Sentimental Rules: On the Natural Foundations of Moral Judgment, by Shaun Nichols. Oxford: Oxford University Press, 2004. Pp. xi + 226. H/b £35.00.

This fascinating book is not only philosophically important, but also a pleasure to read. Nichols contrasts *rationalist* metaethical positions (inspired by Cudworth, Whichcote, and Kant) with *sentimentalist* ones (inspired by Shaftesbury, Hutcheson, and Hume); he briefly attacks rationalism, but his main goal is to formulate and defend a novel version of sentimentalism. To attack rationalism, Nichols distinguishes *conceptual* rationalism, according to which 'It is a conceptual truth that a moral requirement is a reason for action', from *empirical* rationalism, according to which 'It is an empirical fact that … our moral judgments derive from our rational faculties' (67). To attack conceptual rationalism, Nichols argues that the existence of psychopaths who make moral judgments but are not motivated by them is conceptually possible. (Strictly speaking, this refutes at most not conceptual rationalism, but rather the claim that 'it is a conceptual truth that people who make moral judgments are motivated by them'; Nichols notes this is in a footnote (72), but to my mind the point deserves greater emphasis.) To attack empirical rationalism, Nichols argues that *actual* (as opposed to conceptually merely *possible*) psychopaths have a 'defective capacity for moral judgment' which 'seems not to derive from a rational deficit, but rather from a deficit to an affective system' (82). Nichols thus rejects both versions of rationalism.

Considering sentimentalism, Nichols distinguishes traditional sentimentalist views like *emotivism* (defended by Ayer and Stevenson), according to which 'moral judgments are really expressions of one's feelings' (85), from more recent *neosentimentalist* views (defended by Blackburn, Gibbard, and Wiggins), according to which 'to think that X is morally wrong is to think it appropriate to feel some emotion [following Gibbard, Nichols focuses on *guilt*] … in response to X' (87-8). Nichols argues that neosentimentalism is 'an impressive achievement' (88) because it solves two problems that plague traditional sentimentalism: the problem of explaining how 'A person can judge something wrong even if he has lost all feelings about it', and the problem of explaining how 'Reasoning plays a crucial role in moral judgment' (86). Nevertheless, Nichols argues that 'the empirical evidence on moral judgment poses a serious problem for neosentimentalism' (96), namely the *dissociation problem*: 'According to neosentimentalism, the capacity for moral judgment depends on the capacity for judging the appropriateness of guilt. However, there are large populations of individuals [namely young children and children with autism] who have the capacity for moral judgment and lack the capacity for judging the appropriateness of guilt.' (92) So Nichols proposes his own version of sentimentalism, the *Sentimental Rules Account*, which is supposed to have the virtues of neosentimentalism while avoiding the dissociation problem.

As I understand it, the Sentimental Rules Account is an account of what Nichols calls the 'capacity for core moral judgment' (7). To explain what this capacity is, Nichols considers psychological experiments in which people are presented with canonical examples of morally impermissible actions, like hitting or pulling hair, and with canonical examples of purely conventionally impermissible actions, like talking out of turn or drinking soup out of a bowl. It is found that (even young) children, unlike psychopaths, on average attribute certain characteristics to a greater extent to the morally than to the purely conventionally impermissible actions. The characteristics in question are: seriousness, generalizability (i.e., being impermissible even in other cultures), authority independence (i.e., being impermissible even in the absence of any relevant prohibition by the authorities), and being impermissible on account of causing harm. Let us say then that a person has the *capacity for core moral judgment* exactly if the person attributes the above characteristics to morally impermissible *harmful* actions (to a greater extent than to

purely conventionally impermissible actions; it is not clear whether Nichols adopts this qualification, but for the sake of simplicity I omit it in what follows). As Nichols in effect notes (6-7), actions that are morally impermissible but not (directly) harmful, like cheating on one's taxes, fall outside the scope of the capacity for *core* moral judgment.

A problem that I see with the above understanding of the capacity for core moral judgment is that it is possible for a person who lacks the concept of morality to have this capacity. It is possible, for example, for a person who lacks the concept of morality to find actions like hitting or pulling hair disgusting *because* they are harmful. Evidence that Nichols himself adduces on how people judge actions that they find disgusting (21-4) suggests that such a person would judge actions like hitting or pulling hair to be impermissible because they are disgusting (and thus, by hypothesis, because they are harmful), and would attribute to such actions the characteristics of seriousness, generalizability, and authority independence. Such a person would thus have the capacity for core moral judgment despite lacking the concept of morality.

In any case, given the above understanding of the capacity for core moral judgment, the Sentimental Rules Account can be formulated as the empirical thesis that the following two conditions are individually necessary and jointly sufficient for having that capacity:

(1) Possession of a *normative theory* that prohibits certain harmful actions.
(2) Possession, at least during some developmentally critical period, of an *affective mechanism* that is activated by (observing or imagining) suffering.

Concerning the affective mechanism, Nichols notes that it differs from Hutcheson's 'moral sense', understood as 'the source of distinctive feelings … triggered by the perception of virtue and vice' (62): the affective mechanism can be activated for example by observing an accident victim, 'in the conspicuous absence of any judgment that a transgression [or even an action] has occurred' (63). Concerning the normative theory, Nichols notes that it need not prohibit *all* harmful actions (e.g., it need not prohibit *unintentionally* harmful actions): 'the normative theory provides the basis for distinguishing wrongful harm from acceptable harm' (17). Moreover, 'even a motley set of rules prohibiting certain behaviors will count as a normative theory' (16). Because the rules that comprise the normative theory prohibit harmful actions, and according to Nichols harmful actions are *affect backed* in the sense that observing or imagining such actions is 'likely to elicit strong negative affect' (at least in psychologically normal people), Nichols calls these rules 'Sentimental Rules' (18). In general, Sentimental Rules are rules prohibiting affect-backed actions; Nichols claims that not only harmful actions, but also disgusting actions are affect backed, so that rules prohibiting disgusting actions are also Sentimental Rules (25).

Why should one accept the Sentimental Rules Account? I will address this question by examining successively what I take to be the three components of the account: (a) the (alleged) necessity of condition (1), concerning the normative theory, (b) the necessity of condition (2), concerning the affective mechanism, and (c) the joint sufficiency of the two conditions (for having the capacity for core moral judgment).

(a) Concerning condition (1), Nichols in effect argues that people in 'all of the populations studied' who have the capacity for core moral judgment 'have knowledge of the conventional rules' and thus possess a normative theory (16). It does not follow, however, that possessing a normative theory is *necessary* for having the capacity for core moral judgment: maybe some *particularists*, understood as people who make moral judgments on a case-by-case basis and reject general moral rules, have the capacity for core moral judgment despite lacking a normative theory. Nichols in effect responds that apparently people in general are not particularists (18 n. 6). But even if people *in general* are not particularists, the existence of just *a few* particularists

having the capacity for core moral judgment would suffice to refute the claim that possessing a normative theory is *necessary* for having that capacity. My point is not that such particularists are just *conceptually* possible: as Nichols notes, his aim is 'to provide an empirical account of what moral judgment is, rather than a semantic or conceptual analysis of what moral terms and concepts mean' (110). (So necessity and sufficiency in the Sentimental Rules Account are to be understood not conceptually, but presumably—though Nichols is not clear on this—causally or nomologically.) My point is rather that, for all Nichols has said, such particularists may in fact exist. Maybe they are exceptional, or even pathological, but Nichols does take exceptional or pathological people like children with autism (10-11) and psychopaths to be relevant to an empirical account of moral judgment. So it seems that Nichols needs more of an argument to support the first component of the Sentimental Rules Account.

(b) Concerning condition (2), Nichols adduces the following reasoning:

> Since the experiments indicate that the disgust system provokes nonconventional responses to questions about permissibility, seriousness, authority contingency and justification, we have evidence that nonconventional responses to these questions can be induced by affective response. There is independent reason to think that suffering in others inspires considerable affective response … [To repeat:] Harm-scenarios generate affective response, and affective response can provoke nonconventional answers to the standard moral/conventional questions. So it is reasonable to suppose that the affective response to harm-scenarios does play a crucial role in leading subjects to judge that hitting others and pulling hair is impermissible, very serious, and not authority contingent. (25)

There are several gaps in the above reasoning. First, given the premise that disgusting actions generate a negative affective response, some argument is needed to secure the conclusion that the 'nonconventional answers' (provoked by disgusting actions) are *induced* by this negative affective response. Second, even if that conclusion is somehow secured, it is quite a leap to infer from it the general claim that *all* (kinds of) actions—including harmful ones—which generate a negative affective response provoke nonconventional answers induced by this negative affective response. (Nichols in correspondence has in effect stated that he does not want to commit to the above general claim. But the weaker claim that nonconventional answers '*can* be induced by affective response'—empasis added—does not suffice to reach a conclusion about the affective response generated by harmful actions in particular.) Third, even if somehow the claim is secured that in most people harmful actions provoke nonconventional answers induced by the negative affective response that the actions generate, it does not follow that generating a negative affective response is *necessary* for provoking nonconventional answers. As in my discussion of condition (1), the point is that, for all Nichols has said, some (exceptional, or even pathological) people may exist who have the capacity for core moral judgment without possessing (or having ever possessed) an affective mechanism activated by suffering. So it seems that Nichols needs more of an argument to support the second component of the Sentimental Rules Account.

(c) Concerning finally the joint sufficiency of conditions (1) and (2), I should note first that Nichols does not make such a sufficiency claim explicit. But such a claim is implicit in his exposition; for example, he does not consider the possibility that some *third* condition is also necessary for having the capacity for core moral judgment. I don't see, however, how Nichols can exclude this possibility. Maybe, to take a hypothetical example, the capacity to feel pain is such a third condition: maybe people who lack this capacity (e.g., people with congenital analgesia) also lack the capacity for core moral judgment despite possessing both a normative theory and an affective mechanism activated by suffering. The question is empirical, and the example is purely hypothetical; but the general point is that, for all Nichols has said, any number of conditions may be necessary for having the capacity for core moral judgment. (Nichols in

correspondence has in effect agreed, and has stated that he is not committed to the sufficiency claim.)

It should be clear by now that I find inadequate the *direct* support that Nichols provides for the Sentimental Rules Account. Nichols, however, also tries to support the account less directly, by arguing that it has the virtues of neosentimentalism while avoiding the dissociation problem. More specifically, among other things, Nichols argues that the Sentimental Rules Account explains (i) 'how emotion plays a role in linking moral judgment to motivation' and (ii) 'how moral judgments can be made in the absence of emotional response' (98). Concerning (i), however, it is not clear to me how the Sentimental Rules Account explains the link between moral judgments and motivation when the moral judgments concern only actions that are *not* (directly) harmful, like cheating on one's taxes: as I indicated above, Nichols does not claim that such actions are likely to elicit strong negative affect. Concerning (ii), and more generally, note that an account of how core moral judgments are in fact made (or caused) is one thing, and an account of what conditions are (causally) necessary and sufficient for having the capacity for core moral judgment is another thing. (This is so even if the material equivalence is true that one has the capacity for core moral judgment if and only if one in fact makes core moral judgments.) In accordance with the thrust of the book, I formulated the Sentimental Rules Account as an account of the latter kind (about necessary and sufficient conditions), but in his treatment of (ii)—and elsewhere—Nichols seems to understand it as an account of the former kind (about causal mechanisms). I am not sure how to resolve this tension; maybe Nichols implicitly takes the Sentimental Rules Account to have two parts, a part about necessary and sufficient conditions and a part about causal mechanisms, each part being partly inspired by the other. (Nichols in correspondence has conceded that he was not very clear on the point, and has stated that he intended to make not the necessity claims, but rather the claim that 'two factors are causally implicated in the *normal* acquisition of the capacity [for core moral judgment]'—emphasis added.)

Nichols devotes a large portion of the book to what he takes to be 'a significant shortcoming' (115) of his theory, namely the *coordination problem*: 'The Sentimental Rules account has no principled explanation for the coordination between the norms we have and the emotions we have' (116). Nichols tries to provide (the beginning of) such an explanation by defending an 'Affective Resonance' hypothesis: 'Normative prohibitions against action X will be more likely to survive if action X elicits … negative affect' (129). He argues that this hypothesis can explain the 'characteristic evolution of harm norms', namely that 'harm norms tend to evolve from being restricted to a small group of individuals to encompassing an increasingly larger group' (143). For example, 'in European culture the prohibition against harming others seems to have been expanded to prohibit cruelty to animals' (144). I will not go over the details of Nichols's argument, but I wish to raise a worry about his claim that the characteristic evolution of harm norms can be explained better by the Affective Resonance hypothesis than by the appeal to 'moral progress' made by moral realists (like Brink and Sturgeon). The worry is that the Affective Resonance hypothesis seems hard pressed to explain the virtual extinction of normative prohibitions against actions like inoculating children: as Nichols himself notes, 'Knowing that inoculations are for the best does not eliminate the discomfort one feels witnessing a child get inoculated' (155). Moral realists, on the contrary, can provide an explanation by arguing that people have 'gradually come to recognize' (150) the moral fact that 'inoculations are for the best'. (Note that the Affective Resonance hypothesis, being probabilistic, is not *refuted* by my example; but maybe one could also find other examples of

actions which elicit negative affect and yet are no longer prohibited because they are for the best.)

In the last chapter of the book, Nichols proposes a 'Humean' argument against the objectivity of morality:

> 1. Rational creatures who lack certain emotions would not make the moral judgments that we do.
> 2. There is no principled basis for maintaining that … all rational creatures *should* have the emotions.
> [Thus: 3. There is no principled basis for maintaining that all rational creatures should make the moral judgments that we do.] (185)

Concerning premise 1, Nichols claims that it is 'plausible, though not certain, that Martians who lacked entirely an affective response to suffering would not share our harm norms' (186). This claim apparently relies on what I called the second component of the Sentimental Rules Account, namely the thesis that possessing (or having possessed) an affective mechanism activated by suffering is *necessary* for having the capacity for core moral judgment. As we saw, however, the necessity in question is presumably causal, not conceptual, so from the above thesis nothing follows about Martians who would not share our causal make-up. In response Nichols might replace 'would not' with 'might not' in premise 1. But this would wreck his reasoning: the resulting premise may well be true, but some thought will show that in conjunction with 2 it does not entail 3.

In conclusion I would like to make clear that, to my mind, the problems which I raised above detract only slightly from the value of Nichols's book. The book is in general clearly and carefully written, and it holds the reader's attention even during the—sometimes lengthy—digressions. The book is also amazingly interdisciplinary. I find particularly noteworthy the fact that, in order to address empirical questions that arise at several places, Nichols has performed his own experiments; he uses the results to good account. On the whole, the book is a splendid contribution to the small but burgeoning field of 'empirically informed ethics'. It deserves to be widely read.

(I am grateful to Justin D'Arms, Stephen Darwall, Aviv Hoffmann, Peter Railton, and especially Shaun Nichols for comments on an earlier draft of this review. Thanks also to Neera Badhwar, Jeanette Kennett, and Stephen Stich for help.)

Department of Philosophy and Religious Studies PETER B. M. VRANAS
Iowa State University
402 Catt Hall
Ames, IA 50011
USA
vranas@iastate.edu