Forthcoming in W. Sinnott-Armstrong (ed.) *The Psychology and Biology of Morality*. Cambridge, MA: MIT Press.

Sentimentalism Naturalized[1]

Shaun Nichols

Sentimentalism, the idea that the emotions or sentiments are crucial to moral judgment, has a long and distinguished history. Throughout this history, sentimentalists have often viewed themselves as offering a more naturalistically respectable account of moral judgment. In this paper, I'll argue that they have not been naturalistic enough. The early, simple versions of sentimentalism met with decisive objections. The contemporary sentimentalist accounts successfully dodge these objections, but only by promoting an account of moral judgment that is far too complex to be a plausible account of moral judgment on the ground. I argue that recent evidence on moral judgment indicates that emotional responses do indeed play a key role in everyday moral judgment. However, the emotions themselves are only one part of moral judgment; internally represented rules make an independent contribution to moral judgment. This account of moral judgment is grounded in the empirical evidence, but it can also handle a cluster of desiderata that concern philosophical sentimentalists. If emotions and rules do make independent contributions to moral judgment, this raises a puzzle. For our rules tend to be well coordinated with our emotions. In the final section, I'll argue that this coordination can be partly explained by appealing to the role of cultural evolution in the history of norms.

1. Sentimentalist metaethics

The history of sentimentalist metaethics is a history of increasing sophistication. On perhaps the most prominent contemporary account, Allan Gibbard's, to judge an act morally wrong is to "accept norms that prescribe, for such a situation, guilt on the part of the agent and resentment on the part of others" (1990, 47). In section 2, I'll argue that the sophistication of the philosophical accounts is their undoing. But it's worth reviewing a bit of the history to see how we ended up with such a dazzlingly complex theory of moral judgment. Along the way, we'll accumulate several desiderata for an adequate sentimentalist metaethics.

The early, relatively simple sentimentalist accounts were met with crushing counterexamples. On one prominent version of the history (e.g., Stevenson 1937), Hobbes promoted a first-person subjectivism, according to which "X is bad" just means "I disapprove of X." This runs up against the familiar problem of disagreement – when one person says X is bad and another says X is not bad, according to first-person subjectivism there is typically no conflict since they are both reporting their own psychological states. Yet it's clear that typically people *are* disagreeing when one claims that X is morally bad and the other says that it isn't. Hume is sometimes viewed as offering a community-based subjectivism according to which "X is bad" just means

[1] This paper is a *précis* of the case for a naturalistic sentimentalism presented in Nichols (2004). I'm grateful to Walter Sinnott-Armstrong for very helpful suggestions on a previous draft.

"Most people in my community disapprove of X." This allows for the possibility of *some* disagreement, since we can disagree about which views prevail in our community. But, as Stevenson points out, this account doesn't allow for disagreement between communities. And it seems implausible that the very meaning of the term 'bad' should exclude the possibility of intercommunity moral disagreement (Stevenson 1937). Thus, an adequate sentimentalist account must be able to accommodate the possibility of moral disagreement.

"Emotivism" emerged as the prevailing view that offered a solution to this problem. According to emotivism, in giving voice to a moral commitment, one is not merely reporting one's feelings, but *expressing* them. So, when we say that it's wrong to steal, what we are really saying is something like "I disapprove of stealing; do so as well" (see e.g., Stevenson 1944). This account can more easily accommodate disagreement – if you and I express different attitudes about an action, we are promoting conflicting agendas.

Although emotivism was widely viewed as a major improvement on subjectivism, emotivism was beset by problems too. One problem that arose early in the discourse focused on the fact that emotivists maintained that a person must actually have the emotion that he is expressing when he utters a moral condemnation. However, as Darwall, Gibbard and Railton put the problem, "it seems… that a person can judge something wrong even if he has lost all disposition to feelings about it" (Darwall et al. 1992, 17-8). Again, sentimentalists took this problem to provide a constraint on future theorizing – an adequate sentimentalist account must allow for the possibility that a person can still judge an action wrong even if he has lost all the relevant feelings about the action.

The final problem posed against the early sentimentalist theories concerns the role of reasoning in moral judgment (e.g., Toulmin 1950, Brandt 1950, Falk 1953, Baier 1958, Geach 1965). Moral reasoning seems to play an important part in our moral lives, and if moral judgment is simply reporting or expressing one's feeling, it's unclear how the reasoning could proceed as it does. Even simple examples of moral reasoning served to make the point. For instance, Toulmin offers the following bit of ordinary moral reasoning from principles:

> [S]uppose that I say, "I feel that I ought to take this book and give it back to Jones"…. You may ask me, "But ought you really to do so?"… and it is up to me to produce my "reasons"…. I may reply…"I ought to, because I promised to let him have it back". And if you continue to ask, "But why ought you really?", I can answer… "Because anyone ought to do whatever he promises anyone else that he will do" or "Because it was a promise" (Toulmin 1950, 146).

If moral judgments merely express feelings, it is hard to see how to explain these apparently rational transitions from general principles to specific judgments. Geach (1965) presents a more direct attack on emotivism. Emotivists will have difficulty explaining the fact that we accept conditionals with embedded moral statements. People can accept the conditional, "if spanking your own children is wrong, then spanking other people's children is wrong" without ever feeling or reporting any disapproval for spanking one's own children. This means, according to Geach, that emotivists cannot accommodate simple instances of moral reasoning like the following:

If doing a thing is bad, getting your little brother to do it is bad.

Tormenting the cat is bad.

Ergo, getting your little brother to torment the cat is bad (Geach 1965, p. 463).

An adequate sentimentalist account must be able to accommodate the manifest role of moral reasoning.

We now have quite a diverse list of desiderata. An adequate sentimentalist account needs to accommodate the following:

i. *sentimentalism*: emotion plays a crucial role in moral judgment

ii. *moral disagreement*: individuals and communities sometimes have moral disagreements

iii. *absent feeling*: a person can judge something wrong even if he has lost all feelings about it

iv. *moral reasoning*: reasoning plays a crucial role in moral judgment.

Given the disparate nature of these constraints, a sentimentalist theory that manages to meet all the constraints would be an impressive achievement indeed.

There is a basic move that manages to solve all these problems at once. Rather than identify moral judgment with the expression or reportage of emotions, contemporary "neosentimentalist" accounts identify moral judgment with the judgment that it is normatively appropriate to feel a certain emotion in response to the action (D'Arms & Jacobson 2000, 729; see also Blackburn 1998, Gibbard 1990, Wiggins 1991). Although contemporary sentimentalists widely agree on this move, few sentimentalists provide an account that is sufficiently clear and detailed to permit thorough evaluation. In particular, few sentimentalists provide a detailed account of *which* emotion is at the heart of moral judgment. Gibbard (1990) is the most obvious and visible exception. As noted at the beginning of the section, Gibbard maintains that to judge an action morally wrong is to judge that it would be warranted to feel guilty for performing the action. He writes, "what a person does is *morally wrong* if and only if it is rational for him to feel guilty for doing it, and for others to resent him for doing it" (Gibbard 1990, 42). The subsequent discussion will focus on Gibbard, since his theory is rich and detailed enough to evaluate systematically.

The striking feature about neosentimentalism is that it satisfies *all* of the desiderata. Sentiments are integral to moral judgment, indeed emotions are part of the *meaning* of moral judgments. However, even if one has lost any disposition to feel guilty about a certain action, one can still think that feeling guilt is *warranted.* Thus the problem of absent feeling is addressed as well. Furthermore, the problem of moral disagreement is met handily – moral disagreement is really disagreement about whether it is appropriate to feel guilt for doing a certain action. Obviously that kind of disagreement is possible, indeed actual. Finally, the account can accommodate moral reasoning. When we argue about moral matters, we are arguing about the appropriateness of feeling guilt in response to doing certain actions.

This brief review of 20th century metaethics is intended both to show up the relevant constraints on an adequate account and to illustrate why the history of metaethics led us to such a complex account of moral judgment. The simpler stories ran into major difficulties. The neosentimentalist approach provides ingenious solutions to the diverse array of problems and constraints that emerged over the century. Indeed, what I'll argue is that the problem with the most prominent version of neosentimentalism is that it is *too* ingenious to be a plausible account of normal moral judgment.

2. Core moral judgment and the dissociation problem

At least since Hume, sentimentalists have often self-identified as naturalists. Sentimental accounts are supposed to give a more accurate rendering of moral judgment on the ground, as opposed to the disconnected, emaciated characterization of moral judgment promoted by some in the rationalist tradition (e.g., Cudworth, Locke; see Gill forthcoming). Many contemporary sentimentalists continue to embrace naturalistic strictures. Gibbard again provides a prominent example: “The ways we see norms should cohere with our best naturalistic accounts of normative life” (Gibbard 1990, 8).

Although sentimentalists often side with naturalism, it has been notoriously difficult to evaluate sentimentalism empirically, and neosentimentalists have rarely suggested experimental evidence that might confirm or undermine their theory. There is, however, one crucial place where the theory seems to have an empirical commitment. If moral judgments are judgments of the normative appropriateness of certain emotions, then the capacity for moral judgment should not be dissociable from the capacity to make judgments about the normative appropriateness of those emotions. More specifically, if moral judgments are judgments of the appropriateness of guilt, then *an individual cannot have the capacity to make moral judgments unless she also has the capacity to make judgments about the appropriateness of guilt.*

In due course, I’ll argue that there *are* such dissociations between the capacity for moral judgment and the capacity for normative assessment of the appropriateness of guilt. This, I will argue, presents a serious problem for naturalistic neosentimentalism. But first, I need to say a bit about the empirical investigation of the capacity for moral judgment.

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; Tisak, 1995). Turiel explicitly draws on the writings of several philosophers, including Searle, Brandt and Rawls to draw the moral/conventional distinction (Turiel 1983). But the attempt to draw a categorical distinction between morality and convention has been notoriously controversial. We needn’t take sides in the controversy, for we can see the import of the evidence just by considering how people distinguish between canonical examples of moral violations and canonical examples of conventional violations. Canonical moral violations include pulling hair, pushing, and hitting. Canonical examples of conventional violations include violations of school rules (e.g., talking out of turn) and violations of etiquette (e.g., drinking soup out of a bowl). 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, for example, 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. 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.*[2]

Children begin to display a capacity for core moral judgment surprisingly early. Smetana & Braeges (1990) found that at 2 years and 10 months, children already tended to think that moral violations (but not conventional violations) generalized across contexts when asked, "At another school, is it OK (or not OK) to X?" (p. 336). Shortly after the 3rd birthday, children recognize that conventional violations but not moral violations are contingent on authority (Smetana & Braeges 1990; Blair 1993). Thus, the evidence suggests that from a very young age, children can make these distinctions in controlled experimental settings. In addition, studies of children in their normal interactions suggest that from a young age, they respond differentially to moral violations and social-conventional violations (e.g., Dunn & Munn 1987; Smetana 1989).

Although children have a strikingly early grasp of core moral judgment, their understanding of guilt seems to emerge significantly later. According to developmental psychologists, children don't understand complex emotions like guilt, pride and shame until around age 7 (Harris 1989; 1993; Harris et al. 1987; Nunner-Winkler & Sodian 1988; see also Thompson & Hoffman 1980). Gertrude Nunner-Winkler and Beate Sodian asked children to predict how someone would feel after intentionally pushing another child off of a swing. Children under the age of 6 tended to say that the pusher would feel happy. Children over the age of 6 on the other hand, tended to say that the pusher would have some negative feelings. In another study, the experimenters showed children images of two individuals, each of whom had committed a moral violation. One of the children had a happy expression and the other had a sad expression. The subjects were asked to rate how "bad" the children were. While most 4-year old children judged the happy and sad transgressors as equally bad, "the majority of 6-year-olds and almost all 8-year-olds judged the person who displayed joy to be worse than the one who displayed remorse" (1329). So, between the ages of 4 and 8, children are gradually

[2] Most of the research on the moral/conventional distinction has focused on moral violations that involve harming others. It's clear, however, that harm-centered violations do not exhaust the moral domain. To take one obvious example, adults in our society make moral judgments about distributive justice that have little direct bearing on harm. Nonetheless, it's plausible that judgments about harm-based violations constitute an important core of moral judgment. For the appreciation of harm-based violations shows up early ontogenetically (as we will see below), and it seems to be cross-culturally universal. The capacity to recognize that harm-based violations have a special status (as compared to conventional violations) is a crucial part of this core moral judgment.

developing the idea that moral transgressions are and *should be* accompanied by some negative affect. But the findings make it seem quite unlikely that 3 and 4 year old children are capable of making a normative evaluation of whether guilt is an appropriate response to a situation.

Thus, it seems like young children have the capacity for core moral judgment while lacking the capacity to judge when it is appropriate to feel guilt. This dissociation seriously threatens the neosentimentalist view that for S to think that X is morally wrong is for S to think that it would be appropriate to feel guilty for having done X. For young children apparently make moral judgments but lack the capacity to judge whether guilt is normatively appropriate for a situation. In this light, the developmental sequence that neosentimentalism suggests begins to look implausibly demanding. To make moral judgments, one must be able to

i. Attribute guilt
ii. Evaluate the normative appropriateness of emotions
iii. Combine these two capacities to judge whether guilt is a normatively appropriate response to a situation.

This seems seriously overintellectualized as an account of children's moral judgments. Perhaps older children and adults do come to see that the actions that they judge as morally wrong are those for which guilt is appropriate, but the dissociation argument suggests that this is likely a peripheral feature, not a necessary component of moral judgment.

Of course, a natural response to this is to maintain that children don't understand morality after all. What I've called "core moral judgment" is better labeled "*ersatz* moral judgment". But this move carries a number of dangers. First, neosentimentalism is supposed to capture everyday normative life (Gibbard 1990, 26; Blackburn 1998, 13), and it's an *empirical assumption* that most adult moral judgment is radically different from core moral judgment. The basic kind of moral judgment we see in children might be preserved without revision into adulthood, and it might well guide a great deal of adult moral judgment. As a result, if neosentimentalists cede core moral judgment, they risk neglecting a central part of our everyday normative lives.

Furthermore, several of the conditions set out in section 1 on an adequate sentimentalism apply to children's core moral judgment as well. Children enter into disagreements with each other and with their parents over matters in the moral domain. The emotions seem to figure importantly in children's moral judgment. It's likely that the emotions play a key role in leading children to regard moral violations as especially serious and authority independent; the emotions also seem to play a role in subserving a connection between moral judgment and motivation – children find rule violations to be emotionally upsetting, and they find it especially upsetting to witness another being harmed.[3] Finally, and as we will see in more detail in section 3, from a young age,

[3] The situation with motivation and moral judgment is rather complicated, as evidenced by the debate over internalism. For present purposes it's important to note that there are at least two different ways in which moral judgment is connected with motivation. First, moral violations fall into the class of rule violations and people are generally motivated not to violate rules. Secondly, moral rules often prohibit actions that are inherently upsetting, and hence to be avoided. Most saliently in the present context, harm in others

children engage in moral reasoning of the sort appealed to by moral philosophers like Toulmin and Geach. Thus, many of the central constraints for an adequate sentimentalist account must be addressed by an account of core moral judgment, and the dissociation problem suggests that the most prominent neosentimentalist solution is unavailable for young children. It's plausible that whatever the right account *is* for moral judgment in young children, that account will also apply to adults, with no radical changes along neosentimentalist lines.[4]

3. Towards a naturalistic sentimentalism

Thus far I've argued that the most prominent neosentimentalist view is an implausible account of moral judgment. Although this neosentimentalist account falls prey to its own sophistication, the dissociation problem cannot dislodge neosentimentalism unless there is a plausible alternative to take its place. I would hardly urge that we resuscitate the less sophisticated accounts like emotivism or subjectivism. Philosophers rightly abandoned those theories, and for the right reasons. However, I think that an approach that leans more heavily on psychology will give us a more promising account.

Philosophical sentimentalists in the 20th century tended to maintain that emotions are part of the *content* of a moral judgment. This isn't all that surprising really, since philosophy of language reigned supreme. What else is there to do but give an account in terms of the content of moral terms? However, if we approach this question with an eye to psychology rather than semantics, we will find a different way that emotion comes into play in moral judgment. Emotion concepts do not figure into the content of a moral judgment, rather, emotions play a role in leading us to treat as distinctive certain violations, including many of those we consider 'moral', like violations of harming others.

The basic idea is that core moral judgment depends on two mechanisms, a body of information prohibiting harmful actions and an affective mechanism that is activated by suffering in others. After sketching this approach, I'll argue that the account might deliver the explanatory goods promised by neosentimentalism without falling victim to the dissociation problem.

3.1. Core moral judgment depends on an *Affect-backed Normative Theory*

The empirical research on moral judgment indicates, in line with sentimentalism,

generates considerable negative affect and so people are motivated not to do those things. This complication doesn't affect the present point though, since both of these strands of motivation – rule-based and emotion-based – seem to be present in young children.

[4] Of course, it's possible that a neosentimentalist might maintain that the relevant emotion is something other than guilt. But there remains a serious challenge for this approach. Neosentimentalists would need to show that there is some emotion that fits into the neosentimentalist schema and which is sufficient to exclude non-moral cases. Further, in order to address the dissociation problem, the neosentimentalist would need to provide evidence that children have an early understanding of this emotion and of when the emotion is normatively appropriate. The difficulty of this project provides good reason to look elsewhere for a theory of moral judgment.

that core moral judgment is mediated by affective response. In a series of provocative experiments, James Blair found that psychopaths do not perform the way normal people do on the moral/conventional task (Blair 1995). Psychopaths, unlike normal adults, young children, autistic children, and non-psychopathic criminals, failed to draw a significant moral/conventional distinction in Blair's experiments. In addition, children with psychopathic tendencies were more likely to regard moral violations as *authority contingent* than other children with behavioral problems (Blair 1997). Furthermore, psychopaths were less likely than nonpsychopathic criminals to appeal to the victim's welfare when explaining why the moral violations were wrong. Rather, psychopaths typically gave conventional-type justifications (e.g., "it's not the done thing") for all transgressions. Blair and colleagues also found that psychopaths have a distinctive deficit to their capacity to respond to the distress cues of other people (Blair et al. 1997). This affective deficit is not found in non-psychopathic criminals or in autistic children (Blair 1999). Thus, apparently the one population with a deficit in moral judgment also has a deficit in affective response. This provides evidence that emotional response somehow mediates performance on the moral/conventional task.[5] It's not yet clear which affective mechanism is implicated in core moral judgment, but it is presumably some mechanism that responds to harm or distress in others. Two such mechanisms have been identified, one subserving "personal distress" (see, e.g., Batson 1991), and another subserving "concern" (see, e.g., Nichols 2001). Both of these mechanisms emerge quite early in development, well before the 2nd birthday. So, Blair's evidence on the psychopath's response to distress cues suggests that psychopaths have a deficit to either the Personal Distress Mechanism or the Concern Mechanism (or both). And it is this affective deficit that explains their deficit in core moral judgment.

Although core moral judgment seems to be mediated by affective response, the affective response alone does not capture core moral judgment. For there are many cases in which another person's harm or distress has considerable affective consequences for a witness, but in which one does not make a corresponding moral judgment. For instance, victims of natural disasters often lead us to feel both personal distress and concern without also leading us to judge that a transgression has occurred. We also respond to other people's suffering when the suffering is a result of an accident or when the suffering is inflicted for some greater benefit (as in inoculations). In these cases too, we often respond affectively without drawing any moral judgment.

Appeal to an affective response like personal distress or concern does not suffice, then, to explain moral judgment. One natural way to fill out the account is to maintain that core moral judgment also depends on a body of information that specifies a class of

[5] The claim that affect mediates performance on the moral/conventional task is corroborated by research on transgressions that are not moral but are affectively charged. In one experiment, a standard moral/conventional task was carried out in which the moral transgressions were replaced with disgusting transgressions (e.g., spitting in one's glass before drinking from it). The disgusting transgressions were distinguished from affectively neutral conventional transgressions on all the classic dimensions – disgusting transgressions were judged to be less permissible, more serious and less authority contingent than conventional transgressions (Nichols 2002a). In normal people, then, affect seems to infuse normative judgments with a special authority.

transgressions. For present purposes, the important prohibitions are those that focus on harmful actions. We can think of this as a "Normative Theory," a body of mental representations proscribing harmful transgressions that is present in individuals who are capable of core moral judgment. Among other things, this Normative Theory provides the basis for distinguishing wrongful harm from acceptable harm.

Although core moral judgment plausibly depends on both an affective mechanism and a Normative Theory, these two mechanisms are at least partly independent. The affective mechanisms responsive to others' suffering emerge very early, before the second birthday, but few if any researchers would maintain that children make core moral judgment before the age of two. This is plausibly because they haven't yet developed the Normative Theory. More interestingly, it's likely that much of the Normative Theory can be preserved even when the affective system is damaged. Despite their deficits in core moral judgment, psychopaths are, in a sense, perfectly fluent with normative argument – they are quite capable of identifying which actions are proscribed, and they can marshal reasons for why certain actions count as violations and other, superficially similar, actions don't count as violations. The problem with psychopaths seems to be that the affective contribution to moral judgment is seriously diminished and this shows up in their deficit at making the distinction that is at the heart of core moral judgment.

The proposal, then, is that there are two mechanisms underlying the capacity for drawing the moral/conventional distinction. One of these mechanisms is an affective mechanism, responsive to others' harms; the other mechanism is a body of information, a Normative Theory, proscribing a set of harmful actions.   On this account, core moral judgment derives from an Affect-backed Normative Theory. This quick sketch leaves open a number of important questions about the nature of the Normative Theory, the nature of the affective mechanism, and about how these two mechanisms conspire to produce the distinctive pattern of moral judgment that we see in the experimental work. But the sketch does suffice, I hope, to gesture towards a broadly sentimentalist account of moral judgment.

3.2. Meeting the constraints

We began by considering the constraints on an adequate sentimentalism. Moral judgments typically involve the emotions, but online emotional processing isn't required to make a moral judgment. Furthermore, moral disagreement and moral reasoning play important roles in our normative lives. Neosentimentalism offers a sophisticated account that meets these constraints, but the very sophistication of the account leads to the dissociation problem. The account of core moral judgment outlined in the previous section is obviously underdescribed. But it will be worth taking a brief look at how the account compares with neosentimentalism.

First, to return to the dissociation problem, the Affect-backed Theory account is, perfectly consistent with the evidence on young children. The affective mechanism that plausibly underwrites core moral judgment is present early in children. And the Normative Theory containing information about harm violations is also present in young children (see below). So the fact that core moral judgment emerges when it does poses no problem. Unlike neosentimentalism, the Affect-backed Theory account makes no commitments about the child's *understanding* of emotions.   But can the account address the constraints that neosentimentalism handles so impressively? I'll suggest the Affect-backed Normative Theory account does indeed provide the beginnings of an account that can meet the constraints.

On the Affect-backed Theory account, an affective mechanism plays a crucial role in moral judgment. Sentiments play a key role in leading us to treat norms prohibiting harmful actions differently from other norms. This fits. We care more about harm norms, we are more upset when they are flouted, our emotions are more closely attuned to these kinds of transgressions.[6] The emotional mechanisms that give harm norms this distinctive status are defective in psychopathy, and as a result, the capacity for core moral judgment is seriously compromised in psychopaths. Psychopaths also, famously, seem to lack the normal motivation associated with prohibitions against harming others. That is, these prohibitions seem to carry less motivational weight for them than they do for the rest of us. On the account I've sketched, it's plausible that the affective deficit is responsible both for the deficit in moral judgment and for the deficit in moral motivation. Thus, the account of core moral judgment falls comfortably in line with the sentimentalist claim that emotions play a crucial role in moral judgment.

The affective mechanism is thus crucial to core moral judgment. To explain how the theory accommodates the other constraints I will appeal to the role of the Normative Theory. To explain this, it will be useful to consider cases that involve a body of information specifying a set of transgressions that *don't* involve emotions in the way that core moral judgment does. Once again, evidence on children provides an instructive starting point.

Young children have an impressive facility with normative violations in general. As most parents of young children can testify, children often disagree with each other about what the rules in a given domain are and how those rules apply. This is most apparent when children dispute rules of games. But it emerges in many other domains as well, including etiquette, school rules, and rules of the house. Moreover, even three year olds are quite good at detecting transgressions of familiar rules as well as arbitrary novel transgressions. For instance, in one experiment on 3 and 4 year olds, the experimenter said "One day Carol wants to do some painting. Her Mum says if she does some painting she should put her helmet on" (Harris & Nunez 1996, 1581). Children were shown 4 pictures: 2 pictures depicted Carol painting, one with and one without a helmet; in the other 2 pictures, Carol is not painting, but in one of these pictures she has a helmet on. Children were asked, "Show me the picture where Carol is being naughty and not doing what her Mum told her?" The children tended to get the right answer. In addition to children's success in identifying transgressions, children are also able to give some justification for their choice. For instance, in the task described above, after they answer the question about which picture depicts a transgression, the children are asked "What is Carol doing in that picture which is naughty?" (1581). The children in these experiments tended to give the right answer even here – they invoked the feature of the situation that was not present, e.g., they noted that Carol isn't wearing her helmet. So, children are good at detecting violations of unfamiliar and arbitrary rules as well as violations of familiar rules (Harris & Nunez 1996).

In earlier work on the moral/conventional distinction, Judith Smetana (1985)

[6] There are other rules, like the rules prohibiting disgusting actions, that also seem to share these features with harm norms (Nichols 2002a).

presented preschool children with transgression scenarios in which the actual transgression is not specified. Rather, in lieu of transgression terms, she used nonsense words. Some transgressions were modeled on the criteria associated with conventional transgressions (context specific, explicit appeal to rules), other transgressions were modeled on criteria associated with moral transgressions (generalizable; child cries). For instance, in one "conventional" scenario children are told that Mary shouldn't piggle at school, but it is okay for her to piggle at home. For this scenario, children tend to infer that in another country, it's okay to piggle. Now, in order to move from the information to the conclusion, the child presumably relies on some inductive premise of the sort, "If an action is okay at home but not at school, it is likely that the action is okay in another country." That is, children seem to do something very like the reasoning in Geach's example:

> If an action is okay at home but not at school, then it's probably okay in another country.
> Piggling is okay at home but not at school.
> *Ergo*, piggling is okay in another country.

Indeed, what is especially striking about Smetana's finding is that children undertake this reasoning without knowing what piggling is!

If we return to the remaining constraints – disagreement, absent feeling, and reasoning – the above evidence suggests that children can treat nonmoral rules in ways that answer to all these constraints. Children can disagree about what the rules are, they can recognize the rules without having distinctive affect, and they can reason about the rules. The reasoning here is not about feelings, of course. Rather, children accept certain rules and they reason about their application.

In the case of moral judgment, I've argued that core moral judgment depends on a body of rules, a Normative Theory. That is, rules make an independent contribution to moral judgment. And it is the rules, on this account, that allow us to explain disagreement, absent feeling, and reasoning.[7] Just as children can disagree and reason over rules that don't excite emotion, so too can we all disagree and reason about rules that prohibit harming others. Toulmin gives the example of reasoning that I should return a book because I promised to return it and I should keep my promises. Young children do something that seems quite analogous when they judge that it's wrong to pull hair because it *hurts* the other person. Presumably the child reasons that it's wrong to pull hair because the pulling hair hurts the person and hurting people is prohibited. So, while emotions play a key role in moral judgment, the emotions are not invoked to solve this range of constraints. Disagreement and reasoning are features that moral judgment shares with the vast array of normative thinking that we find in children – it's driven by internalized rules.[8]

[7] Actually the situation with absent feeling is a bit more complex. For it's not yet clear whether core moral judgment really is preserved in the absence of all dispositions to feel. But in any case, just as one can embrace etiquette norms in the absence of all dispositions to feel, so too can one continue to embrace the harm norms in the absence of all dispositions to feel.

[8] Note that the kind of reasoning considered here, the kind that has been important to many metaethicists, is actually consistent with Jonathan Haidt's recent attack on moral

Obviously this brief sketch of a theory of moral judgment leaves open a huge range of questions. But we don't need to await the answers to these questions to see a stark difference between this relatively simple account and the spectacularly intellectualized account proffered by neosentimentalism. According to neosentimentalism, in moral discourse, we are arguing about the appropriateness of feeling some emotion. No doubt we sometimes do engage in such discussions over when emotions are appropriate. But much of the disagreement and argument we find in moral discourse can be accounted for more simply by adverting to the content of the Normative Theory. Of course, this would also allow us to explain why it is that when we disagree and argue about moral issues, it doesn't *seem* like we're talking about emotions at all. That's because typically we're *not* talking about emotions. We're talking about the content and implications of a largely shared Normative Theory.

4. Sentimentalism and the evolution of norms

Philosophers in the sentimentalist tradition are fond of pointing out that there is a striking connection between our emotions and our norms. We have norms prohibiting harming others and these norms are closely connected to our responses to suffering; we have norms against the gratuitous display of bodily fluids, and these norms are closely connected to our disgust responses.

Part of the story here, of course, is that there is a *consequent* connection between emotions and norms. Once a rule is established, we often find it upsetting to see the rule broken, even when the rule bears no direct relation to our emotional repertoire. For instance, if the school rule is that you can't have snack until you've finished your picture, children who saw Johnny starting his snack without finishing his picture, this would upset the children. And children tend to say that people should be punished for violating conventional rules. So there might be some emotional response that is easily elicited by rule-breaking. But this is something that the affect-backed theory account can easily take on board. For what the affect-backed theory says is that harm norms get a special status because of their connections with specific kinds of emotions.

On the above issue, traditional sentimentalists have no particular advantage over the theory I've been promoting. But there is a different connection between emotions and norms as well. The kinds of things that we are independently likely to find upsetting (e.g., disgusting actions, harmful actions) also happen to be the kinds of actions that are proscribed. Traditional sentimentalists had an independently motivated answer to this – the norms just *are* the relevant emotions. On the early sentimentalist accounts, subjectivism and emotivism, moral judgment is just reporting or expressing the feelings

reasoning (Haidt 2001). Haidt's claim is that moral judgment typically doesn't depend on *conscious deliberate* reasoning. That's consistent with the possibility that in some key cases, moral judgment *does* depend on conscious deliberate reasoning. More importantly, Haidt's claim is consistent with the possibility that a great deal of moral judgment depends on quick, nondeliberative reasoning over rules. And in the examples from Toulmin and Geach, it's plausible that this kind of reasoning is not typically a deliberative conscious process. Indeed, it would be a bad thing for Haidt's theory if it did exclude the possibility of the kind of reasoning promoted by Toulmin and Geach. For the evidence indicates that even children engage in this kind of moral reasoning.

that you have. So, since we have feelings of revulsion at harmful actions and at disgusting actions, it follows that we would have norms against these kinds of actions. The norms just are the emotions. In the more sophisticated neosentimentalist account, emotional activation isn't required, but the emotion concepts are still part of the very semantics of moral concepts. Either way, for philosophical sentimentalists, emotions are deeply, inextricably embedded in moral concepts.

On my view the norms are *not* the emotions. Nor are emotion concepts implicated in the semantics of moral judgment. Rather, norms make an independent contribution to moral judgment. But now this leaves a bit of a puzzle. If the rules are independent of the emotions, why is it that the rules happen to fit so well with our emotional endowment? Why do we have rules that prohibit actions that we are independently likely to find emotionally aversive? Call this the *coordination problem*. To address the problem, I want to look away from semantics, to history.

To explain the coordination between emotions and norms, I'll appeal to the role of cultural evolution. The hypothesis I want to promote is that emotions played a role in determining which norms survived throughout our cultural history. In particular, norms prohibiting actions likely to elicit *negative* affect will have enhanced cultural fitness. We can put this as an "Affective Resonance" hypothesis:

> Norms that prohibit actions to which we are predisposed to be emotionally averse will enjoy enhanced cultural fitness over other norms.

It's worth emphasizing that *obviously* there are other important factors in cultural evolution. The hypothesis is only that Affective Resonance will be one of the factors that influences cultural evolution.

There are general theoretical reasons to favor the Affective Resonance hypothesis. For instance, emotionally salient cultural items will be attention-grabbing and memorable, which are obvious boons to cultural fitness. In the case of norms, we also know that affect-backed norms, like the norms prohibiting disgusting actions, are regarded as more serious than other norms (Nichols 2002a). Again, the fact that we take these norms more seriously provides reason to think they would be more robust across the ages.

Despite these general theoretical virtues, the Affective Resonance hypothesis would be much more compelling if we had evidence for it. Ideally, we want *historical* evidence, since the hypothesis is that norms that are affect-backed will be more likely to survive throughout the changes wrought through history. The Affective Resonance hypothesis predicts that, *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.

The cultural evolution of etiquette bears out the prediction. Disgust is widely regarded as a basic emotion (Ekman 1994; Izard 1991; Rozin et al. 2000, 638-9), and, while there is cultural variation in the things that provoke disgust, bodily fluids are very common elicitors for disgust responses across cultures (Rozin et al. 2000, 647). Indeed, Haidt and colleagues maintain that it's useful to recognize "core disgust", which is elicited by body products, food, and animals (especially animals associated with body products or spoiled food) (Haidt et al. 1994).

Given this view of core disgust, the Affective Resonance hypothesis generates a specific prediction about the evolution of norms. Norms that prohibit core disgusting actions should be more likely to succeed than norms that are not connected to affective response. This prediction is impressively confirmed by a glance at the history of etiquette. In *The Civilizing Process,* Norbert Elias charts the history of etiquette in the West by reviewing etiquette manuals from the middle ages through the 19th century (Elias 1939/2000). He reports numerous instances in which the culture came to have prohibitions against some action involving bodily fluids (e.g., norms involving spitting and nose blowing), and in each case, these norms were preserved in the culture. A closer look at the most important manual, Erasmus' *On Good Manners for Boys,* corroborates our prediction more effectively. In Erasmus we find several norms that are not connected to core disgust and that did not survive:

> "When sitting down [at a banquet] have both hands on the table, not clasped together, nor on the plate" (281)
>
> "If given a napkin, put it over either the left shoulder or the left forearm." (281)

On the other hand, most of the norms in Erasmus' manual that prohibit core disgust actions are now so deeply entrenched that they seem too obvious to mention. Consider, for example, the following:

> "It is boorish to wipe one's nose on one's cap or clothing, and it is not much better to wipe it with one's hand, if you then smear the discharge on your clothing." (274)
>
> "Withdraw when you are going to vomit" (276).
>
> "Reswallowing spittle is uncouth as is the practice we observe in some people of spitting after every third word" (276).

I had independent coders evaluate a representative sampling of norms from Erasmus book, and their responses confirmed that the norms prohibiting disgusting actions were much more likely to survive than the other norms found in Erasmus (Nichols 2002b).

We can turn now to the norms at the center of our moral worldview, norms prohibiting harming others. The Affective Resonance hypothesis would predict that harm norms should have an advantage in cultural evolution. For 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.

Just as we've seen that norms prohibiting disgusting actions have been extremely successful, so too have harm norms done well historically. 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, Reiman 1985, Smith 1994).[9] Since we are disposed to respond aversively to even low level signs of distress, the trend in moral evolution further confirms the Affective Resonance hypothesis that norms will have enhanced cultural fitness when they prohibit actions which we're predisposed to find emotionally aversive.

Thus, it seems that we can explain the impressive coordination between emotions and norms by appealing to history rather than semantics. Emotional mechanisms prove to be a potent factor in the cultural evolution of norms. Norms are more likely to be preserved in the culture if the norms resonate with our affective systems by prohibiting actions that are likely to elicit negative affect. We find confirmation for this both in the history of etiquette norms and in the history of norms prohibiting harming others. Norms prohibiting disgusting and harmful actions seem to have thrived in our culture, whereas affect-neutral norms have proved much more feeble.

5. Conclusion

Emotions do make vital contributions to moral judgment, as sentimentalists have always maintained. However, the contribution of emotions isn't something that can be adequately gleaned from philosophical analyses. Rather, the case for the role of emotion is best made by looking at psychological evidence on moral judgment. Emotions drive a wedge between two different classes of normative judgment. Affect-backed normative judgment shows systematic differences from other kinds of normative judgment. But the basic capacity for moral judgment cannot be explained solely in terms of emotional responses. Internally represented rules make an independent contribution to moral judgment. The emotions play a crucial role in making some of these rules psychologically distinctive. Furthermore, emotion plays an important historical role in the fixing of norms in the culture. Norms that fit with our emotions have a greater cultural resilience.

The naturalized sentimentalism promoted here does not look much like the earlier philosophical accounts. Nonetheless, it's clear that Hutcheson, Hume, and subsequent sentimentalists were right to think that emotions are at the heart of moral judgment. Our normative lives would be radically different if we had a different emotional repertoire.

[9] Even Nietzsche makes a (rather wittier) observation in this spirit: "in those days, pain did not hurt as much as it does today" (1956, 199). Thanks to Walter Sinnott-Armstrong for reminding me of this line.

References:

Baier, K. (1958). *The Moral Point of View*. Ithaca: Cornell University Press.

Baron-Cohen, S., A.M. Leslie and U. Frith (1985). "Does the Autistic Child Have a 'Theory Of Mind'?" *Cognition*, 21, 37-46.

Baron-Cohen, S. (1995). *Mindblindness*. Cambridge, Mass.: MIT Press.

Batson, C. (1991). *The Altruism Question.* Hillsdale, N.J.: LEA.

Blackburn, S. (1984). *Spreading the Word: Groundings in the Philosophy of Language.* Oxford: Oxford University Press.

Blackburn, S. (1985). "Errors and the Phenomenology of Value." In T. Honderich (ed.), *Morality and Objectivity*. London: Routledge & Kegan Paul.

Blackburn, S. (1998). *Ruling Passions: A Theory of Practical Reason*. Oxford: Oxford University Press.

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, 1-29.

Blair, R. (1997). "Moral Reasoning and the Child with Psychopathic Tendencies," *Personality and Individual Differences*, 26, 731-739.

Blair, R. (1999). "Psychophysiological Responsiveness to the Distress of Others in Children with Autism," *Personality & Individual Differences*, 26, 477-485.

Blair, R., L. Jones, F. Clark, M. Smith and L. Jones (1997). "The Psychopathic Individual: A Lack of Responsiveness to Distress Cues?", *Psychophysiology*, 34, 192-198.

Brandt, R.(1950). "The Emotive Theory of Ethics." *Philosophical Review*, 59, 305-318.

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

D'Arms, J. and D. Jacobson (2000). "Sentiment and Value". *Ethics,* 110, 722-748.

Darwall, S., A. Gibbard, and P. Railton (1992) "Toward Fin de siecle Ethics: Some Trends," *Philosophical Review*, 115-189. Reprinted in Darwall et al. 1997. Page numbers from the reprinted version.

Dunn, J. and P. Munn (1987). "Development of Justification in Disputes with Mother and Sibling," *Developmental Psychology*, 23, 791-798.

Ekman, P. (1994). "All Emotions Are Basic," in P. Ekman and R. Davidson (eds.), *The Nature of Emotion*. New York: Oxford University Press, 15-19.

Elias, N. (1939/2000). *The Civilizing Process*. Translated by E. Jephcott. Malden, Mass: Blackwell.

Erasmus, D. (1530). *On Good Manners for Boys*. Translated by B. McGregor, in *Collected Works of Erasmus*, Vol. 25, ed. by J. Sowards. Toronto: University of Toronto Press, 1985.

Falk, W. (1953). "Goading and Guiding," *Mind* 53, 145-171.

Geach, P. (1965). "Assertion," *The Philosophical Review*, 74, 449-465.

Gibbard, A. (1990). *Wise Choices, Apt Feelings*. Cambridge, Mass: Harvard University Press.

Gill, M. (forthcoming). *The Human Nature Question*. Cambridge University Press.

Haidt, J. (2001). "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment," *Psychological Review*, 108, 814-834.

Haidt, J., C. McCauley, and P. Rozin (1994). " Individual Differences in Sensitivity to Disgust: A Scale Sampling Seven Domains of Disgust Elicitors," *Personality and Individual Differences*, 16, 701-713.

Harris, P. (1989). *Children and Emotion: The Development of Psychological Understanding.* Oxford: Blackwell.

Harris, P. (1993). " Understanding Emotion," In M Lewis and J. Haviland (eds.), *Handbook of emotions*. New York : Guilford Press, 237-246.

Harris, P. and M. Núñez (1996). "Understanding of Permission Rules by Preschool Children," *Child Development*, 67, 1572-1591.

Harris, P., T. Olthof, M. Meerum Terwogt, and C. Hardman (1987). Children's knowledge of the situations that provoke emotions. *International Journal of Behavioral Development*, 10, 319-344.

Izard, C. (1991). *The Psychology of Emotions.* New York : Plenum Press.

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

Nichols, S. (2001). “Mindreading and the Cognitive Architecture Underlying Altruistic Motivation,” *Mind & Language,* 16, 425-455.

Nichols, S. (2002a). “ Norms with Feeling: Towards a Psychological Account of Moral Judgment,” *Cognition,* 84, 221-236.

Nichols, S. (2002b). “ On the Genealogy of Norms: A Case for the Role of Emotion in Cultural Evolution,” *Philosophy of Science*, 69, 234-255.

Nichols, S. (2004). *Sentimental Rules*. New York: Oxford University Press.

Nietzsche, F. (1956). *The Birth of Tragedy and The Genealogy of Morals*. Trans. F. Golffing. New York: Doubleday, 1956.

Nunner-Winkler, G. and B Sodian (1988). “Children’s Understanding of Moral Emotions,” *Child Development*, 59, 1323-38.

Railton, P. (1986). “Moral Realism.” *Philosophical Review* 95.

Reiman, J. 1985. “Justice, Civilization, and the Death Penalty,” *Philosophy & Public Affairs*, 14, 115-148.

Rozin, Paul, J. Haidt and C. McCauley (2000). “Disgust,” in M. Lewis & J. Havilland-Jones (eds.), *Handbook of Emotions*, 2nd Edition. New York: Guilford.

Ruse, M. and E. Wilson (1986). “Moral Philosophy as Applied Science,” *Philosophy*, 61, 173-192.

Simner, M. (1971). “Newborn's Response to the Cry of Another Infant,” *Developmental Psychology 5*, 136-150.

Smetana, J. and J. Braeges (1990). “The Development of Toddlers’ Moral and Conventional Judgements,” *Merrill-Palmer Quarterly*, 36, 329-346.

Smetana, J. (1985). “Preschool Children’s Conceptions of Transgressions: Effects of Varying Moral and Conventional Domain-related Attributes,” *Developmental Psychology*, 21, 18-29.

Smetana, J. (1989). “Toddler’s Social Interactions in the Context of Moral and Conventional Transgressions in the Home,” *Developmental Psychology,* 25, 499-508.

Smetana, J. (1993). “Understanding of Social Rules,” In M. Bennett (ed.) *The Development of Social Cognition : The Child as Psychologist.* New York: Guilford Press, 111-141.

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

Stevenson, C. (1937). “The Emotive Meaning of Ethical Terms,” *Mind* 46, 14-31.

Stevenson, C. (1944). *Ethics and Language*. New Haven: Yale University Press.

Thompson, R. and M. Hoffman (1980). “Empathy and the Arousal of Guilt in Children,” *Developmental Psychology,* 15, 155-6.

Tisak, M. (1995). “Domains of Social Reasoning and Beyond,” In R. Vasta (ed.), *Annals of Child Development*, Vol. 11 . 95-130 London: Jessica Kingsley.

Toulmin, S. (1950). *An Examination of the Place of Reason in Ethics*. Cambridge, UK: Cambridge University Press.

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

Wiggins, D. (1991). “A Sensible Subjectivism,” In *Needs, Values, Truth: Essays in the Philosophy of Value*. Oxford: Blackwell. Page reference to reprint in Darwall, et al. 1997.