and this approach can be entirely neutral on the psychological details about how (or whether) this grammar is internally represented (see Stich 1972). By contrast, on the internal approach to linguistics, the goal is not just to come up with a set of principles that *fit* the observed intuitions, but to divine the set of principles that are causally responsible for, *inter alia*, the production of the grammatical intuitions (e.g., Fodor 1981).

Now, as with linguistic theory, we need to distinguish between two approaches to the trolley cases. On an external approach, the goal is to produce a unified set of principles that would capture most of the trolley intuitions. On an internal approach the goal is to determine

[7] Harman credits unpublished work by John Mikhail here. Mikhail (forthcoming) makes an extensive empirical case that people's intuitions about trolley cases conform to the DDE.

the psychological elements that actually subserve the trolley intuitions. Many of the philosophers engaged in the trolley debates are clearly pursuing the externalist project of producing a set of principles that fits with the intuitions, and if it turns out that their favored set of principles is not psychologically realized in the average person, this is not a particular problem.

There is, of course, considerable disagreement in the trolley literature, and a number of philosophers deny that the DDE is externally adequate to our intuitions (e.g., Foot 1967, Thomson 1986). Nonetheless, the DDE, or something very like it, has an impressive cadre of admirers (e.g., Harman 1977, Nagel 1986; Quinn 1989), and I'll simply grant the moral principle nativist the assumption that the DDE is externally adequate to trolley intuitions. However, this does not entail that the DDE is part of an innate universal moral grammar. While the explicit goal of the external project is to develop a single theory that accommodates the trolley intuitions, it is a bold assumption that internally there is a single unified set of principles that subserves trolley intuitions.

So even if we assume that the DDE is externally adequate to a core set of trolley intuitions, we still need to determine the best internal account of those trolley intuitions. It's by no means clear that the appeal to an innate DDE principle is the best explanation of the pattern of intuitions. Here I want to sketch just one alternative internal account. The intuitions might implicate multiple cognitive mechanisms rather than a single unified set of complex principles.

It's independently plausible to think that people have both a set of nonhypothetical moral rules (like the prohibition against murder) and a separate, general capacity to reason about how to minimize bad outcomes.[8] It's natural to think of these two systems as deontological and utilitarian systems, respectively. These systems are at least partly independent. For the utilitarian system is deployed in thoroughly nonmoral domains, including the merely prudential; furthermore, the nonhypothetical rules of the deontological system are expressly *not* utilitarian – the rules apply independently of our wants and interests.

Acknowledging both a deontological and a utilitarian system also helps us to explain some apparently irresolvable tensions in commonsense moral thought. We have deeply conflicting intuitions about cases in which catastrophic utilitarian consequences – say, the destruction of a civilization – will follow unless we perform an action that is obviously

[8] Evolutionary psychologists have similarly proposed independent mechanisms for cheater detection and hazard management (e.g., Fiddick et al. 2000).

forbidden, such as murdering a child. It seems wrong to murder the child, and it also seems wrong to allow the catastrophe. (Nagel 1972). The two system approach would explain why we have this tension in our moral intuitions. The deontological system rebels at defying the moral rule; the utilitarian system balks at the catastrophic cost of sparing the child.

We can now exploit the two-system proposal to generate an internal account of the trolley intuitions: An action (or possible action) is assessed by the deontological system for whether it violates deontological prohibitions against, e.g., intending to harm innocents.[9] If the action violates such a deontological principle, then the action is judged as impermissible.[10] Even if the action violates no deontological principle it still gets assessed by the utilitarian system. If the action has not violated a deontological principle and does not run afoul of utilitarian considerations, then it is judged permissible. This two-system model might explain why trolley intuitions would fit with the DDE. According to the DDE, an action that has a foreseen effect that would be wrong to intend is permissible only if:

1. the intended action is permissible
2. the foreseen bad effect is not intended
3. there is no way to achieve the good effect without also causing the bad effect
4. the bad effect is not disproportionate to the good effect (e.g. Uniacke 1998, 120).

[9] Of course, to fit the DDE, the formulation here is important. The prohibition is against actions intended to produce bad effects rather than against actions that cause unintended but foreseeable bad effects. But this is likely a feature even of many nonmoral prohibitions. Consider the following nonmoral variant of the trolley cases. Susie and Billy's mom says "you are forbidden from breaking any of the cups." Billy subsequently sets up his train set so that the train is about to plow through five cups, then he calls for Susie as he leaves the scene. In one case, Susie can divert the train so that it will break only one cup; in another case, Susie must smash a cup in front of the train to prevent the train from breaking the five cups. It's plausible that only in the latter case would Susie be breaking the rule (even though her mother will presumably forgive the transgression).

[10] I am supposing here that the deontological system is typically privileged in an important way over the utilitarian system, but the above account doesn't explain why one system is privileged or how the systems might interact. An adequate account would obviously need to address these issues. One interesting possibility is that emotions play a role in the deontological system that they do not play for the utilitarian system (cf. Greene et al. 2001).

The deontological system will ensure that whenever conditions 1 and 2 are not met, the action will be judged as impermissible. The utilitarian system, on the other hand, will deem the action impermissible when 3 or 4 is flouted.

So the two-system model would provide an internal explanation for why the DDE is externally adequate to our intuitions. However, the two-system model does not require that the principle itself is internally represented at all. Rather, the two-system model is aimed at elucidating how we could have intuitions that can be externally captured by the DDE, even while the DDE itself does not correspond to any internal item in our moral psychology.

Of course, the two-system model I've suggested is a bare sketch. It hardly counts as a serious psychological account. The goal here has not been to deliver a definitive internal account of trolley intuitions but rather to provide a model that explains people's intuitions without invoking the universal moral grammar. The two-system model does, I suggest, provide a viable alternative to the idea that the DDE is a part of a universal moral grammar. Indeed, to the extent that it's independently plausible that the mind includes separate deontological and utilitarian evaluative systems, the two-systems account of the trolley intuitions provides a significantly better explanation than the appeal to a universal moral grammar.

4.2 Universality

Even if the argument from 'unlearned moral principles' fails, the nativist can still exploit the linguistic analogy to explain the universality of moral principles. My earlier discussion just takes for granted the rather striking fact that virtually everyone thinks that it's wrong to intentionally harm or kill innocent people. The moral nativist might complain that I've simply helped myself to a large part of what makes the nativist account attractive. For, as in the case of language, nativism provides an obvious explanation for why the range of moral systems seems to be significantly constrained. Indeed, as Mikhail notes, "even the most superficial comparison of morality and language suggests the development of moral competence is *more* constrained than the development of linguistic competence" (Mikhail 2002, 1110). For instance, it would seem that in every culture there are prohibitions against rape, violence, and murder (Brown 138-9; see Mikhail 2002, 1107-10). The hypothesis of a universal moral grammar explains the universality that I have simply assumed.

Moral nativism does offer one explanation for the universality of moral principles. However, in many instances of universally held beliefs, nativism is not the *best* explanation for universality. A standard empiricist alternative is that some beliefs are universal because the relevant information is readily available in everyone's environment. So, for instance, the universal belief that many birds fly comes from the *fact* that many birds fly and that this fact is readily accessible through our experience. This empiricist explanation might be offered to explain why we have universal moral principles. The normative information is readily available in the environment: parents systematically instruct their children that it's wrong to hurt others. However, this parry only defers the question – why is it that the norms themselves are so widely present? Why do parents in every culture have these norms? The appeal to a universal moral grammar provides an answer.

If we concede that the standard empiricist explanation of universal moral beliefs is incomplete, does the universal moral grammar proposal count as the best explanation for the universality of norms prohibiting intentional harm?[11] It's far from obvious, and I want to sketch an alternative explanation for why prohibitions against intentional harm are virtually ubiquitous.

Why does every culture have norms prohibiting hurting others? Nativism does seem a natural answer, and it is at home both with the Chomskian approach and with various evolutionary accounts of morality (e.g., Ruse 1993). Another alternative, however, is that harm norms are ubiquitous because they have an edge in cultural evolution. There are a number of cultural explanations for why harm norms arose. And it's quite possible that such norms arose for different reasons in different communities. But what seems clear is that once the harm norms did arise, they would find a powerful ally in the emotions. Accordingly, we might explain the universality of harm norms as follows:

> i. Harm norms prohibit actions to which we are predisposed to be emotionally averse
>
> ii. Norms that prohibit actions to which we are predisposed to be emotionally averse enjoyed enhanced cultural fitness over other norms.

If these two claims are right, we should expect harm norms to become widely prevalent, and we thus would have an explanation for the ubiquity of harm norms.

[11] Due to space considerations, I'm focusing on harm norms. But there are other candidates for moral universals including the widely prevalent notion of fairness. A discussion of the hypothesis that there is an innate principle of fairness exceeds the ambitions of this chapter. But it's worth noting that there are important cultural evolutionary and game theoretic explanations for the ubiquity of fairness principles that needn't appeal to an innate notion of fairness (e.g. Skyrms 1996).

Each of the two claims enjoys considerable support. Normal humans have strongly aversive emotional responses to suffering in others. These responses show quick onset, and they emerge quite early in development. Indeed, even newborn infants respond aversively to some cues of suffering (e.g., Simner 1971). As with “basic emotions” like sadness, anger, disgust, and fear, there is good reason to suppose that the emotional response to suffering in others is universal and innately specified. As a result, we should expect that in all cultures, harming people will tend to produce seriously aversive affect. Thus harmful actions themselves will be likely to arouse negative affect, all else being equal.

As for claim (ii), it’s independently plausible that emotional responses would contribute greatly to the cultural viability of norms. For instance, emotionally salient cultural items will be attention-grabbing and memorable, which are obvious boons to cultural fitness. In addition to these general theoretical virtues, (ii) also makes a clear prediction about the pattern of normative cultural evolution. *Ceteris paribus*, norms that prohibit actions that are independently likely to excite negative emotion should be more likely to survive than norms that are not connected to emotions. In some recent work on the cultural evolution of etiquette, this prediction was borne out. In Western European culture, 16th century etiquette norms that prohibited disgusting actions were much more likely to survive than other 16th century etiquette norms (Nichols 2002b).

The predicted pattern of normative evolution is also found in moral norms themselves. It has become a commonplace in discussions of moral evolution that, in the long run, moral norms exhibit a characteristic pattern of development. First, harm norms tend to evolve from being restricted to a small group of individuals to encompassing an increasingly larger group. That is, the moral community expands. Second, harm norms come to apply to a wider range of harms among those who are already part of the moral community – i.e., there is less tolerance of pain and suffering of others. The trends are bumpy and irregular, but this kind of characteristic normative evolution is affirmed by a fairly wide range of contemporary moral philosophers (e.g., Brink 1989, Nagel 1986, Railton 1986, Smith 1994). Since we are disposed to respond aversively to even low level signs of distress, the trend in moral evolution further confirms the prediction that norms will have enhanced cultural fitness when they prohibit actions which we’re predisposed to find emotionally aversive (see Nichols 2004).

Thus, one doesn’t need to appeal to innate moral principles to explain the ubiquity of harm norms. A cultural evolution account that appeals to the role of emotions can provide an explanation that is at least as promising as the moral nativist explanation. Indeed, given that the affect-based cultural evolution story is independently motivated there is reason to think it's a *better* explanation than the nativist explanation. Of course, on this account, we still explain the ubiquity of harm norms as a function of innate biases, but the biases are innate affective systems rather than innate moral principles.

5 The case for the innateness of moral judgment

Finally, let's turn to the child's capacity for core moral judgment. As we saw in section 2, from a young age children treat moral transgressions as distinctive on a number of dimensions – seriousness, authority contingence, generalizability, and justification-type. The early emergence and the multidimensionality of this capacity makes it an extremely attractive candidate for a nativist explanation. And, indeed, recently Susan Dwyer has taken up this charge. Dwyer characterizes the child's competence with the moral/conventional distinction much as we saw above, and she goes on to develop a kind of POS argument for moral nativism (Dwyer 1999 171-7). According to Dwyer, "the fundamental mistake" of empiricist accounts like social learning theory is "the assumption that all the information the child needs to achieve moral maturity is available in her environment" (172). More fully, she writes:

> Absent a detailed account of how children extrapolate distinctly moral rules from the barrage of parental imperatives and evaluations, the appeal to explicit moral instruction will not provide anything like a satisfactory explanation of the emergence of mature moral competence. What we have here is a set of complex, articulated abilities that (i) emerge over time in an environment that is impoverished with respect to the content and scope of their mature manifestations, and (ii) appear to develop naturally across the species
>
> (173).

Thus Dwyer draws the negative, anti-empiricist conclusion from her POS argument. According to Dwyer, just as empiricist accounts can't explain the child's linguistic competence, empiricist accounts can't explain the child's *moral* competence (as revealed by their grasp of the moral/conventional distinction). Dwyer also goes on to propose an answer similar to the positive conclusion set out above (section 1) for language. She suggests that "we all come into the world equipped with a store of innate moral knowledge which, together with our experience, determines our mature moral competence" (176-7). Given the universality of the moral/conventional distinction, she speculates that children are "in possession of some knowledge that primes them for recognizing two normative social domains" (177). So Dwyer draws both a negative and a positive conclusion from her POS argument. The negative conclusion is that the child's moral competence exceeds what an empiricist learner would be able to achieve given the information available in the environment. The positive conclusion is that moral competence depends on innate domain-specific information, viz., knowledge of the moral domain.

Let's allow Dwyer the negative conclusion that there isn't enough information in the environment to explain the child's capacity for moral judgment. To assess Dwyer's positive conclusion we need to consider whether there is an alternative to innate moral knowledge that provides a better explanation of the capacity for moral judgment. Recent work suggests that a better explanation of this capacity adverts to affective response. In a series of important studies, James Blair found that psychopaths and children with psychopathic tendencies perform abnormally on the moral/conventional task. For instance, psychopaths tend to give social-conventional explanations for why moral transgressions are wrong (Blair 1995). And children with psychopathic tendencies are more likely than other children with behavioral problems to judge moral transgressions as authority contingent; for example, they are more likely to say that hitting others would be okay if the teacher said it was okay (Blair 1997). Blair also found that psychopaths tend to have diminished response to distress cues in others. Over a series of studies, Blair and colleagues found that normal children, autistic children and non-psychopathic criminals all show considerably heightened physiological response both to threatening stimuli and to cues that another is in distress; psychopaths, on the other hand, show considerably heightened physiological response to threatening stimuli, but show abnormally low responsiveness to distress cues (Blair et al. 1997; Blair 1999). The fact that the population that shows a deficit in moral judgment also shows a distinctive affective deficit suggests that the moral deficit might derive from the affective deficit.

Blair's explanation of the psychopath's deficit in moral judgment appeals to what he calls a "Violence Inhibition Mechanism" or VIM (Blair 1995). The idea derives from Lorenz' (1966) proposal that social animals have evolved mechanisms to inhibit intra-species aggression. When a conspecific displays submission cues, the attacker stops. Blair suggests that there's something analogous in our cognitive systems, the VIM, and that this mechanism underlies both our response to distress cues and our capacity to distinguish moral from conventional violations. This mechanism is damaged in psychopathy, according to Blair, and this explains the psychopath’s failure on the moral/conventional task. In normals, the VIM produces negative affect which generates moral judgment.

I think that there are a number of problems with Blair’s VIM account of moral judgment (Nichols 2002a). On the model that I prefer, the capacity for drawing the moral/conventional distinction depends on two quite different mechanisms. First, there is a body of information, a normative “theory” that specifies a set of harm-based normative violations. The child’s knowledge of these rules presumably depends on the general capacity for rule comprehension (see section 3). Secondly, Blair’s data suggest that affect also plays a role in mediating performance on the moral/conventional task. In the normal population, the affective response to suffering in others bestows the harm norms with a distinctive, nonconventional status. Since psychopaths have a deficiency in their affective response to harm in others, this plausibly explains why they show a diminished tendency to treat harm norms as distinctive.

The proposal that emotions play a crucial role in generating nonconventional judgment gains further support from recent work on judgments about disgusting transgressions (e.g., spitting into a glass of water before drinking from it). In recent experiments, disgusting transgressions were treated as nonconventional along the same dimensions as moral transgressions. Disgusting transgressions are regarded by children as generalizably wrong (Nichols & Folds-Bennett 2003). Adults regard disgusting transgressions as less authority contingent and more serious than conventional transgressions. Furthermore, low disgust-sensitivity subjects are more likely than high disgust-sensitivity subjects to judge a disgusting action as authority contingent (Nichols 2002a).

Although there are differences between Blair’s proposal and the one just sketched, if either of these accounts is right, then the capacity for core moral judgment can be explained without appeal to innate moral knowledge. Of course, there is still a crucial innate contribution to distinctively moral judgment, but the contribution comes from innate affective systems rather than innate propositional knowledge.

After setting out her case for moral judgment nativism, Dwyer actually considers the possibility that emotions play a crucial role in the acquisition of moral judgment:

> the moral environment might be richer that I supposed earlier. Indeed, it is quite plausible that affective cues help children distinguish between moral transgressions and conventional transgressions. But it is hard to see how the deployment of

> emotional capacities could facilitate children’s grasp of the distinction between rule-governed behavior and accidentally-regular behavior
>
> (Dwyer 1999, 182).

Of course, I think that Dwyer is right to acknowledge that emotions might play a critical role in the development of moral judgment. However, Dwyer’s initial concession here that “the moral environment might be richer” than she had supposed looks to abandon her POS argument altogether. For it threatens to give up entirely even on the negative conclusion of the POS argument against empiricist accounts of core moral judgment. I think that this concession is too early. Even if the moral/conventional distinction doesn’t derive from innate moral knowledge, there might still be an important sense in which the tendency to treat the moral domain as distinctive is ‘unlearned’. In the affect-based accounts sketched above, the contribution of affect is in the mind of the judger rather than in the cues in the environment, and on both accounts affect influences the emergence of moral competence in a way that doesn’t conform to empiricist learning processes.

Thus, one can perfectly well accept the negative conclusion of Dwyer’s POS argument while rejecting her positive proposal that we have innate moral knowledge. The emotion-based proposals above do just that. However, Dwyer maintains that while emotion-based accounts might explain the child’s capacity to distinguish moral from conventional violations, emotion-based accounts will not explain the child’s appreciation of rule-governed behavior. This claim seems plausible, and it might bolster the kind of rule nativism discussed in section 3. However, it’s worth emphasizing that this is a serious retreat for the moral nativist. If rule nativism remains the only stronghold, then there is no longer a case for innate moral knowledge or even for innate capacities that are distinctively moral. For, as we saw above, rule nativism does not entail moral judgment nativism; the capacity for rule comprehension is by no means a distinctively moral capacity.

6 Affective constraints on cognition

Over the last several sections I’ve maintained that none of the arguments for innate moral knowledge succeeds. It’s plausible, however, that innate affective mechanisms shape our moral capacities. In this final section, I will consider how this proposal reflects on broader issues about nativism. Clearly the influences of affective responses to suffering constitute innate biases that fall on the nature side of the nature/nurture divide. However, I’ll suggest that the role of affective mechanisms in structuring the mind complicates the standard picture about poverty of the stimulus arguments and nativism.

We've assumed that some of the emotions that influence moral judgment are innately specified. As noted earlier, affective responses to suffering emerge very early and would seem to be culturally universal. It is also plausible that these emotion systems were designed by evolution. Presumably having these emotional reactions generated motivation that enhanced biological fitness in some way. It is currently unclear exactly why these responses were fitness enhancing. Nonetheless, given that these innate emotion systems are tied so closely to behavioral response, it is prima facie plausible to take them to be adaptations to some problem in the ancestral environment.

Now let's return to the capacity for moral judgment. The evidence on the moral/conventional distinction suggests that the moral realm is organized into a domain that is quite distinct from the conventional domain. Transgressions apparently get sorted into cognitive domains of moral and conventional. Hence, moral judgment has the marks of domain specificity. However, the claim in the preceding section was that these domains are generated by emotion systems – in particular, affective systems that respond to suffering in others. If that's right, emotions can have a cascading influence on information-bases, imposing important cognitive structure onto domains of knowledge. Of course an emotion-based explanation for the acquisition of a cognitive capacity can displace the appeal to innate propositional knowledge. Indeed, that was the thrust of the argument in section 5. But the role of affective mechanisms in structuring the mind has more interesting implications about domain specificity.

As noted above, it's plausible that the emotion systems that react to suffering in others evolved to address some problem in our ancestral environment. However, these emotion systems constrain cognitive structures in ways that are not domain specific. Let's suppose, in line with the proposal in section 5, that the affective response to suffering in others does marshal a division between conventional and moral transgressions. There's no reason to think that this emotion system affects *only* this set of cognitive states. That is, the emotions that influence the character of moral judgment are probably not specific to the domain of moral judgment; these emotions might influence the acquisition of other areas of knowledge. For instance, our responses to suffering in others might also play an important role in the way we think about natural disasters that cause immense human suffering.

Although the effects of these emotions on cognition are probably not domain specific, neither are they perfectly domain general. There are lots of knowledge structures that will be utterly unaffected by the emotional response to suffering in others. The class of conventional transgressions provides one obvious candidate. But these emotions don't affect our cognitions about mathematics, about music, or about growing vegetables either.

The case of moral judgment thus suggests that innate affective elements of the mind can shape cognitive structures in ways that do not fit the traditional distinction between domain general and domain specific. For the influence of emotion systems might be neither domain specific nor domain general but rather domain *diverse*. The emotional responses to suffering affect the development and character of certain cognitive structures, like the rules against intentional harm, but have no intercourse with other cognitive structures, like folk astronomy.

The existence of innate domain-diverse factors alters the landscape of nativist arguments. For a POS argument might succeed in showing that a given capacity can't have been reached by general purpose learning mechanisms, but it won't follow immediately that the capacity depends on a mechanism that is devoted to the domain of that capacity. The acquisition might depend rather on a *domain diverse* mechanism like an emotion system.

There is one further implication of the account of moral judgment that I'd like to spin out. On the proposal set out in section 5, both the capacity for rule comprehension and the emotional response to suffering are implicated in core moral judgment. I've assumed that affective mechanisms that respond to suffering in others has an innate basis and that it is the product of natural selection. I have also been allowing throughout that we have an innate capacity for rule comprehension. It's quite possible that this capacity for rule comprehension is also an adaptation. Indeed, there is a range of intriguing adaptationist proposals about the capacity for rule comprehension (see e.g., Cummins 1996; Sripada & Stich forthcoming). Thus, both of the mechanisms that I've suggested contribute to moral judgment might well be adaptations. However, it is distinctly less plausible that the capacity for core moral judgment itself is an adaptation. It's more likely that core moral judgment emerges as a kind of byproduct of (*inter alia*) the innate affective and innate rule comprehension mechanisms.[12] That is, if the emotion system and the rule system are innate adaptations, core moral judgment is plausibly a kind of cognitive spandrel. It isn't an adaptation, but it is a natural byproduct of psychological mechanisms that are adaptations.

[12] This view is bolstered by the findings on disgusting transgressions (Nichols 2002a). For again, distinctively nonconventional judgments apparently emerge as a byproduct of rules and emotions.

7 Conclusion

The linguistic analogy provides an attractive basis for advancing the idea that we come with innate moral knowledge. However, none of the recent arguments for innate moral knowledge is at all convincing. There is reason to think that moral psychology is profoundly shaped by innate biases. But the innate biases plausibly come in the form of affective mechanisms rather than propositional information. The human mind comes loaded with a set of affective systems which seem to shape cognitive structures in ways that are neither domain general nor specific to a particular domain. So if we are to understand the innate factors that influence the acquisition of knowledge structures and other cognitive capacities, we must attend to the distinctive role of our innate affective endowment.

Acknowledgements

I'd like to think Peter Carruthers and Philip Robbins for very helpful comments on an earlier draft of this paper.

References

Blair, R. (1993). *The Development of Morality*. Unpublished Ph.D. thesis, University of London.

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

Blair, R. (1997). Moral reasoning and the child with psychopathic tendencies. *Personality and Individual Differences*, 26.

Blair, R. (1999). Psychophysiological responsiveness to the distress of others in children with autism. *Personality & Individual Differences*, 26.

Blair, R., L. Jones, F. Clark, M. Smith and L. Jones (1997). The psychopathic individual: A lack of responsiveness to distress cues? *Psychophysiology*, 34.

Botterill, G. and P. Carruthers (1999). *The Philosophy of Psychology*. Cambridge University Press.

Brink, D. (1989). *Moral Realism and the Foundation of Ethics*. Cambridge University Press.

Brown, D. (1991). *Human Universals*. Temple University Press.

Carruthers, P. (forthcoming). Practical reasoning in a modular mind.

Cowie, F. (1999). *What's Within?* New York: Oxford University Press.

Cummins, D. (1996). Evidence of deontic reasoning in 3- and 4- year old children. *Memory and Cognition,* 24.

Dwyer, S. (1999). Moral competence. In K. Murasugi and R. Stainton (Eds.), *Philosophy and Linguistics*. Westview Press.

Fiddick, L., Cosmides, L. and Tooby, J. (2000). No interpretation without representation: The role of domain-specific representations and inferences in the Wason selection task. *Cognition,* 77.

Fodor, J. (1981). Introduction: Some notes on what linguistics is about. In N. Block (Ed.), *Readings in the Philosophy of Psychology,* vol. 2. Harvard University Press.

Fodor, J. (1983). *Modularity of Mind*. MIT Press.

Fodor, J. (2001). Doing without *What's Within. Mind,* 110.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review,* 5. Reprinted in *Virtues and Vices,* Oxford University Press. All page references to the reprinted version.

Foot, P. (1972). Morality as a system of hypothetical imperatives, *The Philosophical Review,* 81, 305-316.

Greene, J., Sommerville, R., Nystrom, L., Darley, J., & Cohen, J. (2001). An fMRI investigation of emotional engagement in moral judgment, *Science,* 293, 2105-08.

Harman, G. (1999). Moral philosophy and linguistics. In K. Brinkmann (Ed.), *Proceedings of the 20th World Congress of Philosophy: Volume 1: Ethics*. Philosophy Documentation Center, 107-115. Reprinted in his *Explaining Value,* Oxford University Press. All page references to the reprinted version.

Harris, P. and Núñez, M. (1996). Understanding of permission rules by preschool children. *Child Development,* 67.

Karmiloff-Smith, A. (1992). *Beyond Modularity*. Cambridge, MA: MIT Press.

Laurence, S. and Margolis, E. (2001). The poverty of the stimulus argument. *British Journal for the Philosophy of Science,* 52.

Lorenz, K. (1966). *On Aggression,* New York: Harcourt, Brace, Jovanovich.

McIntyre, A. (2001). Doing away with double effect. *Ethics,* 111.

Mikhail, J. (2002). Law, science, and morality: A review of Richard Posner's *The Problematics of Moral and Legal Theory*. *Stanford Law Review,* 54.

Mikhail, J. (forthcoming). Aspects of the theory of moral cognition.

Nagel, T. (1972). War and massacre. *Philosophy and Public Affairs,* 1.

Nagel, T. (1986). *The View from Nowhere*. Oxford University Press.

Nichols, S. (2002a). Norms with feeling: Towards a psychological account of moral judgment. *Cognition,* 84.

Nichols, S. (2002b). On the genealogy of norms: A case for the role of emotion in cultural evolution. *Philosophy of Science*, 69.

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

Nichols, S. and Folds-Bennett, T. (2003). Are children moral objectivists? Children's judgments about moral and response-dependent properties. *Cognition*.

Núñez, M. and P. Harris (1998). Psychological and deontic concepts: Separate domains or intimate connection? *Mind and Language*, 13.

Quinn, W. (1989). Actions, intentions, and consequences: The doctrine of double effect, *Philosophy and Public Affairs*, 18.

Railton, P. (1986). Moral realism. *Philosophical Review* 95.

Ruse, M. (1993). The significance of evolution. In P. Singer (Ed.) *A Companion to Ethics*. Blackwell Publishers.

Samuels, R. (2002). Nativism in cognitive science, *Mind & Language*, 17

Samuels, R., Stich, S., Tremoulet, P. (1999). Rethinking rationality. In E. LePore and Z. Pylyshyn (Eds.), *What Is Cognitive Science?* Blackwell Publishers.

Simner, M. (1971). Newborn's response to the cry of another infant, *Developmental Psychology 5*.

Skyrms, B. (1996). *The Evolution of the Social Contract*. Cambridge University Press.

Smetana, J. (1993). Understanding of social rules, In M. Bennett (Ed.), *The Development of Social Cognition : The Child as Psychologist.* Guilford Press.

Smith, M. (1994). *The Moral Problem.* Blackwell.

Stich, S. (1972). Grammar, psychology, and indeterminacy, *Journal of Philosophy*, 69.

Stich, S. (1993). Moral philosophy and mental representation. In M. Hechter, L. Nadel, and R. E. Michod (Eds.), *The Origin of Values*. Aldine de Gruyter.

Stich, S. and Ravenscroft, I. (1994). What *is* folk psychology? *Cognition,* 50.

Thomson, J. (1986). *Rights, Restitution, and Risk*. Harvard University Press.

Tisak, M. (1995). Domains of social reasoning and beyond, In R. Vasta (Ed.), *Annals of Child Development*, Vol. 11. Jessica Kingsley.

Uniacke, S. (1998). The principle of double effect. In E. Craig (ed.) *Routledge Encyclopedia of Philosophy*, Vol. 3.