Rejecting the Excuse of Incapacity

Country singer Hank Williams III seeks to excuse his sorry state with the following lyric: "I can't help the way that I am, 'cause the whiskey, weed, and women, had the upper hand." When caught at his drunk and rowdy antics, Hank pleads incapacity: because he is unable to meet the standards of behavior that are being applied to him, he says he should be let off the hook.

Hank doesn't reveal to whom his plea is addressed, but it probably didn't get him very far. He isn't alone in this. Saying "I couldn't help it" works sometimes, but like every excuse, it has its limits. Indeed, I think these limits are more unforgiving than some moral philosophers think they should be. For example, consider the much-discussed "Principle of Alternate Possibilities" (PAP), according to which a person is morally responsible for what he has done only if he could have done otherwise. On this principle, the excuse "I couldn't help it" is valid whenever true.

Some philosophers, such as Harry Frankfurt (1969, 2003), have sought to undermine PAP by offering sophisticated counterexamples. In these debates, both sides presuppose that the excuse of incapacity works the same way no matter the context or seriousness of the accusation. PAP entails that incapacity *always* excuses. On this view, the *only* way to reject the excuse of incapacity is to show that the accused actually could have done otherwise, despite his claim to the contrary. Opponents of PAP exploit its alleged universality across all contexts, constructing fanciful counterexamples to establish the falsity of PAP as a universal generalization.

I'll try a different tack. My goal is to understand when and why the excuse of incapacity fails in ordinary contexts. To be clear, by the excuse of incapacity, I mean any effort to escape

moral responsibility by asserting one's literal inability to comply with others' expectations. This alleged inability may be due to one's settled personality traits, psychological compulsion, genetics, socialization, habit, mental defect, or any other chain of causes that allegedly precludes alternate possibilities. Note that the excuse does not always involve an assertion of psychological defect; in some incarnations, the excuse seeks to cancel moral responsibility by asserting that any normal person subject to the same influences would have acted in the same way. To understand how and why the excuse of incapacity is evaluated in actual practice, I proceed in three steps. First, I introduce examples in which the excuse of incapacity is rejected by reviewing some recent empirical studies that probe lay intuitions about moral responsibility. This evidence shows that ordinary people do not in fact ascribe responsibility as PAP dictates. Second, I corroborate these experimental findings with examples of responsibility attribution in legal cases. Then, by drawing on recent moral psychology, I consider why people assess the excuse of incapacity as they do. Finally, I defend the reasonability of these assessments. Having done all this, I hope to lend credence to Frankfurt's claim that the absence of alternate possibilities is not always relevant to ascriptions of moral responsibility. In other words, rather than refuting PAP, I aim to provide an interpretation of the actual practice of evaluating the excuse of incapacity that rationalizes our tendency to disregard it.

I. Empirical studies of moral responsibility attribution

In this section, I review some recent empirical studies conducted under the banner of "experimental philosophy." Carried out by philosophers seeking a new angle on the venerable debate over free will and determinism, these studies seek to determine what lay intuitions about moral responsibility really are. They show that survey respondents are often quite willing to attribute moral responsibility to individuals whose actions are causally determined. Indeed, there

is some evidence that the more severe the offense, the more likely it is that responsibility will be attributed despite causal determination. Translating these results into the central concern of this paper, these studies show that the excuse of incapacity is not in fact assessed by focusing exclusively on the alleged inability of the accused to do otherwise: the excuse is often rejected even when the inability of the accused to do otherwise is stipulated in the description of the case. I'll seek to rationalize this result later, but the goal of this section is just to establish it.

The free will debate pits "compatibilists" against "incompatibilists." The former believe that free will can exist even if all human action is causally determined. The latter believe that causal determination is incompatible with free will, and consequently cancels moral responsibility. Incompatibilists frequently claim to have common sense on their side, dismissing compatibilist definitions of freedom as revisionist departures from ordinary practice. Eddy Nahmias and his coauthors (2006) decided to test this claim. They reasoned that, if ordinary people are really incompatibilists, the following prediction should be empirically confirmable:

> (P) When presented with a deterministic scenario, most people will judge that agents in such a scenario do not act of their own free will and are not morally responsible for their actions. (ibid. p. 86)

Contrary to (P), most test subjects in Nahmias's studies attributed both free will and moral responsibility to agents in deterministic scenarios. For example, in one scenario test subjects read that all the laws of nature had been discovered and programmed into a supercomputer that predicts all future events with 100% accuracy (ibid. p. 87). The computer applies the laws of nature to predict a bank robbery committed by a man named Jeremy twenty years before his birth. When asked whether Jeremy "acts of his own free will," 76% of respondents answered

affirmatively. 83% judged that Jeremy was "morally blameworthy for robbing the bank." Similar results were obtained when Jeremy was described as performing a positive action (saving a child) or a neutral action (going jogging): 68% judged that Jeremy acted freely when saving the child, while 79% judged that he went jogging of his own free will. In this and other scenarios tested by Nahmias, ordinary people did not take causal determination to prevent free will or cancel moral responsibility.[i] Nahmias concludes that incompatibilists do not have common sense on their side, and must now provide some additional theoretical justification for principles like PAP. For my purposes, a somewhat more modest conclusion suffices. Even if incompatibilists can provide some other form of evidence that their theories are intuitive to ordinary people, Nahmias's results show that incompatibilist principles do not in fact prevent people from attributing moral responsibility.

As it happens, there is in fact survey evidence that ordinary people will agree to statements of incompatibilist theoretical principles. Even this evidence, however, corroborates my claim that agreement with incompatibilist principles does not in fact prevent people from attributing moral responsibility in violation of those principles. Nichols and Knobe conducted a study with the uneasy finding that ordinary people simultaneously embrace abstract incompatibilist principles, and violate those principles in response to particular cases (2007). They presented test subjects with descriptions of a deterministic universe, in which "everything that happens is completely caused by whatever happened before it," including human decisions (ibid. p. 110). Respondents were then split into two groups, the "abstract" condition and the "concrete" condition. In the abstract condition, subjects were asked "is it possible for a person to be fully morally responsible for their actions?" 86% responded negatively, showing a commitment to incompatibilism. However, in the concrete condition, subjects were given a

vignette describing a man named Bill who murdered his family in order to be with his secretary. 72% of respondents agreed that Bill was "fully morally responsible" for his action, despite the fact that his action "had to happen" the way it did.

Seeking to explain these conflicting moral intuitions, Nichols and Knobe hypothesized that a strong emotional reaction to Bill's crime caused respondents to judge his case in violation of their own theoretical commitments about moral responsibility. To test this hypothesis, they ran an additional study in which some respondents were given a "high affect" vignette in which Bill stalks and rapes a stranger, and a "low affect" vignette in which Mark cheats on his taxes. In both cases, test subjects were told (using non-technical language) that these misdeeds occurred in a deterministic universe. 64% of subjects judged Bill to be "fully morally responsible" for his action, but only 23% said the same about Mark. Noting that the latter result is more consistent with incompatibilist intuitions, Nichols and Knobe conclude that the high affect condition manifests a "performance error" on the part of test subjects, whose emotional reaction to Bill's crime causes their normal moral competence to misfire.

Although I disagree with Nichols and Knobe's claim that responses to the high affect condition manifest performance errors, I needn't dispute their *explanation* of compatibilist intuitions at this point.[ii] My goal in this section is to establish the *existence* and indeed typicality of compatibilist intuitions about moral responsibility, and their study helps achieve this goal. In the third section, I will be considering contemporary theories in moral psychology that explain why compatibilist intuitions are typical in this way. For present purposes, however, the point about this study I would like to feature is this: ordinary people are more likely to reject the excuse of incapacity for more serious moral offenses. In this case, although three out of four

people accept Mark's excuse that he couldn't help but cheat on his taxes, only a third proved willing to accept the same excuse for Bill's violent crime.

Nichols and Knobe's study corroborates results obtained by psychologist Mark Alicke. According to Alicke's "culpable control" model of blaming, agents are blamed when they are judged to be in control of their actions. In particular, Alicke maintains that agents are blamed when they are perceived to cause a harmful outcome intentionally with foresight of the result (2000, p. 563). An incompatibilist may find Alicke's model copacetic due to its focus on judgments about control. However, Alicke also notes that the perception of control is not based solely on a dispassionate investigation of the capacities of the accused, but is strongly influenced by emotional reactions:

> Observers who spontaneously evaluate the actor's behavior unfavorably, for example, will exaggerate evidence that establishes her causal or volitional control and de-emphasize exculpatory evidence. Conversely, observers who evaluate the actor's behavior favorably will exaggerate exculpatory evidence and de-emphasize incriminating evidence. (ibid., p. 566)

Alicke's work thus provides additional evidence that ordinary people are less likely to accept excuses, including the excuse of incapacity, the more serious the transgression.[iii] Like Nichols and Knobe, Alicke sees this result as evidence of emotional bias in moral judgment, and he certainly has a point in many cases. For example, negative emotional reactions towards racial minorities increase the likelihood that they will be blamed for offenses, affecting both conviction rates and severity of punishment. While I agree that racial bias distorts judgments of moral responsibility in these cases, I am less convinced that the same can be said for negative reactions to serious consequences. At this point, however, I need not dispute Alicke's *explanation* of his

findings. My goal remains to establish the widespread existence of compatibilist intuitions – to show that moral judgment does not in fact operate consistently with PAP, particularly for more serious offenses. Alicke's work helps establish this point. In section III, I will consider theories of moral psychology according to which emotion-induced departures from PAP are often functionally normal operations of moral competence rather than performance errors, as Alicke and Nichols and Knobe contend.

II. Legal pleas of incapacity

A quick look at the legal system buttresses my claim that the excuse of incapacity does not always turn on the literal ability of the accused to do otherwise. As in the previous section, I continue to construe the excuse of incapacity broadly, rather than as a narrowly circumscribed legal notion. In particular, the excuse needn't involve the claim of a mental defect, but includes, for example, duress, indoctrination, the effects of a bad upbringing, appeals to settled personality traits, or any other defense that rests on the claim that the accused couldn't do otherwise. This broad usage admittedly cuts across traditional legal categories, but the goal here is to corroborate the central claims of the previous section: that the excuse of incapacity is not in fact assessed consistently with PAP, as a straight-forwardly factual question about the ability of the accused to do otherwise; and that the excuse of incapacity is less likely to be accepted for more serious offenses. As in the previous section, my initial goal is not to defend the way pleas of incapacity are assessed, but rather to examine how they are in fact assessed in real cases. I will review several cases in which some form of incapacity or inability to do otherwise is cited as an excuse, but the alleged fact of incapacity is nevertheless not the central focus in the legal determination of guilt.

First, however, a brief aside is in order about my use of legal examples. Philosophers often brush aside legal examples with the remark that legal and moral responsibility are different concepts answering to different standards of assessment. While this is indisputable, it is equally certain that legal and moral responsibility overlap. This is true, I submit, for the cases I'll discuss, involving such unproblematically immoral crimes as murder and robbery. In cases like these, the burden of proof shifts to those who assert a divergence between legal and moral responsibility. In any event, I don't presuppose an inference from legal to moral responsibility. Rather, I regard the following observations about legal practice as corroborating the claim that the standards for assessing the excuse of incapacity that people actually apply (and, indeed, institutionalize) don't follow incompatibilist principles. Together with the evidence cited in the previous section, the legal examples reveal that the excuse of incapacity is not always assessed by focusing on the literal ability of the accused to do otherwise. To challenge this point, a critic would need to give some evidence that ordinary people do in fact judge legal and moral responsibility differently in the cases below, rather than providing a philosophical argument that moral responsibility *should* be assessed differently from legal responsibility. In other words, my main goal in this section, like the previous one, is to establish a claim about the contours of actual practice, not to entertain critiques thereof.

The law recognizes several categories of excuse. When characterizing these categories, lawyers tend to say things that superficially suggest commitment to incompatibilist principles like PAP. Consider the following passage from a criminal law textbook:

> If there is a general explanation of excuses, it is this: our judgment that the actor could not reasonably have been expected to have done otherwise… The current formulations of

> disability excuses typically use the requirements of (1) a recognized disability (2) causing a recognized excusing condition. (Robinson, 2005, p. 739)

For example, the insanity defense requires the identification of a recognized mental illness that causes the defendant to be unaware that his action is wrong.

However, although inability to avoid one's action is sometimes an excuse, it would be hasty to conclude that it is always an excuse. Take duress, for instance. The law excuses criminal conduct committed due to coercion – "a threat of force that 'a person of reasonable firmness in the actor's situation would have been unable to resist'" (ibid., p.741). On this standard, the question is whether the defendant acts as a "reasonable person," not whether his action is compelled by his psychological state. For example, suppose a pain-averse, cowardly person is threatened with a thorough pinching if he does not participate in an identity theft scheme. He cannot establish a duress defense by demonstrating that his past record of pain-aversion and cowardliness caused an "inability to resist," because he would in the process show himself to be unreasonable. Thus, the actual assessment of his plea of incapacity does not turn on his allegedly compulsive pain aversion, but rather on the reasonability of being so pain averse.

Now consider a real case. Heiress Patty Hearst was kidnapped by the Symbionese Liberation Army, a far-left militant group. After weeks of captivity in a closet during which she was repeatedly raped and browbeaten, Hearst became sympathetic to the cause of her captors. As Hearst increasingly identified with the SLA, she participated in an armed bank robbery and other criminal acts. At trial, the defense argued that she was not responsible for her actions because she had been brainwashed by her captors. Although Hearst's sentence was later commuted due to her family connections, her excuse did not hold up at trial and she was convicted of bank robbery.

For short of an insanity defense,[iv] the law does not excuse people for acting on their beliefs and values, no matter how involuntarily those beliefs and values were formed. The same legal textbook quoted earlier makes this point clearly, but with some trepidation:

> While the brainwashed actor presently suffers neither coercion nor mental illness, the actor's choices nonetheless have been highly influenced by others, and against his will, through the coercive indoctrination that the actor experienced. For that reason, it seems awkward to hold such an actor accountable for conduct in the same way as someone who has (more) freely chosen or developed his own beliefs and values. But the law presumes that each person is accountable for one's own personality, including one's beliefs and values. There simply is no mechanism under current doctrine for rebutting this presumption. (ibid., p.749)

Here again, we see that the actual standards for assessing the excuse of incapacity do not center on the question of whether the defendant could have done otherwise. Just as some forms of evidence, such as hearsay, are inadmissible at trial, some forms of alleged causal determination are excluded from consideration.

It may be suspected that this fact about the legal system reflects some peculiarly legal institutional pressure that is irrelevant to moral responsibility. But the standard prevents the exculpation of the kinds of criminals that tend to draw the strongest condemnation from the public, both legally and morally. If there were a mechanism for rebutting the presumption of responsibility for one's beliefs and values, not only sympathetic figures like Hearst, but gang members, radicalized terrorists, and others who can rightly claim a history of involuntary indoctrination would have access to the same mechanism. However, membership in a gang or a

terrorist cell, far from serving as an excuse, is actually an aggravating factor.[v] And it remains an aggravating factor even if we find, as happens often enough, that the defendant's indoctrination took place when he was a child. Children are no more at liberty to avoid indoctrination than Patty Hearst was. If anything, children have a better excuse for not resisting indoctrination because they lack any counter-balancing life experience. Nevertheless, the law does not allow pleas of incapacity based on the details of one's upbringing.[vi] Ordinary assessments of moral responsibility are similar, as the sketch of Alicke's research in the previous section shows: a long history of involvement with negatively perceived groups such as gangs does not support an excuse of incapacity, but instead intensifies moral blame.

The law also refuses to excuse actions rooted in settled personality traits, regardless of how those traits may have originated. For example, consider *Kansas v. Borman* (1988).[vii] Borman admitted to killing his girlfriend in a rage after she went on a date with another man. But he sought to avoid criminal responsibility by claiming diminished capacity: given his record of poor impulse control and explosive anger, he claimed that he couldn't control his murderous reaction to his girlfriend's infidelity. His defense thus rested on the assertion that no one with his particular personality traits could have done otherwise. However, the judge's instructions to the jury made clear that they were not to assess Borman's excuse as his lawyer suggested. In the law, diminished capacity requires more than causal necessitation by one's character or personality:

> The criminal law concept of diminished capacity requires the presence of a mental disease or defect not amounting to legal insanity which a jury may consider in determining whether the defendant has the specific intent required for the crime charged. Mere personality characteristics such as poor impulse control, a short temper, frustration, feelings of dependency, "snapping," lack of concern for the rights of other people, etc., do

> not constitute a mental disease or defect bringing the doctrine of diminished capacity into play. (ibid.)

Borman's excuse was rejected, and he was convicted of second degree murder (intentional, but unpremeditated murder). Borman's case is typical. The law doesn't ask what your particular personality causes you to do, but rather what you did, whether you intended to do it, and whether your action was reasonable.

When the law does take incapacity into account, it's often as a mitigating factor in sentencing. Even here, however, the effectiveness of the excuse often varies with the seriousness of the crime. For example, consider the treatment of juveniles. Due to their immaturity, juveniles are typically tried separately from adults. Recent evidence suggests that the brains of minors are not fully developed until the early twenties, leaving them with poor impulse control and underdeveloped empathic responses (Ortiz, 2004). Although the courts do not treat these neurological factors as exculpatory, they do sometimes serve as a mitigating factor in sentencing. For example, the Supreme Court recently ruled it unconstitutional to impose the death penalty on minors.[viii] However, in actual practice immaturity does not always achieve even this mitigating role. Indeed, juvenile offenders are far more likely to be tried as adults for serious offenses such as murder. For example, in the state of Florida, juveniles over the age of 14 who are charged with a violent crime are automatically tried as adults if they have a prior conviction for violent crime, including "murder, sexual battery, armed or strong-armed robbery, carjacking, home-invasion robbery, aggravated battery, aggravated assault, or burglary with an assault or battery."[ix] Repeat offenses involving firearms draw similar treatment. Here we see that the seriousness of the offense automatically exposes juveniles to the more severe penalties typical of

the adult penal system, despite their immaturity. This reinforces the conclusion that the excuse of incapacity is less likely to be accepted for more serious offenses.

The foregoing observations about the legal system dovetail with the findings of experimental philosophers: the excuse of incapacity does not always cancel or even mitigate responsibility, even when the truth of the accused's plea of incapacity is substantially granted. Second, the excuse is less likely to be accepted for more serious transgressions. The defensibility of these practices is of course contestable, but so far I am content merely to provide examples of how the excuse of incapacity fails to cancel the attribution of responsibility in actual practice. In the next section, I'll turn to consider why the excuse is not as effective as incompatibilist principles suggest it should be.

III. Moral psychology

We've seen that the excuse of incapacity has its limits, especially for serious offenses. Compatibilists and incompatibilists alike owe an explanation of this fact. For example, we saw above that Nichols and Knobe explain failures to abide by PAP as emotion-induced performance errors. In contrast, I will argue that compatibilist assessments of responsibility are just what recent theories in moral psychology and evolutionary biology would lead us to expect from our normal moral competence. Of course, purely causal psychological considerations don't show that our normal assessments of responsibility are justified. However, by reflecting on why we are the way we are, we have a fresh perspective from which to ask whether we should want to change. I'll postpone that question until the next two sections.

We've already seen that emotional reactions to serious consequences make people more likely to blame and prone to inflict more severe punishments (Alicke, 2000).[x] The work of

psychologist Jonathan Haidt contributes to an explanation of this fact (Haidt, 2001). According to his "social intuitionist" model of moral judgment, moral judgments are typically caused by emotional reactions rather than moral reasoning.[xi] Consequently, for excuses like incapacity to change minds, they must overcome the initial emotional responses provoked by perceived transgressions. This will be harder to do when the initial emotional reaction is intense, as it tends to be for more serious transgressions.

To see why, consider Haidt's theory of the role of moral reasoning. Subjectively, we often take our moral reasoning to be the cause of our moral judgments. Haidt calls this the "wag the dog" illusion: "We believe that our own moral judgment (the dog) is driven by our own moral reasoning (the tail)" (ibid., p. 823). Instead, Haidt argues, moral reasoning provides *post hoc* rationalizations of the judgments already induced by emotional reactions. In other words, we make our judgments first, then (if challenged) search for reasons to justify them. For example, consider how Alicke's culpable control model of blaming looks from the perspective of Haidt's theory: the tendency of emotional reactions to intensify the perception of control is just what we should expect, because emotional reactions cause judgments of moral responsibility which in turn cause the search for a *post hoc* rationalizations. The judgment of culpable control is thus an effect of moral judgment, not its cause. That is not to say that moral reasoning is merely epiphenomenonal: rather, it serves an inter-personal function – to influence other people. But moral debate, on Haidt's theory, isn't just the cool weighing of reasons. If it were, moral argumentation would change minds much more frequently than it does (an expectation Haidt dubs the "wag-the-other-dog's-tail illusion"). Through his studies of "moral dumbfounding," Haidt found that people ordinarily stick to their moral judgments even after the justifications they are able to formulate for those judgments are undermined. According to Haidt, moral reasoning

changes minds, when it does, by inducing new emotional responses in the other person. For example, when focusing on an injury, one's immediate emotional reaction is likely to be negative and prone to blaming. However, this reaction may be changed by moral reasoning that leads one to focus on more positively perceived factors such as the good intentions of the actor. Thus, on this model, for an excuse such as incapacity to succeed, it must reframe the perceived offense in a more emotionally favorable light. More serious offenses will be harder to reframe due to the intensity of the negative emotional reaction they generate.

Such references to the causal role of emotion raise the suspicion of bias, as we've seen in the previous discussion of Alicke and Nichols and Knobe. However, in order to assess whether a charge of bias is well-placed, it's helpful to consider the functional role of emotion in moral judgments. To that end, I'll now consider some recent evolutionary theories of the origin of moral competence. These theories help explain why it is that we have an emotionally-based disposition to be less forgiving of more serious offenses, and to so often fail to abide by PAP.

From an evolutionary perspective, morality functions to promote mutually beneficial cooperation or, as the biologists call it, reciprocal altruism. Evolutionary theorists have been at considerable pains to explain how reciprocal altruism could evolve, given that cooperation often imposes at least short-term costs on the individual. However, short-term costs can be out-weighed by the long-term benefits of mutual cooperation. The trick is to devise a system that is truly reciprocal, promoting cooperation while preventing exploitation by cheaters. In terms of evolutionary game theory, what's needed is an "evolutionarily stable strategy" (ESS) for achieving relationships of reciprocal altruism. An ESS is a strategy that can persist in a population amidst competition with other strategies. For example, unconditional helping is not an ESS because unconditional helpers are exploited by those deploying non-reciprocating strategies.

However, conditional helping is an ESS (Axelrod, 1984): the strategy named "tit for tat" takes cooperation as the initial move, and continues cooperation as long as it is reciprocated. Cheaters are met with non-reciprocation, although the cheater can renew a cooperative relationship by being the first to initiate cooperation.

Tit for tat was developed in computer simulated Prisoner's Dilemma tournaments, which are highly simplified compared to real social interactions. However, evolutionary psychologists Leda Cosmides and John Tooby have argued that a version of tit for tat governs human social interactions (Cosmides and Tooby, 1989, 2008). They postulate innate, modular psychological mechanisms for "cheater-detection" and reasoning about social exchange. Although there is considerable disagreement about their theory of the psychological mechanisms involved – particularly regarding modularity and innateness (see, for example, Prinz (2008)) – much less controversy attends their claim that normal human beings are psychologically disposed to reciprocate with reciprocators, while shunning or punishing cheaters. Indeed, game-theoretic considerations give us reason to believe that some version of tit for tat must be implemented psychologically to explain the emergence of reciprocal altruism as an ESS. In any event, if it is true that some version of tit for tat is incorporated in normal moral competence, it considerably clarifies some of the facts about moral practice discussed earlier.

Take, for example, the legal standard illustrated by the Patty Hearst case, according to which defendants are accountable for acting on their own beliefs and values even if they were indoctrinated involuntarily. Tit for tat helps explain both why Hearst was punished, and why many feel a reluctance to blame her despite a willingness – indeed, a heightened willingness – to blame other indoctrinated offenders such as gang members. Tit for tat keys on two main factors: episodes of uncooperative behavior, and subsequent indications of cooperative behavior. Hearst

predictably triggers blame by robbing a bank, but her subsequent "reprogramming" (which was, incidentally, as involuntary as her first round of brainwashing) and renunciation of the SLA show her to be once again a fit partner for cooperative relations. A gang member, in contrast, is less likely to be perceived as a rehabilitated cooperator, and will accordingly continue to be excluded from cooperative exchange no matter how involuntarily his gang-land beliefs and values were acquired. In short, what matters psychologically in this context is not whether the beliefs and values an offender acts upon were acquired voluntarily or involuntarily, but the prospect that she can be counted on as a future partner in cooperative exchange.

Furthermore, if some version of tit for tat is part of our moral competence, it should be no surprise that recidivism intensifies assessments of responsibility. If anything, recidivism is additional evidence of incapacity to change. Therefore, if incompatibilist principles such as PAP were really driving us psychologically, we should expect recidivism to make us more open to the excuse of incapacity. But the reverse is the case. Recidivism intensifies rather than mitigates attributions of responsibility, canceling excuses such as immaturity that effectively mitigate a first offense. This is entirely predictable if our moral psychology incorporates some version of tit for tat. Recidivism reinforces the expectation that the accused cannot be counted on as a future partner in cooperative exchange. It consequently elicits the usual responses of moral disapproval, ranging from cold shoulders to institutionalized standards like Florida's rule for trying repeatedly violent minors as adults.

For similar reasons, a moral psychology incorporating tit for tat also helps explain why we would be disposed to adopt the legal standard illustrated by the Borman case, according to which actions flowing from settled personality traits such as a short temper do not support a plea of diminished capacity. This makes sense psychologically if we take the judgment to turn on the

prospects of future cooperation: Borman's claim that he is susceptible to explosive anger predictably does not cancel or even mitigate blame because it provides reason to expect continued failures to engage in cooperative relations. Psychologically speaking, the allegedly involuntary origin of Borman's explosive anger is less of a factor in the assessment of his responsibility than perceptions of his capacity for cooperative conduct.

The above considerations about tit for tat gain additional support from experimental philosophy findings. A study by Woolfolk and his colleagues (2006) went further than establishing the existence of compatibilist intuitions about particular cases, showing that an agent's identification with his action elicits attribution of moral responsibility even in conditions of complete causal determination. The experiment involved a vignette about a man forced by terrorists to execute a friend who, he has just learned, had an affair with his wife. Different versions of the vignette altered the degree of constraint, and the degree to which the man identifies or willingly goes along with the terrorists' orders. The study found that respondents were much more likely to attribute responsibility in the high identification condition than the low identification condition, even when the degree of constraint was described as absolutely irresistible. This makes sense if moral competence keys on prospects for reciprocal altruism because the causally constrained agent nevertheless shows himself to be a willing "cheater" of moral rules.

Taken together, the above considerations suggest that, psychologically speaking, the plea of incapacity is not in fact assessed as a factual question about the degree of causal determination brought to bear on the accused. Rather, in cases like Hearst's, Borman's and the vignettes in Woolfolk's study, the excuse of incapacity succeeds to the extent that it cancels the normal expectation that the agent identifies with her own action and is likely to continue acting

uncooperatively. [xii] Psychologically speaking, we're interested in who we can count on as partners in mutually beneficial social relations, not in how they got that way.

IV. Should we want to change?

I've spent a lot of time detailing the way we actually assess the plea of incapacity. As Nahmias argues (2006), these facts about actual patterns of moral judgment challenge incompatibilist claims that PAP is a common sense platitude, shifting the burden of proof onto the incompatibilist to provide a more robust theoretical justification of her principles. The points about our legal system adduced in section II add weight to Nahmias's burden-shifting argument. The law reflects our culture's most systematic effort to think about responsibility. As such, the fact that legal proceedings do not follow incompatibilist principles puts further pressure on the incompatibilist to defend her principles independently of common sense. Although I agree with this burden-shifting argument, there is still more to say about why and how ordinary practice makes sense. For no amount of describing our actual practices shows by itself that we ought to continue them. For example, it is a widespread practice, which evidently comes quite naturally psychologically, to more readily blame and more severely punish members of racial minorities (Alicke 2000). When we learn of such unfortunate dispositions, the obvious response is not to acquiesce in them as natural or inevitable, but to devise strategies for counteracting them. The question before us, then, is whether our actual practice of assessing the excuse of incapacity should be changed, or is instead worthy of endorsing.

To focus the discussion, I will concentrate on a feature of ordinary practice established in sections I and II -- the tendency to be less accepting of the excuse of incapacity for more serious offenses. I will begin by sketching a moral justification of this practice. In this, I follow the

example of R. Jay Wallace, who argues that the debate between compatibilists and incompatibilists arises from within ordinary moral practice, as a question concerning the *fairness* of blaming people who lack alternate possibilities (Wallace, 1994, p. 109).

So how could it be fair to be less accepting of the excuse of incapacity when faced with a more serious offense? No doubt the question is too clumsy as posed. I have no intention of claiming that it is *always* fair to reject a plea of incapacity for serious offenses. For example, in cases of duress, it is sometimes excusable to impose significant harms in order to avoid even worse harms. And there are certainly varieties of incapacity, such as insanity, that morally excuse any offense by disqualifying the perpetrator as a genuine moral agent. However, there remain an interesting range of cases in which an excuse that may have been accepted for a less serious offense will not be accepted for a more serious one. As J.L. Austin notes, inadvertence may excuse stepping on a snail – but not on a baby (Austin, 1956-7, p. 194-5). He goes on to explain this fact by appeal to "standards." To suggest that inadvertence excuses such a serious transgression is to endorse an indefensible standard of reasonable care toward babies, whose extreme vulnerability renders heightened carefulness necessary.

I suggest that the same point about standards justifies the difference between many cases in which we are, and cases in which we are not, prepared to accept incapacity as an excuse. For example, consider Borman, who sought to excuse his violence by claiming he literally could not control his temper. Had Borman refused to return his girlfriend's calls or smashed his own dishes, his excuse would probably suffice. But he killed her. That is among the actions for which there is "no excuse," as we say. But why not? Because standards of fairness do not focus on the accused alone: they involve both the agent and those affected by him. It may be unfortunate that someone has an explosive temper through no fault of his own, but it's more fair that he should

bear the burden of incarceration than that others should suffer further fits of violence. Fairness, in practice, involves weighing the rights and interests of all concerned parties. If you are so temperamental that you can't engage in romantic relationships without violence, then you shouldn't engage in romantic relationships because potential partners have a stronger right to safety than you do to companionship. Comparisons of value such as these drive the evaluation of excuses like incapacity, not assessments of whether the accused possessed alternate possibilities.

Borman, to his discredit, was at least partially at fault for his own incapacity. Knowing of his explosive history, he could have taken anger management classes, sought medication, or joined a monastery. Aristotle famously uses this point to explain how people can be responsible for actions that flow from settled traits of character: we're all responsible for shaping our own character, so prior choices render us responsible for later actions that flow from our character (Aristotle, Nic Ethics III.5, 1114.b22). But Aristotle's explanation doesn't cover all cases. Take Patty Hearst. She was held responsible for bank robbery despite the fact that she bore no responsibility for the coercive indoctrination that ultimately led to her radicalization. Hearst's case is admittedly controversial. But as I noted above, she cannot be excused morally or legally unless childhood indoctrination of the sort suffered by many gang members and other paradigm offenders is also counted as a legitimate excuse.

For example, we may understand why a child indoctrinated in racial bigotry becomes a bigot, but his lack of responsibility for his upbringing does not excuse him from responsibility for acting on his prejudices by, say, firing racial minorities at his workplace. Here again, moral assessment focuses on interpersonal standards of fairness, not the vicissitudes of the indoctrinated bigot's personal history. Regardless of how his prejudiced beliefs and values were formed, he is blameworthy for acting on them because the victims of his discriminatory action do

not deserve the treatment they receive. For the bigot to say that he cannot help being a bigot is not an excuse, but an irrelevancy.

Admittedly, this claim about the irrelevance of incapacity to cases like the indoctrinated bigot's is the very point at issue between compatibilists and incompatibilists. According to the incompatibilist principle PAP, an inability to do otherwise is *always* relevant – indeed, decisively so – to moral debate. I'll now challenge this claim about relevance by considering how a principle like PAP applies to claims made within moral debate itself. By doing so, I aim to provide a *reductio ad absurdum* argument against the assumption that incapacity is always relevant within moral debate.

Luckily, for my purposes, I needn't settle the vexed question whether particular claims to incapacity are *true*. Rather, the question is whether such claims are always in order: *if* true, would certain claims to incapacity be relevant within moral debate?

Suppose, after catching the indoctrinated bigot at his discriminatory employment practices, we punish him. Punishment deserves scrutiny for its moral justification, and those who suffer it are often the first to voice objections. Because his plea of incapacity was rejected, the indoctrinated bigot may claim that he is being treated unfairly. Suppose that we have been reading up on the role of emotion in moral psychology and the evolutionary origins of moral judgment, and have reached the conclusion (rightly or wrongly – it doesn't matter for present purposes) that our moral judgments, and the sanctions we impose, are causally determined by our innate psychological mechanisms and personal history. In short, we are convinced that we literally cannot do otherwise than punish the bigot. So we say: "We can't help it, either. Our censorious attitudes towards you are rooted in involuntary emotional responses that we can't

control, so we can't do otherwise than inflict your punishment." If PAP is true, this rejoinder is relevant to the bigot's claim that we are wrong to punish him. However, such a comment, even if true, obviously fails miserably as a moral justification. Indeed, it's not even relevant. It marks a refusal to continue discussing the case in moral terms. Pragmatically, the message to the accused of the comment "We can't help it, either" is that his excuse won't be accepted. But the *reason* it won't be accepted is the standard that he violated – the rule against discrimination. Our psychological characteristics as judges of the situation are not directly relevant unless they show that we have failed to abide by that rule due to prejudice, inattention to relevant details, or the like. Even here, however, the issue is our consistency in applying the rule and the justifiability of the rule itself, not whether it's true that we "can't help" but apply it.

Incompatibilists may object that they were all along trying to question the justification of the standards applied in ordinary cases. But I think they miss the point. After all, if they are right, so long as a bigot's attitudes are so deeply entrenched that he cannot resist them, he cannot be blamed for adopting his bigoted standards. But a digression into his childhood history is simply irrelevant to a moral debate about the beliefs and values he acts upon. Debate about moral standards is and should be focused on the standards themselves, not the capacity of the participants in the debate to resist them.

V. Post hoc rationalization?

In the last two sections, I've sought to explain why we assess the excuse of incapacity as we do, both by applying standards of fairness internal to moral practice, and from an external psychological perspective. But at this point, it may be objected that the facts about moral psychology revealed in section III should undermine our confidence in the moral reasoning

presented in section IV. In particular, if Haidt is right that moral reasoning just offers *post hoc* rationalizations of emotion-induced judgments, isn't our tendency to focus on standards within moral debate based on false preconceptions about the psychology of moral reasoning? How can we trust moral reasoning at all if it's just *post hoc* rationalization?[xiii]

To dramatize the worry, consider how the arguments of the previous section look from the standpoint of Haidt's moral psychology. For example, take my discussion of Borman's murder. On Haidt's model, my moral condemnation of Borman stems from a negative emotional reaction. I then "rationalize" this reaction by emphasizing the harm done to his girlfriend. The main purpose of such moral argument is interpersonal – an effort to reach consensus in moral judgment with others. The people to whom I address my argument will have made their moral judgments on an emotional basis as well, so my moral argument will fail to convince them unless it also causes them to share my negative emotional reaction to Borman.[xiv] My appeals to standards of fairness are thus attempts to elicit similar judgments in others by drawing attention to the suffering of Borman's victim and equating it with other cases to which my audience is (if the argument is to succeed) already disposed to react as I do. On this model of moral psychology, appeals to moral standards are not a purely rationalistic exercise, drawing only on our abstract cognitive capacities, but rather efforts to engage the emotional dispositions of those with whom we enter into moral argument.

Once we learn that moral reasoning is *post hoc* rationalization, it may seem that we can no longer trust our own moral reasoning – including the moral arguments advanced in section IV. For "rationalization" involves glossing over emotional motivations with reasoning that wouldn't be found acceptable if it weren't for its emotional impetus. And that's exactly what Haidt's theory has us doing when we engage in moral reasoning. Doesn't that show the whole

enterprise to be biased and irrational? How can we understand the causes of our moral practices on a psychological level, including especially the heavy role of emotion, without losing confidence in their rationality?

The first step to reconciliation here is to acknowledge that armchair auto-psychology encourages flawed preconceptions about the causal processes underlying our rational competence. Rationality is, after all, a cognitive achievement of human beings. So we should expect that some causal explanation will be forthcoming of the psychological processes underlying rational competence. Furthermore, we should be prepared for these psychological explanations to depart substantially from common-sense speculation about how rationality operates. The proper response to these surprises is not to declare that we aren't really rational, but rather to correct our preconceptions about how rational competence really works. In particular, despite the long-standing preconception that emotion invariably interferes with or biases reason, we should not declare ourselves irrational upon discovering that emotions play a necessary role in normal rational competence.

Such a discovery has already been made. Antonio Damasio famously studies patients with injuries to the ventro-medial pre-frontal cortex, a part of the brain necessary for emotions such as shame, guilt, and embarrassment (Damasio, 1994, 2003). Lesions in this area leave mathematical ability, IQ scores, and other abstract reasoning abilities intact. But in the absence of emotional cues, these abstract cognitive capacities alone do not enable the unfortunate victims of brain injury to behave rationally. For example, Damasio's patients fail miserably at such paradigmatically rational tasks as managing their finances, and are basically incapable of living on their own. Damasio concludes that emotions play a crucial role in normal rational

competence, and that the long tradition of starkly distinguishing reason from emotion is mistaken.

That is not to say that explicit appeals to emotion have a new privileged place *within* rational argumentation. To say that emotions are causally necessary to rational competence is not to say that discussion of the emotions involved is relevant to us as we participate in rational argumentation. In other words, the process of weighing reasons may involve emotional underpinnings as a psychological matter of fact, but that doesn't mean that those very emotions can therefore be cited *as* reasons. For example, Damasio did not prove that financial reasoning is about feelings instead of finances. Rather, emotions play a critical causal role in the normal psychological processes underlying financial reasoning – reasoning that does not typically include emotions themselves as part of its subject matter. If a particular investment is unwise due to its high degree of risk, good financial reasoning will focus on the risk itself as the basis for avoiding the investment, not the negative emotional reaction that psychologically enables one to heed that risk.

Furthermore, to say that emotions play an important causal role in rational competence is not to deny that emotion can also disrupt that competence. Gnawing envy of your more affluent neighbors may well interfere with sound financial judgment. In short, some emotional reactions bias financial reasoning, despite the fact that others are essential to its proper operation. Failure to appreciate these points contributes to the exaggerated fear that any psychological process involving emotion as part of its normal operation is therefore irrational or hopelessly biased.[xv] But the fear is misplaced: no amount of cataloging emotionally induced performance errors can make it the case that good financial reasoning proceeds independently of emotion, nor render the functionally essential contribution of emotion somehow dysfunctional.

Now return to moral competence: emotions are essential to its operation, but that means neither that all emotional reactions facilitate sound moral judgment, nor that explicit appeals to emotion always have a place within moral reasoning. Take the second point first. As I argued above, moral reasoning appeals explicitly to moral standards, not to one's psychological dispositions to apply those standards. In other words, although emotions play a crucial causal role in the normal psychological processing that underlies appeals to moral standards within moral debate, explicit appeals to those very emotions are not for that reason relevant within moral debate.[xvi] You won't succeed in a moral argument by citing the degree of your indignation, even if an emotional reaction of indignation is causally necessary both for your own moral judgment and any change in moral judgment you hope to elicit from those you are trying to convince. Nor can you change another's judgment by pointing out that he wouldn't make his moral judgment if it weren't for his indignant emotional reaction. Even if true, such a point of fact isn't relevant to the evaluation of moral reasoning. Only appeals to moral standards shared by the interlocutor, calling to mind relevantly similar cases, or other means of causing (rather than mentioning) indignation will lead to the desired moral judgment. For example, violations of standards of fairness predictably cause indignation in those who value fairness, but it is the violation itself, not the indignation caused, that is the focus of moral debate. After all, we routinely discount the indignation of others when we think it issues from partiality or bias rather than a commitment to fairness or some other moral standard.

Admittedly, it is possible to cause indignation by fanning prejudices, prodding fears, or otherwise exploiting the emotions of those one is trying to convince. In short, moral judgment can be manipulated. This is no surprise – any theory of moral judgment that denied this possibility would thereby be inadequate. But the possibility of manipulation by emotional

appeals or skillful rhetoric does not dissolve the distinction between good moral judgment and biased moral judgment. This reflects the other important lesson from the analogy to financial reasoning: although emotions can bias judgment, they are also essential to normal competence. Accordingly, we can acknowledge that emotions *often* bias moral judgment without concluding that *all* emotional influence is proof of bias. Psychologically, the distinction between biased and unbiased moral judgment will advert to a theory of normal function. Relative to such a theory, the emotional reactions that facilitate moral competence can be distinguished from dysfunctional emotional reactions that cause performance errors. In this respect, emotions are no different from anything else: even so fundamental a nutrient as water can be poisonous in the wrong amounts or in the wrong places, but no one would be tempted to conclude that any physiological process involving water is therefore dysfunctional.

But consider the question morally, not psychologically. How do we tell the difference between biased and unbiased moral judgment? My central example of biased moral judgment has been the tendency to blame racial minorities more readily, and to subject them to more severe penalties. Compare this biased judgment to the unbiased judgment that such race-based discrimination is morally wrong. What is the difference? As Alicke notes, negative emotional reactions are responsible for the tendency to more readily blame, and more severely punish, minorities. But the upshot of recent discoveries in moral psychology is that our moral disapproval of racial discrimination, like our moral disapproval of any moral wrong, also requires a negative emotional response. So the morally relevant difference between the cases is not whether a negative emotional response is involved: it always is and must be. The moral question is what prompts the negative emotional reaction, and whether the ensuing moral judgment is defensible in moral terms. Was one's judgment prompted by the violation of a sound

moral standard like the rule against discrimination? Or was it prompted by some morally indefensible motive such as antipathy to racial minorities? Which moral standards are "sound," and which motives "indefensible," is of course a moral question that is regularly contested within moral debate. But the important point here is that the distinction is not erased by the acknowledgment that emotional reactions are causally involved in moral judgment.

I have endeavored to blunt the worry, that the role of emotion in moral judgment renders all moral reasoning unjustified rationalization, by showing the worry to rest on the unwarranted assumption that the mere presence of emotion is evidence of bias or irrationality. In actual practice, we reject motives as biased when they cannot be defended by appeal to moral standards we accept. "Rationalization," in the pejorative sense of the term, applies to moral judgments that are biased in this sense: such judgments are motivated by self-interest, envy, prejudice, or other non-moral emotional reactions. Judgments that survive moral scrutiny by being grounded in moral principles are not "rationalizations" in this pejorative sense, even if they are "rationalizations" in the purely psychological sense that they involve judgments that wouldn't be made if it weren't for their emotional impetus. A similar point goes for the pejorative sense of rationalization in other domains. For example, rationalization in financial reasoning consists in accepting a conclusion that cannot be defended by reference to sound financial principles, but is instead motivated by envy, over-confidence, wishful thinking, etc. The fact that emotion is causally involved in all financial reasoning does not erase the distinction between good financial reasoning and unsupportable financial rationalizations.

The foregoing distinction between pejorative and innocent senses of rationalization does not of course provide a "rational defense" of moral reasoning, in the sense of a justification external to moral debate that shows the practice as a whole to meet some preconceived standard

of rationality (e.g., a standard that would be acceptable to some imagined rational being that doesn't share our emotional make-up). But I don't think this should be troubling. Where, after all, are we to find such standards of rationality? We've seen that it's unwise to declare in advance that the absence of emotion is the standard of rationality, given the role of emotion in normal rational competence. The fact that emotions causally support our concern for fairness means, among other things, that brain damage to the ventro-medial pre-frontal cortex will extinguish that concern. But this observation does not, and I submit should not, impair our commitment to fairness. Indeed, one lesson of the previous section is that such facts about moral psychology, considered as moves within moral debate, are irrelevant. From within moral debate, an accusation of bias cannot be sustained on the grounds that it is motivated by a concern, emotional or otherwise, for fairness.

VI. Conclusion

I'll close by comparing my approach to that of Strawson, who famously argued that our practices of assessing moral responsibility involve a susceptibility to inter-personal emotional reactions ("reactive attitudes") that we cannot, and should not want to, give up:

> This commitment [to ordinary inter-personal attitudes] is part of the general framework of human life, not something that can come up for review as particular cases can come up for review within this general framework. And…if we could imagine what we cannot have, viz. a choice in this matter, then we could choose rationally only in the light of an assessment of the gains and losses to human life, its enrichment or impoverishment; and the truth or falsity of a general thesis of determinism would not bear on the rationality of *this* choice." (Strawson, 1982, p. 70)

Strawson's argument distinguishes between the "participant" attitude, which involves susceptibility to reactive attitudes, and the "objective" attitude, which views the world without emotional attachment. When adopting the objective attitude, causal explanation predominates, and moral issues simply don't arise. In Strawson's view, it wouldn't be rational to always view the world objectively, because we would thereby lose valuable aspects of human life, including our moral practices. I agree with this point, but I want to clarify what I take to be the relationship between the objective and participant attitudes.

A shallow reading of Strawson may see the participant attitude as completely insulated from objective inquiries. However, objective inquiries can both help us understand the practices in which we participate, and alert us to our failures to abide by the standards of those practices. For example, we've seen that emotional reactions based on racial prejudice can lead to lower standards of evidence for blame, and more severe penalties. This objective discovery should lead us to be wary of our normal "participant attitude" judgments about minorities. However, the reason we should be disturbed by this discovery is that it shows us to violate our own fundamental moral commitment to fairness. Here we see objective inquiry helping us be better participants in our own moral practices, warning us of a moral blind spot.

Objective inquiry can also help us better understand what we are doing when we uphold our own standards. For example, Haidt's work provides empirical support for Strawson's view that reactive attitudes are central to ordinary moral practice. Were it not for our capacity to feel the injustice of discrimination, for example, we would be unlikely to care about standards of fairness that preclude it. Knowing the role these emotions play in the maintenance of valuable ends such as social cooperation, we can gladly accept and indeed encourage our own susceptibility to such emotional reactions.

Our usual standards for assessing the excuse of incapacity can be endorsed for similar reasons. The facts about moral psychology reviewed in section III help us understand why we assess the excuse as we do. By itself, these facts do not distinguish our failure to abide by incompatibilist principles from our susceptibility to prejudice. That's why it's also necessary to consider the way these practices are rooted in standards of fairness. Even here, objective inquiries, like those of the experimental philosophers, can help us identify with more confidence which moral principles are really driving us. In the end, the best result we can hope for is to find that the objective description of our practices has us doing things, such as focusing on fairness (or, as the evolutionary game theorists call it, reciprocity), that we can endorse as participants. I find such a happy synergy in our compatibilist practices of assessing the excuse of incapacity.

References

Alicke, M. (2000). Culpable Control and the Psychology of Blame. *Psychological Bulletin*, 126(4), 556-574.

Appiah, K. A. (2008). *Experiments in Ethics*. Cambridge, MA: Harvard University Press.

Austin, J.L. (1956-7). A Plea for Excuses. In J.G. Warnock (ed.) *J.L. Austin: Philosophical Papers* (3d Ed.) (pp. 175-204). Oxford: Oxford University Press.

Axelrod, R. (1984). *The Evolution of Cooperation*. New York, NY: Basic Books.

Cosmides, L. and Tooby, J. (1989). Evolutionary Psychology and the Generation of Culture: I. Case Study: A Computational Theory of Social Exchange. *Ethology and Sociobiology*, 10, 51-97.

Cosmides, L. and Tooby, J. (2008). Can a General Deontic Logic Capture the Facts of Human Moral Reasoning? How the Mind Interprets Social Exchange Rules and Detects Cheaters. In W. Sinnott-Armstrong (ed.), *Moral Psychology* vol.1 (pp.53-119). Cambridge, MA: A Bradford Book: The MIT Press.

Damasio, A. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York, NY: Gossett/Putnam.

Damasio, A. (2003). *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain*. New York: Harvest Books.

Frankfurt, H. (1969). Alternate Possibilities and Moral Responsibility. *Journal of Philosophy*, LXVI(23), reprinted in Frankfurt (1988), *The Importance of What We Care About* (pp. 1-10). Cambridge: Cambridge University Press.

Frankfurt, H. (2003). Some Thoughts Concerning PAP. In D. Widerker & M. McKenna (Eds.), *Freedom, Responsibility, and Agency* (pp. 339-45). Aldershot: Ashgate Press.

Gilbert, D. and Malone, S. (1995). The Correspondence Bias. *Psychological Bulletin,* 117(1), 21-38.

Greene, J., Sommerville, R., Nystrom, L., Darley, J., and Cohen, J. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science*, 293, 2105-8.

Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review*, 108(4), 814-834.

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

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

Knobe, J. (2006). The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology. *Philosophical Studies*, 130, 203-231.

Knobe, J. and Nichols, S. (eds.) (2008). *Experimental Philosophy*. New York, NY: Oxford University Press.

Nichols, S and Knobe, J. (2007). "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions" *Nous*. Reprinted in Knobe and Nichols (2008), 105-126.

Ortiz, D. (2004). Adolescence, Brain Development and Legal Culpability. American Bar Association Juvenile Justice Center. Retrieved October 15th, 2008, from http://www.abanet.org/crimjust/juvjus/Adolescence.pdf.

Prinz, J. (2008). Is Morality Innate? In W. Sinnott-Armstrong (ed.), *Moral Psychology* vol.1 (pp.367-406). A Bradford Book: MIT Press.

Robinson, Paul H. (2005). *Criminal Law: Case Studies & Controversies*. New York, NY: Aspen Publishers.

Rutkowski, David S. (1996). A Coercion Defense for the Street Gang Criminal: Plugging the Moral Gap in Existing Law. *Notre Dame Journal of Law, Ethics and Public Policy*, 137-226.

Strawson, P. F. (1982). Freedom and Resentment. In G. Watson (ed.), *Free Will* (pp.59-80). Oxford: Oxford University Press.

Wallace, R. Jay. (1994). *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

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

[i] It may be objected that prediction is not the same thing as causal determination, so the experimental results do not establish compatibilist intuitions on the part of the test subjects. However, the computer does predict the future on the basis of causal laws. Secondly, Nahmias et.al. ran their experiment with alternative descriptions of determinism, achieving similar results (ibid. p.88). One such scenario described a universe that "is re-created over and over again, starting from the exact same initial conditions and with all the same laws of nature." Another invited test subjects to "Imagine there is a world where the beliefs and values of every person are caused completely by the combination of one's genes and one's environment." The experimental results were comparable to the supercomputer case, pre-empting any objection based on the distinction between prediction and causal determination.

[ii] I criticize their explanation of these results elsewhere, featuring two main lines of argument (author 2009). First, I argue that they fail to establish that the role of emotion in generating compatibilist intuitions justifies the charge of performance error. The crux of this argument is that the charge of performance error can only be made relative to a theory of normal function, and that current theories of normal function cannot be used to sustain the charge. Second, I raise doubts about the reliability of the psychological processes that generate incompatibilist intuitions.

[iii] Cf. Knobe (2006), which finds that survey respondents are more likely to judge that harmful side effects of intentional actions are caused intentionally than are beneficial side effects.

[iv] Even for the insanity defense, psychological compulsion alone is not the issue. As I will discuss below, causal determination by one's settled personality traits is not a legally acceptable incapacity excuse.

[v] See Rutkowski (1996) for a review and criticism of statutes that treat gang-membership as an aggravating factor. Rutkowski argues that lower-level gang members should be able to excuse their criminal actions by appeal to the coercion defense. As he recognizes, however, his position puts him at odds with existing law and public opinion.

[vi] The matter was settled in the case of *United States v. Alexander* (471 F.2d 923 (D.C. Cir. 1973), which rejected the "rotten social background" defense.

[vii] I discuss this example more fully in (reference). The text of the decision is available on-line at http://www.kscourts.org/kscases/supct/1998/19980417/76131.htm.

[viii] *Roper, Superintendent, Potosi Correctional Center v. Simmons*, No. 03-633.Argued October 13, 2004--Decided March 1, 2005.

[ix] 2008 Florida Statute 985.556, available on-line at http://www.leg.state.fl.us/Statutes/index.cfm?App_mode=Display_Statute&Search_String=&URL=Ch0985/SEC556.HTM&Title=-%3E2008-%3ECh0985-%3ESection%20556#0985.556.

[x] Another oft-noted source of bias in moral judgment is the "Fundamental Attribution Error" or "correspondence bias," which consists in "the tendency to draw inferences about a person's unique and enduring dispositions from behaviors that can be entirely explained by the situations in which they occur" (Gilbert & Malone 1995). A full discussion of this source of bias is beyond the scope of this paper, but the issues it raises are similar to the sources of bias I do discuss. Just as the role of emotion in moral judgment raises a question about its rationality, so too does the influence of the Fundamental Attribution Error. In both cases, two related but separable questions arise. First, is the alleged "bias" a functionally normal part of moral competence? This is a psychological question. Second, are judgments made under the influence of this mechanism morally justifiable? In contrast to my claims about emotion-induced compatibilist judgments, I

am less optimistic that either of these questions can be answered in the affirmative with respect to the Fundamental Attribution Error. After all, the Fundamental Attribution Error leads to mistaken assumptions about the behavioral dispositions of its targets. So if one of the psychological functions of moral judgment is to identify suitable partners for reciprocal altruism (as I argue below), the Fundamental Attribution Error works at cross-purposes psychologically speaking.

[xi] Neuro-imaging studies (Greene et. al. 2001) and brain-injury studies (Damasio 1994) provide further evidence.

[xii] Insanity pleas or appeals to severe mental defects such as mental retardation are different. In those cases, incapacity is cited to cancel the expectation that the accused is a moral agent at all. In other words, the insane person is not excused because he is expected to be cooperative in the future, but because he is withdrawn from the pool of potential cooperators. Cf. Strawson's (1982) distinction between excusing conditions and excluding conditions.

[xiii] Thanks to Matthew Groe for pressing this question. Cf. Appiah (2008 p.149): "The force of [Haidt's] studies is to make us doubt that there's any deep relation between our moral judgments and the explicit rationales we offer for them."

[xiv] Haidt's study of "moral dumbfounding" may leave the impression that moral reasoning never works. In a study of moral judgments about consensual incest between siblings, Haidt found that refuting a person's moral reasoning does not by itself cause her to change her moral judgment (Haidt 2001). However, the lesson of moral dumbfounding is not that reasoning never works, but that it fails when the underlying emotional response remains unaltered.

[xv] I press a similar point against Nichols and Knobe (2007) in (author 2009).

[xvi] The qualification "for that reason" is important. For example, if disgust is part of the psychological process that leads you to condemn the violation of some norm, moral justification of your judgment would focus on the violated norm rather than your feeling of disgust. That is not to deny that emotions may sometimes be cited as reasons. But when they are, their relevance is based in moral norms, not the mere fact of the emotion's causal role. For example, if you could save only one of two people from a burning building, and were asked to justify your choice to save your spouse over a stranger, it would be relevant to cite your love for your spouse as a reason for your choice. (Thanks to Felipe De Brigard for this example.) However, the relevance of this appeal is not based on the mere fact that love was part of the causal process leading to the choice, because emotions are also causally implicated in many morally indefensible actions. Rather, the appeal to love is relevant to the extent that it connects with a moral norm (e.g., the heightened duty of care towards our loved ones).