Explaining Compatibilist Intuitions about Moral Responsibility: A Critique of Nichols and Knobe's Performance Error Model

***Scott Kimbrough**, Jacksonville University*

An emerging movement called "experimental philosophy" provides a new angle on the venerable debate about moral responsibility and determinism. Philosophers claiming that moral responsibility is incompatible with determinism ("incompatibilists") generally insist that common sense is on their side, that the concept of moral responsibility employed by ordinary people requires an ability to do otherwise. Compatibilists deny this of course, claiming to be the real representatives of common sense; but they are accused of revisionist concepts of moral responsibility that evade rather than solve the problem posed by determinism. Enter "experimental philosophy." Armed with questionnaires, experimental philosophers conduct empirical studies to explore and explain the judgments about moral responsibility that ordinary people actually make.[1] For example, in their paper "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions," Shaun Nichols and Joshua Knobe approach the issue in two steps: first, they devise empirical studies to determine what ordinary people's "intuitions" about moral responsibility and determinism really are; and second, they use the results of the studies to assess hypotheses about the psychological processes responsible for generating those intuitions.[2] The second step is crucially important. For when conflicting intuitions turn up in the questionnaire studies, Nichols and Knobe hope to advance the philosophical debate by assessing the reliability of the underlying psychological processes.

In their paper, Nichols and Knobe dispute the significance of earlier studies that demonstrated the existence of widespread compatibilist intuitions among test subjects.[3] Their critique turns on a contrast between two main kinds of moral intuitions: "concrete" intuitions that are (largely compatibilist) responses to specific cases, and "abstract" intuitions that are (largely incompatibilist) responses to general questions about the relation between determinism and moral responsibility. According to Nichols and Knobe, concrete compatibilist intuitions are the result of "performance errors" caused by the distorting influence of emotion on moral judgment. After describing the relevant data, I will dispute their argument on two main grounds. First, they fail to establish that compatibilist intuitions should be deemed performance errors. Second, I question the reliability of the psychological processes that generate the incompatibilist intuitions expressed by test

subjects. Although these criticisms largely fall within the empirically oriented framework favored by experimental philosophers, they also point to an important limitation of the approach.

Before getting into these criticisms, however, it's necessary to sketch Nichols and Knobe's data and their explanation thereof. Test subjects were presented with a description of two universes, A and B. Universe A is completely deterministic. In the less technical language geared for the test subjects: "everything that happens is completely caused by whatever happened before it."[4] This includes human decisions, which test subjects are told "*had to happen*" as they did. In contrast, human decisions are not determined by previous events in Universe B: "each human decision *does* ***not*** *have to happen* the way that it does."[5] When asked which universe more closely resembles our own, 90% of subjects chose the indeterministic Universe B. Subjects were then randomly assigned to the "concrete condition" or the "abstract condition." In the concrete condition, they were asked to respond to a vignette about a man named Bill in Universe A who kills his family in order to be with his secretary. Seventy-two percent of respondents answered "yes" to the question, "Is Bill fully morally responsible for killing his wife and children?" That's a strong compatibilist majority. However, in the abstract condition, test subjects were asked, "In Universe A, is it possible for a person to be fully morally responsible for their actions?" An even stronger majority (86%) responded "no"—a striking incompatibilist response.

To understand these competing intuitions, Nichols and Knobe suggest investigating the psychological processes that generate them, noting that "there has been remarkably little discussion about *why* people have the intuitions they do."[6] In their view, compatibilist intuitions are suspect due to the role of emotion in generating them:

> Our hypothesis is that when people are confronted with a story about an agent who performs morally bad behavior, this can trigger an immediate emotional response, and this emotional response can play a crucial (distorting) role in their intuitions about whether an agent was morally responsible. In fact, people may sometimes declare such an agent to be morally responsible despite the fact that they embrace a theory of responsibility on which the agent is not responsible.[7]

Note the rhetorical positioning here: compatibilist intuitions are "declared" (presumably contrary to subjects' *real* theoretical convictions) whereas incompatibilist intuitions are "embraced" as revealing "a theory of responsibility" to which respondents are presumably committed. Nichols and Knobe call this explanation the "performance error" model. On this model, incompatibilist theory reflects the criteria that are ordinarily applied in the psychological process that generates

judgments of moral responsibility, while compatibilist intuitions are products of affective interference or bias in that process.

To understand Nichols and Knobe's performance error model, it is important to remember that they intend the term "error" in a purely psychological sense. In calling compatibilist intuitions performance errors, Nichols and Knobe do not pretend to prove that those judgments are *false*.[8] Rather, "error" in the relevant sense is judged relative to moral *competence*, the psychological system functionally responsible for generating judgments of moral responsibility. A performance error is thus a psychological malfunction—a failure to execute one's moral competence without interference from psychological factors disruptive to that competence. Nevertheless, if compatibilist intuitions are "errors" in this purely psychological sense, Nichols and Knobe suggest that we have reason to be suspicious of their *reliability*. As the products of emotional bias, compatibilist intuitions issue from a source that we have independent reasons not to trust. Note that this form of argument conveniently sidesteps any need to presuppose or defend the truth of incompatibilist theories of moral responsibility. The performance error model need only establish that ordinary people do in fact have an incompatibilist theory of responsibility and that their moral competence involves the application of that theory. Then, if that competence is disrupted by a strong emotional response, the reliability of the resulting intuitions is naturally called into question.

Nichols and Knobe also give careful consideration to an alternative explanation of competing moral intuitions: the "affective competence" model, according to which emotion or affect plays a functional role in the competence underlying judgments of moral responsibility.[9] The basic idea is that the normal process of moral judgment involves affect, and that the theoretical beliefs about responsibility elicited in the studies play no significant role in that process. In response, Nichols and Knobe actually concede the point that normal moral judgment involves affect. However, they argue that this concession is consistent with the performance error model: "affect serves *both* as part of the fundamental competence underlying responsibility judgments *and* as a factor that can sometimes lead to performance errors."[10]

Nichols and Knobe support this hybrid model on the basis of a further study that compared test subjects' intuitions in response to "high affect" and "low affect" conditions. In the high affect condition, 64% of subjects answered "yes" when asked whether Bill, who stalks and rapes a stranger in (deterministic) Universe A, is fully morally responsible for his action. In the low affect condition, a comparatively small 23% of subjects answered "yes" when asked whether Mark, who arranges to cheat on his taxes in (deterministic) Universe A, is fully morally responsible for his action.[11] According to Nichols and Knobe, the best way to explain this discrepancy is by reference to the distorting influence of affect. Comparatively cool reflection on tax evasion yields responses that are consistent with the incompatibilist theory of responsibility uncovered by the abstract condition. But

when confronted with the emotionally charged rape vignette, subjects' emotional responses cause them to override their own theoretical convictions. Thus, although affect may play a functional role in judgments of responsibility, high affect can cause subjects to disregard a valid excuse—indeed, the very excuse most test subjects accepted while judging Mark's less emotionally provocative transgression.

In contrast, according to Nichols and Knobe, the affective competence model cannot readily explain the different responses to the high affect and low affect conditions. They consider the possibility that the low affect case may "fail to trigger our competence with responsibility attribution, and so we should not treat those responses as reflecting our normal competence."[12] They go on to note that it would "take significant work to show that such everyday cases of apparent responsibility attribution don't really count as cases in which we exercise our competence at responsibility attribution." Luckily, however, in earlier work Nichols reviews one useful way of distinguishing moral responsibility from conventional responsibility.[13] Already by the age of three, test subjects are able to distinguish between moral rules and conventional rules. Moral rules are distinguished, he notes, by a number of features including "authority independence." For example, when asked whether it would be OK to wear your pajamas to school if the teacher says so, three year olds tend to answer in the affirmative. They do not, however, agree that it would be OK to hit fellow students if the teacher says so. Mark's transgression—cheating on taxes—is a violation of an authority dependent rule like the rule against wearing pajamas. If the government said it was OK not to pay taxes, it would be OK not to pay taxes. This contrasts with moral violations like Bill's, involving the infliction of physical harm. Indeed, Nichols argues that rules against harm are central to morality: "Core moral judgment depends on two mechanisms, then, a normative theory prohibiting harm to others, and some affective mechanism that is activated by suffering in others."[14] Thus, if Nichols is correct in his earlier characterization of moral rules, there is an argument to be made that Mark's case does not involve an exercise of *moral* competence, whereas Bill's case provides a core instance.

Nevertheless, it may well be that Nichols and Knobe could replicate their experiment by providing a different, unproblematically immoral action for the low affect vignette—stealing a tube of caulk, say, or breaking a promise to attend a committee meeting.[15] So for the sake of argument, assume that Mark's transgression is a moral one, not merely a conventional one. The key issue in my view is the supposition that compatibilist intuitions in the high affect condition involve *error* in attributing moral responsibility, while incompatibilist intuitions reflect normal moral competence and hence do not raise the same concerns about reliability. I oppose these claims in two main ways: by arguing that compatibilist intuitions reflect our underlying competence in attributing moral responsibility, and by questioning the reliability of the psychological processes responsible for incompatibilist intuitions.

First of all, it's important to emphasize the functional centrality of affect to moral competence. Nichols and Knobe concede that affect plays a role, but underestimate its importance. Recent theories emphasizing the primary role of emotion in causing moral judgment erode the case for the performance error model. That model, as we've seen, is committed to the claim that an incompatibilist theory of moral responsibility is applied as part of moral competence. Such "rationalist" models of moral competence are increasingly giving way to theories that more centrally feature the causal role of emotion. Haidt's influential social intuitionist model, for example, holds that conscious moral reasoning typically does not play a causal role in the production of moral judgments, but instead provides *post hoc* rationalizations thereof.[16] Haidt's theory is supported by findings such as the emotional deficits of psychopaths[17] and neuro-imaging studies indicating that emotional centers of the brain are active in the normal process of making moral judgments.[18] Nichols and Knobe cite such studies themselves, but again I think they underestimate the importance of their findings. Taken together, the results tell against performance error model's assumption that moral competence typically involves the application of a theory (incompatibilist or otherwise) of moral responsibility. Recent psychological theories such as Haidt's lean more to the side of the affective competence model.

Franz De Waal's work on non-human primates provides further impetus for moral psychology's recent move towards the affective competence model. De Waal argues that human moral judgment is built on the same emotional platform as similar judgments by our fellow primates. In his view, "evolutionary parsimony" dictates that we should explain human capacities in similar terms to our primate relatives whenever possible.[19] Clearly, sophisticated theoretical views like those reflected in incompatibilist intuitions do not play a role in our primate relatives, from which De Waal concludes that the processes generating concrete moral judgments by human beings are similarly insulated from higher theoretical cognition. If De Waal is right, the performance error model is mistaken because it assumes that moral competence involves the application of a tacit theory of moral responsibility. Consequently, the fact that compatibilist intuitions violate test subjects' incompatibilist theoretical intuitions lends little support to the hypothesis that compatibilist intuitions are performance errors.

Reflection on evolutionary explanations of the origin of our moral competence further weakens the claim that compatibilist intuitions are performance errors. Evolutionary theorists seeking to explain the origin of morality focus on such things as the effectiveness of moral codes in fostering group stability and cooperation. As Richard Joyce puts the point, "the evolutionary function of moral judgment is to provide added motivation in favor of certain adaptive social behaviors," serving as "a kind of social glue, bonding individuals together in a shared justificatory structure and providing a tool for solving many group coordination problems."[20] Clearly, moral

judgments can serve this function in a deterministic universe. Indeed, Hume goes so far as to insist that determinism is *necessary* for this function: if punishment did not cause the inhibition of bad behaviors, for example, it would be pointless to inflict it.[21] But we needn't go as far as Hume in order to see that compatibilist intuitions about Bill's moral responsibility subserve the evolutionary function of moral judgment. Furthermore, if people like Bill were not held morally responsible, it is hard to see how judgments of moral responsibility could serve their function of motivating adaptive social behaviors and discouraging anti-social behaviors.[22] This tells against the performance error model. For when determining whether a particular intuition about moral responsibility should be deemed a performance error or a functionally successful exercise of moral competence, one important criterion is surely whether that intuition contributes to fulfilling the evolutionary function of moral competence. This criterion should carry more weight than consistency with incompatibilist theories, even if those theories are tacitly held by the very people who judge Bill morally responsible. In other words, the performance error model assumes that moral competence involves the application of an incompatibilist theory of moral responsibility; but this assumption loses considerable plausibility to the extent that compatibilist intuitions satisfy the evolutionary function of moral competence more effectively than would a successful application of the incompatibilist theory.

Even if Nichols and Knobe are wrong that moral competence involves the application of an incompatibilist theory, it may be suspected that the presence of *high* affect still raises the suspicion of performance error. In "An Experimental Philosophy Manifesto," Nichols and Knobe write:

> Clearly, an intuition developed in a jealous rage is less trustworthy than one developed after calm and careful consideration. Thus, if our hypothetical philosopher discovers that her intuition about a case is driven by such distorting emotional reactions, this will and should affect how much she trusts the intuition.[23]

However, there is no valid reason to suspect that the high affect condition is relevantly similar to a jealous rage. First of all, Nichols and Knobe do not provide any evidence that the affective reaction to the vignette about Bill is actually "high." Perhaps if they had asked survey respondents to rate their degree of outrage, high marks would have been forthcoming due to the seriousness of Bill's moral transgression. But do survey questions about hypothetical strangers in hypothetical universes really cause intense emotional reactions? I have my doubts, but the question would need to be settled by the application of more objective measures of emotional response, such as facial expressions or heart rate.

The strength of respondents' emotional reaction is of course an empirical question that could be studied. However, even if it turned out that the intensity of emotional reactions to cases like Bill's is indeed high, that does not by itself provide evidence of performance error or bias. Admittedly, it is beyond question that affect *can* bias judgment, as in Nichols and Knobe's example of judging whether a poem written by a good friend is moving.[24] But the relevant question is not the *degree* of affect (high or low), but whether its causal role contributes to or disrupts competence. For example, the fact that one is moved to tears by reading a poem is not evidence of performance error in the judgment that the poem is moving. The accusation of performance error must be backed by a specific explanation of how or why the emotional reaction in question interferes with the underlying competence. By this standard, emotional partiality to a friend does raise suspicions about the judgment that a poem is moving; tears at the beauty or sadness of the poem do not, no matter how strong their intensity.

Consider a parallel example involving moral judgment. If a parent is inclined to judge that his child is innocent of a crime despite strong evidence of guilt, there is a clear case to be made that parental love is interfering with the parent's competence for attributing moral responsibility. However, in the case of test subjects judging Bill for stalking and murdering a stranger, the only emotional reactions plausibly involved are the very kind of emotional reactions routinely involved in normal moral competence. After all, Bill is a hypothetical person in a hypothetical universe, so none of the ordinary sources of emotional bias or partiality in moral judgment, such as parental love, could possibly be involved. So even if the emotional reaction to Bill's case is in fact quite strong, the intensity of that reaction lends no credence to the claim that a performance error has occurred. I conclude that the presence of high affect in the psychological process leading to compatibilist intuitions does not support the hypothesis of performance error.[25]

Nichols and Knobe would probably respond by asking how, without attributing error, to explain the difference between the high affect and low affect conditions in their study. But the difference needn't reflect any error. It may instead reflect the way excuses work: the weaker a moral violation, the more likely we are to accept an excuse. As J.L. Austin points out, an excuse that works for one transgression may be unacceptable for another: [26]

> It is characteristic of excuses to be "unacceptable": given, I suppose, almost any excuse, there will be cases of such a kind or of such gravity that "we will not accept" it…We may plead that we trod on the snail inadvertently: but not on a baby—you ought to look where you are putting your great feet. Of course it *was* (*really)*, if you like, inadvertence: but that word constitutes a plea, which is not going to be allowed, because of standards. And if you

> try it on, you will be subscribing to such dreadful standards that your last state will be worse than your first.[27]

Nichols and Knobe miss this point about normal moral competence. As is typical in philosophical discussions of determinism, they treat the fact that an action "had to happen" because of prior conditions as a blanket excuse covering all actions equally. Consequently, they take the difference in how the excuse applies in the high affect and low affect conditions as evidence of a performance error in the high affect condition.[28] However, as Austin's example shows, excuses like "I didn't notice" or "I couldn't help it" or "It's an inevitable product of my genetics and upbringing" get a different reception depending on the seriousness of the offense. Serious violations like Bill's may accordingly mark contexts in which the excuse "It *had* to happen" is found unacceptable, rather than an emotion-induced misfire of normal competence. Incompatibilists would no doubt object that excuses should not be sensitive to the seriousness of an offense in this way, but that complaint in no way contradicts my claim that it is a part of normal moral competence to make such distinctions. As such, the difference between the high affect and low affect conditions can be smoothly accommodated by the affective competence model.

Of course, the points so far establish at most that compatibilist judgments are part of normal moral competence, not necessarily that those judgments are *true*. But it does undermine the specific objection Nichols and Knobe raised about their reliability. Neither the presence of high affect, nor the failure of compatibilist intuitions to comply with incompatibilist theories of moral responsibility, justifies the charge of performance error.

Although Nichols and Knobe question the reliability of the psychological processes that lead to compatibilist intuitions, they do not devote any attention to the processes generating incompatibilist intuitions. However, there is substantial reason to be suspicious of those processes. If De Waal and Haidt are right that moral competence typically operates independently of higher level theoretical beliefs, those beliefs must have some other psychological source. I do not pretend to know what that source is with any confidence. But a consideration of some of the candidate sources is no occasion for optimism.

For example, suppose that incompatibilist intuitions are semantic intuitions: when we reflect on our concept of moral responsibility, it strikes us that causal determination always cancels moral responsibility. But how reliable are such semantic intuitions? Ironically, one of the main motivations for the experimental philosophy movement is skepticism about such introspective reports about our concepts. Jesse Prinz offers reasons for such skepticism about introspective semantic intuitions, which he associates with the traditional philosophical method of "conceptual analysis":

> First, introspection, like all memory retrieval, is a constructive process. What we recall often depends on beliefs, expectations, norms, context, and other factors. It is prone to confabulation and distortion. It can be heavily influenced by our theories, by social pressures, and by background knowledge. Second, conceptual information is often stored in the form of exemplars or paradigm cases. Philosophers often use conceptual analysis to identify necessary and sufficient conditions. If this is done on the basis of stored exemplars, it will inevitably require the addition and subtraction of information.[29]

Prinz's point about exemplars is particularly important. If incompatibilist intuitions derive from exemplars of constraint, such as paradigm cases of coercion, it is no surprise that they contradict normal moral judgments about less top-of-mind cases such as the concrete Universe A vignettes. But this shouldn't cause us to doubt judgments about non-exemplary cases.[30] I'm reminded of an exercise I have my introduction to philosophy students perform as they learn about the Socratic method. One student offers a definition of some ordinary term like "chair" or "knife," and the other students pelt them with counterexamples. A student who begins this process by defining chair as "a piece of furniture you sit on" has not thereby shown a theoretical commitment to classifying sofas as chairs. Rather, he has shown how unreliable abstract semantic intuitions tend to be.[31]

But perhaps incompatibilist intuitions are not semantic intuitions. Where else might our tacit moral theories come from? Haidt proposes that moral reasoning reflects *culturally supplied* "a priori moral theories":

> A priori moral theories can be defined as a pool of culturally supplied norms for evaluating and criticizing the behavior of others. A priori moral theories provide acceptable reasons for praise and blame (e.g., "unprovoked harm is bad"; "people should strive to live up to God's commandments").[32]

He goes on to suggest reinterpreting much of the psychological literature on moral reasoning as a kind of ethnography of a priori moral theories. If Haidt is right about the psychological origin of a priori moral theories, there is ample reason to doubt the reliability of the incompatibilist intuitions elicited in the abstract condition of Nichols and Knobe's study. Given the cultural variability of moral values, the process can clearly lead to mutually inconsistent conclusions. To dispel this particular suspicion about the reliability of the incompatibilist intuitions discovered in their study, Nichols and Knobe would need at the very least to provide evidence that incompatibilist theories of

moral responsibility are found cross-culturally. And even then, the psychological process leading to this cross-cultural consensus would need to be assessed for its reliability.

Assuming this hurdle could be cleared, the reliability of incompatibilist intuitions would still not be established because even a *generally* reliable psychological process may be unreliable in certain contexts. For example, suppose for the sake of illustration that incompatibilist theories of moral responsibility are generated by general inductive learning capacities. These capacities have served us well on the whole; they would never have made it through natural selection otherwise. But consider how inductive learning might work in the case at hand. 90% of Nichols and Knobe's test subjects likened our universe to indeterministic Universe B. Using this assumption, together with the conviction that constraint is often a good excuse, test subjects may have inferred that moral responsibility is impossible in Universe A. Can we infer that such an inference is likely to be correct, because it makes use of generally reliable inductive processes? No. For despite the general reliability of inductive learning, the inductive processes causing 90% of respondents to believe that our universe is indeterministic cannot be independently vouched for with respect to so vexed an issue as the truth or falsity of determinism. And there is good reason to be suspicious of the processes responsible for so bold a theoretical hypothesis: normal human learning notoriously depends upon probability heuristics that work reliably enough in ecological context, but that are prone to fallacies such as hasty generalization and the conjunction fallacy.[33] Faced with pairs of individuals who make different choices despite the same *salient* circumstances, 90% of test subjects have concluded that human choices are not fully determined by prior events. But because we know that such experiential evidence is consistent with the truth of determinism (as the actual determining causes may be unknown), the trustworthiness of our generally reliable inductive processes is called into doubt in this specific context.

Here we run up against the limits of experimental philosophy as a method for advancing philosophical debates. For the attempt to assess the reliability of the processes responsible for particular moral intuitions inevitably leads back to philosophical debates about the truth or falsity of theses like determinism. These debates cannot to be fully resolved by studying which psychological mechanisms lead ordinary people to reach their conclusions, or which of those conclusions reflects our psychological competencies. For even if those competencies are generally reliable themselves, their reliability must also be established for the particular context at issue. For the same reason, meta-ethical questions remain as wide open as ever. For example, the role of emotion in moral competence has led some philosophers to moral skepticism,[34] and others to belief in universal ethics.[35] Furthermore, even when experimental philosophy studies establish facts about the contours of commonsense theoretical commitments, they do not thereby tell us much about the psychology underlying them. As we saw above with Nichols and Knobe's study, even if an experimental

philosophy study establishes the existence of a folk theory on some subject, it does not follow that the folk theory plays any appreciable role in the psychological competence responsible for judgments about that subject.

I don't think Nichols and Knobe would disagree about any of these limitations. Indeed, they see experimental philosophy as adding "another tool to the philosopher's toolbox," not as an attempt to "do away with any of the methods that have traditionally been used for figuring out whether people's intuitions are truly right or wrong."[36] Like them, I think it's fruitful for experimental philosophers to goad their armchair-bound colleagues into considering where their "intuitions" come from, and how widely shared those intuitions are. Furthermore, although survey studies by no means settle questions about psychological competence, they can help zero in on the variables that lead ordinary people to make or withhold attributions of moral responsibility. As we've seen in the case of Nichols and Knobe's study, the interpretation of the results is far from uncontroversial: I have argued that they are wrong to interpret their results as evidence that compatibilist intuitions are performance errors induced by unreliable psychological mechanisms. But despite its limitations, the exciting thing about experimental philosophy is that it offers hope of traction in debates between compatibilists and incompatibilists. For example, my own argument involves a central claim that would be fruitful to study empirically. Is it true, as I claimed above, that it is part of moral competence to be less accepting of a particular excuse (e.g., "I couldn't help it" or "My decision was caused by my genetics and upbringing") the more serious the moral offense? If so, there is more pressure put on incompatibilists to explain why determinism should be treated as a blanket excuse that applies to all offenses in the same way. In other words, incompatibilists must shoulder the additional burden of explaining why we should depart from our usual ways and means of assessing excuses, rather than taking for granted that the inability to do otherwise is always accepted as a good excuse. Granted, even if I am right about moral competence, no amount of survey data can show that survey respondents are *correct* in their judgments. Such studies will typically at best shift the burden of proof.[37] But that is no small accomplishment in a debate that has been as deadlocked as the one about the compatibility, or lack thereof, between moral responsibility and determinism.

Notes

[1] J. Knobe and S. Nichols, S., "An Experimental Philosophy Manifesto," in Knobe and Nichols, eds. *Experimental Philosophy* (Oxford UP, 2008) 3-14.

[2] S. Nichols and J. Knobe, "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions," *Nous* 41.4 (2007): 663-85. Reprinted in Knobe and Nichols, *Experimental Philosophy*, 105-126.

[3] See E. Nahmias et al., "Is Incompatibilism Intuitive?" *Philosophy and Phenomenological Research* 73 (2006): 28-53. Reprinted in Knobe and Nichols, *Experimental Philosophy*, 81-104. See also R. Woolfolk et al., "Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility," *Cognition* 100 (2006): 283-301. Reprinted in Knobe and Nichols, *Experimental Philosophy*, 61-80.

[4] Nichols and Knobe, "Moral Responsibility and Determinism," 110.

[5] Nichols and Knobe, "Moral Responsibility and Determinism," 111, authors' emphases. Here is the full description test subjects received of Universes A and B:

> Imagine a universe (Universe A) in which everything that happens is completely caused by whatever happened before it. This is true from the very beginning of the universe, so what happened in the beginning of the universe caused what happened next, and so on right up until the present. For example one day John decided to have French fries at lunch. Like everything else, this decision was completely caused by what happened before it. So, if everything in this universe was exactly the same up until John made his decision, then it *had to happen* that John would decide to have French fries.
>
> Now imagine a universe (Universe B) in which *almost* everything that happens is completely caused by whatever happened before it. The one exception is human decision making. For example, one day Mary decided to have French fries at lunch. Since a person's decision in this universe is not completely caused by what happened before it, even if everything in the universe was exactly the same up until Mary made her decision, it *did* ***not*** *have to happen* that Mary would decide to have French fries. She could have decided to have something different.

[6] Nichols and Knobe, "Moral Responsibility and Determinism," 105.

[7] Nichols and Knobe, "Moral Responsibility and Determinism," 106. The passage suggests that emotional response "can" cause performance errors, but Nichols and Knobe do not make clear how frequently they take this to happen. For the purposes of their argument, it needn't happen in every

case, but it must be often enough to call into question the reliability of the resulting compatibilist intuitions.

[8] Nichols and Knobe, "Moral Responsibility and Determinism," 119.

[9] Nichols and Knobe, "Moral Responsibility and Determinism," 119. Nichols and Knobe also discuss the "concrete competence" model, according to which normal competence with responsibility judgments is explained by a dedicated psychological module that is independent of both affect and consciously held theoretical beliefs. Because I agree with them that affect does play a role in normal judgments of responsibility, I leave aside consideration of this model.

[10] Nichols and Knobe, "Moral Responsibility and Determinism," 115.

[11] Nichols and Knobe, "Moral Responsibility and Determinism," 117.

[12] Nichols and Knobe, "Moral Responsibility and Determinism," 118.

[13] S. Nichols, *Sentimental Rules: On the Natural Foundations of Moral Judgment* (Oxford UP, 2004) 10. As the source of the moral/conventional distinction Nichols cites J. Smetana and J. Braeges, "The Development of Toddlers' Moral and Conventional Judgments," *Merrill-Palmer Quarterly* 36 (1990): 329-46.

[14] Nichols, *Sentimental Rules*, 18.

[15] Plus, the moral/conventional distinction Nichols deploys is not uncontroversial. For a critique see J. Prinz, "Is Morality Innate?" in W. Sinnott-Armstrong, ed. *Moral Psychology*, Vol.1 (Oxford UP, 2008) 383ff.

[16] J. Haidt, "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment," *Psychological Review* 108 (2001): 814-34.

[17] Haidt, "The Emotional Dog and Its Rational Tail," 824.

[18] J. Greene et al., "An fMRI Investigation of Emotional Engagement in Moral Judgment," *Science* (2001): 293. Greene actually postulates two systems—a more emotionally "hot" system leaning towards deontological judgments, and a comparatively "cold" system leaning more in the direction of utilitarian judgments. For present purposes, we needn't decide the relative merits of these systems for reaching morally *correct* conclusions. The point is that the "hot" pathway is a central part of moral competence, as shown by the fact that damage to the relevant emotional centers of the brain leads the affected individuals to make very different (more utilitarian) moral judgments than their uninjured peers. See M. Koenigs et al., "Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgements," *Nature* (March 21, 2007).

[19] F. De Waal, *Primates and Philosophers* (Princeton UP, 2006).

[20] R. Joyce, *The Evolution of Morality* (A Bradford Book: The MIT Press, 2006) 117. See also M. Hauser, *Moral Minds: How Nature Designed Our Real Sense of Right and Wrong* (HarperCollins, 2006).

Although I've framed my point in terms of the evolutionary function of moral competence, one needn't believe morality specifically evolved to fill that function in order to agree that it currently does do so. For such a theory see J. Prinz, *The Emotional Construction of Morals* (Oxford UP, 2007).

[21] D. Hume, *An Enquiry Concerning Human Understanding*, ed. Steinberg (Hackett, 1977).

[22] Note that the issue here is not whether people like Bill really *are* morally responsible. The issue is whether those who judge him to be morally responsible have made a performance error. This is a purely psychological issue, which is why the discussion is couched in terms of function rather than truth or falsity.

[23] Knobe and Nichols, "An Experimental Philosophy Manifesto," 8.

[24] Nichols and Knobe, "Moral Responsibility and Determinism," 115-16.

[25] My example involves the disruption of moral competence by an emotion external to that competence (parental love). However, emotions that have a role to play in normal moral competence can also induce performance errors. For example, Prinz claims that disgust plays a role in normal moral competence. In particular, he claims that disgust is involved in "transgressions against the perceived natural order," such as incest (Prinz, *The Emotional Construction of Morals*, 73). But if disgust occurs in the absence of a perceived transgression against nature, it can also disrupt moral competence. For example, Prinz discusses an experiment (T. Wheatley J. Haidt, "Hypnotically Induced Disgust Makes Moral Judgments More Severe," *Psychological Science* 16 [2005]: 780-4) in which test subjects are hypnotized to feel disgust upon hearing a particular word such as "often." As a result, when these test subjects hear of a congressman who *often* takes bribes, they condemn his action more strongly than their unhypnotized peers. Hypnotically induced disgust even leads test subjects to condemn a student who *often* picks interesting topics for class discussions. Prinz counts these judgments as performance errors because the causal role disgust plays in them is not normal: specifically, it is not tied to perceived transgressions of the natural order (Prinz, *The Emotional Construction of Morals*, 96). For my purposes, the important point here is not whether Prinz's particular theory of disgust's role in moral competence is correct, but that the charge of performance error can only be made relative to such a theory. The charge of performance error simply does not turn on facts about the *intensity* of emotional reaction alone. On this point, it's worth noting that the performance errors elicited in Wheatley and Haidt's experiment were due to *low* intensity disgust.

[26] Admittedly, I've cited an ordinary language philosopher rather than an empirical study. As Benson Mates complained long ago in "On the Verification of Statements about Ordinary Language" (in *Ordinary Language*, ed. V. C. Chapell, [Prentice-Hall, 1964]), ordinary language philosophers like Austin announce "what we say when" without bothering to confirm their speculations empirically. However, my point is corroborated by psychologist Mark Alicke's "culpable control model" of

blaming, according to which affective reactions to outcomes intensify blame (M. D. Alicke, "Culpable Control and the Psychology of Blame," *Psychological Bulletin* 126 [2000]: 556-74). Alicke notes that these influences often lead to errors, as when emotional reactions typical of racial bias intensify blame. Nevertheless, as I argued above (and in note 25), the intensity of emotional response is not the issue: whether affective reactions cause performance errors in particular cases can only be judged relative to a theory of affect's functional role in moral competence.

[27] J. L. Austin, "A Plea for Excuses" (1956), reprinted in *J. L. Austin: Philosophical Papers*, 3rd Edition, ed. J. G. Warnock (Oxford UP, 1979) 194-95.

[28] Incidentally, if an error must be attributed, it could just as well be the low affect condition that reflects the error. Indeed, if affect is involved in normal moral judgment, too little affect would predictably cause psychologically unusual patterns of moral judgment. This result has been confirmed in cases of patients with brain lesions that disrupt moral emotions (Koenig et al., "Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgements").

[29] J. Prinz, "Empirical Philosophy and Experimental Philosophy," in Knobe and Nichols, *Experimental Philosophy* (Oxford UP, 2008) 191.

[30] In a forthcoming study, Nahmias and Murray provide specific reasons to be dubious of allegedly incompatibilist intuitions. They found that survey respondents who deny moral responsibility, whether in the abstract or concrete conditions, typically misinterpret determinism to involve "bypassing" of agents' decision-making processes. Epiphenomenalism and fatalism, for example, imply that processes of deliberation and conscious decision-making are causally inert, and that the real causes of our actions bypass those processes. However, determinism does not in fact imply bypassing, and test subjects who interpret it to do so are confused. Obviously, any such confusion tells against the reliability of the process leading to incompatibilist intuitions. Thanks to a referee of this journal for directing my attention to this forthcoming study.

[31] Prinz's point about the limits of introspection also provides an empirical basis for the old Wittgensteinian complaint that philosophical problems arise when we get "in the grip of a picture," extending an oversimplified definition or model to examples in ways that contradict our linguistic competence. The point also dovetails with externalist semantic theories, according to which the explications competent speakers offer of their concepts are likely to be inaccurate and/or incomplete. See T. Burge, "Individualism and the Mental," *Midwest Studies in Philosophy* 4 (1979): 73-121 and "Intellectual Norms and Foundations of Mind," *Journal of Philosophy* 83 (1986): 697-720.

[32] Haidt, "The Emotional Dog and Its Rational Tail," 822.

[33] A. Tversky and D. Kahneman, "Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probabilistic Reasoning," *Psychological Review* 90 (1983): 293-315. The forthcoming study by

Nahmias and Murray discussed above in note 30 postulates a specific error in the reasoning that leads to incompatibilist intuitions—the confusion of determinism with bypassing.

[34] Joyce, *The Evolution of Morality*.

[35] M. Gazzaniga, *The Ethical Brain* (Harper Perrenial, 2005).

[36] Knobe and Nichols, "An Experimental Philosophy Manifesto," 10.

[37] As Nahmias et al. seek to do in "Is Incompatibilism Intuitive?"

Bibliography

Alicke, M. D. "Culpable Control and the Psychology of Blame." *Psychological Bulletin* 126 (2000): 556-74.

Austin, J. L. "A Plea for Excuses." 1956. Reprinted in *J. L. Austin: Philosophical Papers*, 3rd Edition. Ed. J. G. Warnock. Oxford UP, 1979. 175-204.

Burge, T. "Individualism and the Mental." *Midwest Studies in Philosophy* 4 (1979): 73-121.

Burge, T. "Intellectual Norms and Foundations of Mind." *Journal of Philosophy* 83 (1986): 697-720.

De Waal, F. *Primates and Philosophers*. Princeton UP, 2006.

Gazzaniga, M. *The Ethical Brain*. Harper Perrenial, 2005.

Greene, J., R. Sommerville, L. Nystrom, J. Darley, and J. Cohen. "An fMRI Investigation of Emotional Engagement in Moral Judgment." *Science* (2001): 293.

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

Hauser, M. *Moral Minds: How Nature Designed Our Real Sense of Right and Wrong*. HarperCollins, 2006.

Hume, D. *An Enquiry Concerning Human Understanding*. Ed. Steinberg. Hackett, 1977.

Joyce, R. *The Evolution of Morality*. A Bradford Book: The MIT Press, 2006.

Knobe, J. and S. Nichols., eds. *Experimental Philosophy*. Oxford UP, 2008.

Knobe, J. and S. Nichols. "An Experimental Philosophy Manifesto." In Knobe and Nichols, eds. *Experimental Philosophy*. Oxford UP, 2008. 3-14.

Koenigs, M., L. Young, R. Adolphs, D. Tranel, F. Cushman, M. Hauser, and A. Damasio, "Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgements." *Nature* March 21, 2007.

Mates, B. "On the Verification of Statements about Ordinary Language." In *Ordinary Language*. Ed. V. C. Chapell. Prentice-Hall, 1964.

Nahmias, E., S. Morris, T. Nadelhoffer, and J. Turner. "Is Incompatibilism Intuitive?" *Philosophy and Phenomenological Research* 73 (2006): 28-53. Reprinted in Knobe and Nichols, eds. *Experimental Philosophy*. Oxford UP, 2008. 81-104.

Nahmias, E. and D. Murray. "Experimental Philosophy on Free Will: An Error Theory for Incompatibilist Intuitions." Forthcoming in *New Waves for Philosophy of Action*.

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

Nichols, S. and J. Knobe. "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions." *Nous* 41.4 (2007): 663-85. Reprinted in Knobe and Nichols, eds. *Experimental Philosophy*. Oxford UP, 2008. 105-126.

Prinz, J. *The Emotional Construction of Morals*. Oxford UP, 2007.

Prinz, J. "Empirical Philosophy and Experimental Philosophy." In Knobe and Nichols, eds. *Experimental Philosophy*. Oxford UP, 2008. 189-208.

Prinz, J. "Is Morality Innate?" In W. Sinnott-Armstrong, ed. *Moral Psychology*, Vol.1. Oxford UP, 2008. 367-406.

Smetana, J. and J. Breaeges. "The Development of Toddlers' Moral and Conventional Judgments." *Merrill-Palmer Quarterly* 36 (1990): 329-46.

Tversky, A. and D. Kahneman. "Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probabilistic Reasoning." *Psychological Review* 90 (1983): 293-315.

Wheatley, T. and J. Haidt. "Hypnotically Induced Disgust Makes Moral Judgments More Severe." *Psychological Science* 16 (2005): 780-4.

Wittgenstein, L. *Philosophical Investigations*. 3rd Edition. Trans. G. E. M. Anscombe. New York: Macmillan, 1958.

Woolfolk, R., J. Doris, and J. Darley. "Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility." *Cognition* 100 (2006): 283-301. Reprinted in Knobe and Nichols, eds. *Experimental Philosophy*. Oxford UP, 2008. 61-80.