

BAD ACTS, BLAMEWORTHY AGENTS, AND INTENTIONAL ACTIONS

Some problems for juror impartiality

Thomas Nadelhoffer

In this paper, I first review some of the recent empirical work on the biasing effect that moral considerations have on folk ascriptions of intentional action. Then, I use Mark Alicke's affective model of blame attribution to explain this biasing effect. Finally, I discuss the relevance of this research—both philosophical and psychological—to the problem of the partiality of jury deliberation. After all, if the immorality of an action does affect folk ascriptions of intentionality, and all serious criminal offenses—e.g., murder and rape—are immoral in addition to being illegal, then a juror's ability to determine the relevant mens rea (i.e., guilty mind) of a defendant in an unbiased way may be seriously undermined.

1. Introduction

In the landmark *Smith* case of 1961, jurors in England had to determine the guilt of a man named Smith who had driven a car containing stolen goods in a zigzag course in order to shake off a policeman who had been clinging to the side of the car. When the policeman was finally shaken off, he rolled into oncoming traffic and sustained fatal injuries (*D.P.P v. Smith* [1961] A.C. 290). Imagine that you are on that jury and your task is to decide whether Smith intentionally killed the policeman. In addition to considerations about Smith's relevant mental states and the relationship between these mental states and his actions—e.g., did Smith foresee that his actions would bring about the policeman's death—what other factors would affect your verdict? Would your decision concerning whether Smith killed the policeman intentionally be influenced by your evaluative belief that Smith brought about bad consequences? On the surface, it seems that the goodness or badness of Smith's actions should be completely *irrelevant* to the question of whether he performed them *intentionally*, but there is growing evidence that ascriptions of intentional actions are often influenced by evaluative considerations.

In this paper, I first briefly review some of the recent empirical work on the relationship between moral judgments and folk ascriptions of intentional action. Then, I shed light on the nature of this relationship by discussing Mark Alicke's affective model of blame attribution (2000). Next, I argue that Alicke's research—when coupled with recent data concerning folk ascriptions of intentional action—gives us reason to worry that jury deliberations in criminal trials involving serious crimes may be partial or biased in a fundamental way. And while psychologists long ago identified how the appearance, gender, race,

occupation, or sexual preference of the defendant and the victim may sometimes bias jury deliberations, the main point of this paper is to suggest that perhaps there is an even more basic sort of partiality that occurs when jurors are asked to make judgments concerning a defendant's mental states—especially when the crime in question is a serious one. After all, if the *immorality* of an action or side effect biases folk ascriptions of intentionality, and all serious criminal offenses such as murder and rape are immoral in addition to being illegal, then a juror's ability to determine the relevant *mens rea* (i.e., guilty mind) of someone like Smith in an unbiased way may be seriously undermined.¹ After considering some possible solutions to the particular type of juror partiality I have identified, I conclude that philosophers, psychologists, and legal theorists will need to continue to work together if we are to minimize the biasing effect that moral judgments have on jurors' judgments concerning the *mens rea* of defendants.

2. Setting the Stage

There is growing empirical evidence that people are more likely to judge that a morally negative action or side effect was brought about intentionally than they are to judge that a structurally similar non-moral action or side effect was brought about intentionally (e.g., Knobe 2003a, 2003b, 2004a; Nadelhoffer 2004a, 2004b, 2004c, 2005). So, for instance, if two individuals *A* and *B* place a single bullet in a six shooter, spin the chamber, aim the gun, and pull the trigger, but *A* shoots a person and *B* shoots a target, people are more likely to say that *A* shot the person intentionally than they are to say that *B* shot the target intentionally—even though their respective chances of success (one-in-six) and their control over the outcome are identical in both cases.

This goes right to the heart of a long-standing debate in the philosophy of action concerning the nature and proper role of ascriptions of intentionality. One of the central issues of this debate is whether moral considerations do—or *should*—affect our application of the concept of intentional action. While some scholars suggest that our use of this concept is often affected by moral considerations (e.g., Bratman 1987; Duff 1982, 1990; Harman 1997), others claim that moral considerations either do not or should not have an effect (e.g., Butler 1978; Katz 1987; Mele and Sverdluk 1996). On this latter view, while we may correctly appeal to the intentionality of an action in our attempt to determine someone's moral or legal responsibility, the converse is not the case—i.e., attributions of blame and praise should not affect our ascriptions of intentional action.²

For now, I want to provide a brief sketch of the recent debate concerning the relationship between moral judgments and judgments of intentionality. The most natural place to start such an investigation is with the work of Joshua Knobe—one of the first philosophers to bring data about folk intuitions to bear on issues in the philosophy of action. In a series of novel experiments, Knobe set out to determine whether folk intuitions about the intentionality of foreseeable yet undesired side effects are influenced by moral considerations (Knobe 2003a, 2003b). Each of the 78 participants in the first of these side effect experiments were presented with a vignette involving either a 'harm condition' or a 'help condition'. Those who received the harm condition read the following vignette:

The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment'. The chairman of the board answered, 'I don't care at all about

harming the environment. I just want to make as much profit as I can. Let's start the new program'. They started the new program. Sure enough, the environment was harmed. (Knobe 2003a, 191)

They were then asked to judge how much blame the chairman deserved for harming the environment (on a scale from 0 to 6) and to say whether they thought the chairman harmed the environment intentionally; 82 percent of the participants claimed that the chairman harmed the environment intentionally.

Participants in the help condition, on the other hand, read the same scenario except that the word 'harm' was replaced by the word 'help'. They were then asked to judge how much praise the chairman deserved for helping the environment (on a scale from 0 to 6) and to say whether they thought the chairman helped the environment intentionally. Only 23 percent of the participants claimed that the chairman intentionally helped the environment (Knobe 2003a, 192). When Knobe first published these surprising results he concluded that people do—and presumably should—rely on their judgments concerning the badness of an action in determining whether the action was performed intentionally, and then they use these intentionality judgments to determine whether the agent deserves blame for having performed the action in question (Knobe 2003a, 2003b, 2004). However, according to Knobe's original analysis of the data, moral goodness does not have a similar influence on folk ascriptions of intentional action.

I have subsequently argued that not only can the moral goodness of an action or side effect influence our judgments of intentionality (Nadelhoffer 2005)—albeit to a lesser degree than badness—but also that judgments concerning the moral blameworthiness or praiseworthiness of agents can have a similar influence on our ascriptions of intentional action (Nadelhoffer 2004b).³ For present purposes, I am going to simply assume that my arguments for the latter claim are correct—although for the problem of juror partiality that I will be examining in this paper, not a lot will hinge on the issue. Keep in mind that if any negative moral considerations influence our judgments concerning the intentionality of an action or side effect, then we have reason to worry that jurors may not be able to make impartial judgments about a defendant's mental state in cases involving bad acts or blameworthy agents. Indeed, it appears that the more negative a case is—morally speaking—the less likely it becomes that jurors will be impartial. But I am getting ahead of myself. Before I discuss the particular problem of jury partiality that I have purportedly identified, I first want to examine a recent model of the psychology of blame that lends additional support to my own view concerning the relationship between blame attribution and ascriptions of intentional action.

3. Mark Alicke's Culpable Control Model of Blame Attribution

One of the key notions underlying much of the recent research on moral psychology is 'automaticity'—i.e., 'the mind's ability to solve many problems, including high-level social ones, unconsciously and automatically' (Greene and Haidt 2002, 517). The seeming ubiquity of these automatic mental processes has led some researchers to reject rationalist models of moral psychology in favor of non-rationalist affective models. According to these affective models, 'moral judgment is more a matter of emotion and affective intuition than deliberative reasoning' (Greene and Haidt 2002, 517). And while these new explanatory models make room for certain types of higher cognition, they nevertheless suggest that emotional

and non-rational processes, rather than deliberative and rational ones, are primarily responsible for our moral judgments. One affective model of moral psychology that is particularly salient for our present purposes is Mark Alicke's model of the psychology of blame.

Alicke develops what he calls the Culpable Control Model (CCM) of blame attribution—a model that purportedly explains, 'the conditions that increase as well as mitigate blame and analyzes the process by which blame and mitigation decisions are made' (Alicke 2000, 557). Unlike other theoretical perspectives on blame and responsibility that focus on *normative* questions concerning how ascriptions of blame and responsibility should be made, the CCM focuses on the cognitive factors that actually influence these ascriptions, i.e. rather than discussing how judgments concerning blameworthiness should *properly* be made in *ideal* circumstances, he examines how they are *actually* made in *ordinary* circumstances.

According to the CCM, the primary factor in ascriptions of blame is the personal control—i.e., 'the freedom to effect desired behaviors and outcomes or to avoid undesired ones' (Alicke 2000, 557)—of the agent who has performed the morally inappropriate act. Alicke identifies three different aspects of personal control: (a) the mental element (e.g., mental states such as desires, plans, motives, etc.), (b) the behavioral element (e.g., actions and omissions), (c) the consequential element (e.g., immediate and extended 'behavioral outcomes'). And these three aspects in turn coincide with the following structural links: 'a link between mind and behavior, one between behavior and consequence, and one between mind and consequence' (Alicke 2000, 557).⁴ On this view, structural links designate the different factors of personal control that affect ascriptions of blame and responsibility. Whenever these factors of personal control are firmly established, ascriptions of blame intensify, whereas if these factors are somehow constrained, blame is mitigated.

While the CCM is similar to traditional rationalist models of blame and responsibility in acknowledging that 'people are socialized to predicate blame on criteria such as intention, causation, and foresight' (Alicke 2000, 557), it differs from these other models in the emphasis it places on the claim that 'personal control judgments and blame attributions are influenced by relatively unconscious, spontaneous evaluations of the mental, behavioral, and consequence elements. Spontaneous evaluations are affective reactions to the harmful event and the people involved' (Alicke 2000, 558). According to Alicke, these spontaneous and relatively unconscious responses can be triggered by both the 'evidential structural linkage information' concerning the three aforementioned factors of personal control and other 'extra-evidential factors' such as a person's appearance, reputation, social status, etc. As he says:

When blame-validation mode is engaged, observers review structural linkage evidence in a biased manner by exaggerating the actor's volitional or causal control, by lowering their evidential standards for blame, or by seeking information to support their blame attribution. In addition to spontaneous evaluation influences, blame-validation processing is facilitated by factors such as the tendencies to over ascribe control to human agency and to confirm unfavorable expectations. (Alicke 2000, 558)

Thus, the CCM suggests that judgments concerning personal control—and hence of moral blameworthiness—are unwittingly influenced by spontaneous affective reactions to the agents and actions involved. This influence can be both direct and indirect.

One way that spontaneous reactions influence structural linkage assessments is by altering perceptions of the evidence itself. When this happens, 'observers who spontaneously evaluate the actor's behavior unfavorably may exaggerate evidence that established her causal or volitional control and de-emphasize exculpatory evidence' (Alicke 2000, 566). Another way that these reactions affect observers' judgments is by engendering blame-validation processing that subsequently increases the observer's 'proclivity to favor blame versus non-blame explanations for harmful events and to de-emphasize mitigating circumstances' (Alicke 2000, 568–69). To the extent that the observer believes that the action in question is immoral, she will be inclined to look for explanations of the action that favor ascriptions of blame while at the same time overlooking explanations that do not. Thus, as a result of both spontaneous evaluations and blame-validation processing, observers tend to 'over ascribe control of human agency and to confirm unfavorable expectations' (Alicke 2000, 558).

To see how these kinds of biases operate, consider the following three studies: the participants in the first study were told that a homeowner shot someone in an upstairs bedroom who was presumed to be an intruder (Alicke, Davis, and Pezzo 1994). In the positive outcome version, the victim was described as a violent criminal who was responsible for other burglaries in the neighborhood. In the negative outcome version, the victim was the boyfriend of the homeowner's daughter who had been packing clothes for a trip. Participants were then asked to rate the causal relevance of a variety of factors, e.g., the fact that the homeowner had two beers to drink shortly before the shooting. Participants who received the negative outcome version found that the beer played a greater causal role in the shooting than participants who received the positive outcome version. This suggests that, 'spontaneous evaluations of the outcome directly affected blame ascriptions, which participants then buttressed by altering their causal control assessments' (Alicke 2000, 565).

In the second study, participants received a vignette that contained an ambiguous story about a subway passenger who was approached by four teenagers asking for money. Feeling threatened, the passenger nervously fired two shots thereby killing one of the teenagers. Upon reading the story, some participants were told that the teenagers were gang members with criminal records whereas others learned that the teenagers were star athletes trying to collect money for their football team. Not surprisingly, the blame ratings from the two respective pools of participants showed the same sort of 'outcome bias effect' that has been found in other studies (e.g., Alicke and Davis 1989)—i.e., the shooter was blamed more in the case involving the star athletes than in the case involving the gang members. Moreover, participants also learned that there were four eyewitness accounts—two for the prosecution and two for the defense—and they were told that owing to time limitations they would each only be able to read the testimony of three of the four. Interestingly, 75 percent of the participants in the star athlete group preferred to read more pro-prosecution testimony whereas 60 percent of the participants in the gang member group preferred to read pro-defense testimony. Alicke concludes that studies such as these show that participants 'who reacted more negatively to the actor for killing innocent victims favored information that supported a blame attribution' (Alicke 2000, 567).

Finally, in the third study, participants read about a driver who got into an accident while speeding (Alicke 1992). Participants learned that the driver was speeding either to hide an anniversary present or a vial of cocaine. Moreover, they learned that the driver encountered a number of environmental obstacles—slippery road, poor visibility, etc. Participants were then asked to say whether the driver's speeding or the environmental factors

played a greater role in causing the accident. The results showed that participants were more inclined to attribute the accident to the driver rather than the environmental conditions when the driver was hiding the cocaine than they were when he was hiding an anniversary gift. Once again, it appears that 'spontaneous evaluations of the actor's motives led participants to exaggerate his causal control over the accident' (Alicke 2000, 567).

Given the results of these kinds of experiments, Alicke concludes that 'cognitive shortcomings and motivational biases are endemic to blame' (2000, 557)—an admittedly disheartening finding. But as disturbing as it is that spontaneous moral intuitions and judgments often have such a negative effect on our ability to impartially consider the evidence surrounding a case, Alicke's CCM of blame attribution nevertheless helps shed light on the aforementioned biasing effect that moral considerations have on folk ascriptions of intentional action.

4. Ascriptions of Intentional Action and the Partiality of Jury Deliberation

In fleshing out the implications of the aforementioned research on moral psychology and folk ascriptions of intentional action for the problem of jury partiality, I will be primarily concerned with serious crimes that are *mala in se* (i.e., both illegal and immoral) such as murder and rape. In order for an agent to be held responsible for these types of crimes, the prosecution must prove two things: first, the agent has to be guilty of having performed the physical element of the offense—i.e. the *actus reus* or guilty act; second, the agent must have acted with the relevant mental or subjective element of the offense—i.e., the *mens rea* or adequately culpable state of mind. And for the types of crimes we are presently concerned with, *mens rea* usually implies that the agent performed the action either purposely, intentionally, designedly, consciously, or knowingly. In its narrowest interpretation—sometimes referred to as the elemental meaning—*mens rea* simply refers to the mental state explicitly required in the definition of the offense in question.⁵

Having briefly discussed the *mens rea* requirement of criminal law, we should now examine the problem that recent research into folk ascriptions of intentional action poses for juror impartiality. After all, to the extent that moral considerations affect folk ascriptions of intentional action, the ability of a defendant who is being prosecuted for a serious crime to receive a fair and unbiased assessment by the jurors is undermined. If the folk—in this case the jurors—are more likely to say that an action was performed *intentionally* if the action was *immoral*, and the defendant whose guilt the jurors are being asked to determine is accused of performing an act that is immoral in addition to being illegal, then the jurors will naturally be more inclined to say that the defendant's act was intentional. This problem is especially pressing in cases where jurors must judge whether the offense was committed with a sufficiently culpable mind.

In first-degree murder trials, for example, jurors are informed that in order for the defendant to be guilty as charged, he must have either (a) committed a murder that involved deliberate meditation, (b) committed a murder that involved extreme atrocity or cruelty, or (c) committed a murder during the commission of a felony (Model Jury Instructions on Homicide).⁶ For our present purposes, the jury instructions for deliberate meditation should suffice. The three elements are (a) that the defendant committed an unlawful killing (i.e., a killing that was not an accident or was not committed in self-defense), (b) that the killing was committed with malice (i.e., the defendant either had

an intent to cause death or caused the death intentionally), and (c) that the killing was committed with deliberate premeditation (i.e., the defendant thought before he acted and decided to kill after deliberation).

Based on these instructions, a juror must believe that the defendant's crime meets all three of these conditions if he is to be found guilty of first-degree murder with deliberate meditation. But if moral considerations—such as the immorality of the *actus reus*—influence juror ascriptions of intentionality, then these considerations will likewise influence whether the jurors judge that the defendant committed the crime with the requisite amount of malice and deliberation, *especially when acting with malice is simply taken to mean acting intentionally*. Similarly, if folk ascriptions of the intentionality of the *side-effects* of actions are affected by the immorality of the action, then in the *Smith* case mentioned earlier, the jurors' verdict may have been affected by the *immoral nature of the outcome* of Smith's actions.

To see whether the moral badness of the policeman's death may have affected the juror's decisions in the *Smith* case, I ran a preliminary study that involved vignettes based on the case. Participants were 126 undergraduates—half of whom received the following vignette:

Case 1 (C1):

Imagine that a thief is driving a car full of recently stolen goods. While he is waiting at a red light, a police officer comes up to the window of the car while brandishing a gun. When he sees the officer, the thief speeds off through the intersection. Amazingly, the officer manages to hold on to the side of the car as it speeds off. The thief swerves in a zigzag fashion in the hope of escaping—knowing full well that doing so places the officer in grave danger. But the thief doesn't care; he just wants to get away. Unfortunately for the officer, the thief's attempt to shake him off is successful. As a result, the officer rolls into oncoming traffic and sustains fatal injuries. He dies minutes later.

They were then asked the following questions. First, did the thief knowingly bring about the officer's death? Second, did the thief intentionally bring about the officer's death? Third, how much blame does the thief deserve for the death of the officer (on a scale from 0 to 6, 0 being no blame and 6 being a lot of blame)? The results were as follows:

- (Q1) 75 percent said that the thief knowingly brought about the officer's death.
- (Q2) 37 percent said that the thief intentionally brought about the officer's death.
- (Q3) The average blame rating was 5.11 on a 6-point scale.

In order to see whether the badness of the death of the officer and/or the perceived moral culpability of Smith was acting expansively on participants' ascriptions of knowledge and intentionality, I gave the other participants a case that is structurally identical to the first case—only this time it is an innocent driver whose actions bring about the death of an attempted carjacker. This case runs as follows:

Case 2 (C2):

Imagine that a man is waiting in his car at a red light. Suddenly, a car thief approaches his window while brandishing a gun. When he sees the thief, the driver panics and speeds off through the intersection. Amazingly, the thief manages to hold on to the side of the car as it speeds off. The driver swerves in a zigzag fashion in the hope of escaping—knowing full well that doing so places the thief in grave danger. But the driver doesn't care; he just

wants to get away. Unfortunately for the thief, the driver's attempt to shake him off is successful. As a result, the thief rolls into oncoming traffic and sustains fatal injuries. He dies minutes later.

The participants were then asked the following questions. First, did the driver knowingly bring about the thief's death? Second, did the driver intentionally bring about the thief's death? Third, how much blame does the driver deserve for the death of the thief (on a scale from 0 to 6, 0 being no blame and 6 being a lot of blame)? The results were as follows:

- (Q1) 51 percent said that the driver knowingly brought about the car thief's death.
- (Q2) 10 percent said that the driver intentionally brought about the car thief's death.
- (Q3) The average blame rating was 2.01 on a 6-point scale.

If we compare the results of C1 and C2, we see that even though the cases are identical in terms of the cognitive and conative considerations of the thief and the driver, the participants in C1 were more likely to say that the thief *knowingly* brought about the officer's death (75 percent) than the participants in C2 were to say that the driver knowingly brought about the death of the car thief (51 percent)—a statistically significant difference [$\chi^2(1, N = 126) = 7.62, p < 0.01$]. Moreover, the participants in C1 were also much more likely to say that the thief intentionally brought about the death of the officer (37 percent) than the participants in C2 were to say that the driver intentionally brought about the death of the car thief (10 percent)—a statistically significant difference [$\chi^2(1, N = 126) = 12.94, p < 0.001$]. And given the difference in the respective blame ratings from the two groups of participants (5.11 versus 2.01), we find *prima facie* evidence that moral considerations do explain the asymmetry of the participants' judgments.

My main goal in this study was to use a scenario based on a famous criminal trial to see whether moral judgments might influence jurors' judgments concerning whether a defendant acted either knowingly or intentionally. And while the results are mostly in line with earlier studies (Knobe 2003a, 2003b, 2004; Nadelhoffer 2004a), there are at least two noteworthy features of my *Smith* study. First, the results suggest that people's judgments concerning whether an agent knowingly brought about a result may also be affected by moral considerations. This is particularly important since for most criminal offenses, a defendant is maximally culpable as long as she either purposely (i.e., intentionally) or knowingly performed the prohibited action (or brought about the prohibited side effects). Hence, if moral judgments affect jurors' deliberations concerning both the intentionality and the foreseeability of a prohibited action or side effect, then the particular type of juror partiality I have identified in this paper is wider in scope than I had originally envisioned. Second, the results suggest that judgments concerning the moral character of either the victim or the defendant (or perhaps both) seem to have had an influence on participants' intuitions as well. However, more studies would admittedly need to be run that tested specifically for these two effects.

If further studies confirm these preliminary results, then the 'cards' were likely stacked against Smith before the trial even began given the moral gravity of the consequences of his actions—which is to say, because his actions brought about bad side effects, the jurors were more inclined to judge that the policeman's death was foreseeable and that Smith brought about his death intentionally. Thus, the influence that moral considerations have on folk ascriptions of intentional action may often undermine a juror's ability to

make impartial judgments concerning whether the defendant satisfies the requisite subjective or mental element of the crime he is being accused of having committed.⁷

Consider, for instance, the following three cases: first, in the frequently quoted *Desmond* case of 1868, a group of Fenian conspirators blew up a prison wall with dynamite in a failed attempt to free some of their imprisoned comrades. Even though their plot failed, the explosion killed a number of people living nearby. The conspirators were subsequently charged and convicted of murder (*Desmond, Barret & Others* [1868] 11 Cox C.C. 146). Second, in the *Hyam* case of 1975, a woman was jealous of a rival who had supplanted her in the affections of a mutual lover. As a result, the defendant went to her rival's house in the middle of the night, poured gasoline through her letterbox, and lit the door of her house on fire. Although the defendant's intention was merely to scare her rival away, the fire got out of hand and killed two of her rival's children. The defendant was subsequently charged and convicted of murder (*Hyam* [1975] A.C. 55). Finally, in *Regina v. Cunningham*, the defendant—who was desperate for money at the time—went into the cellar of the duplex he was renting and illegally removed the gas meter from the gas pipes in order to sell it. Although the switch for the gas was only two feet away from the meter, the defendant did not shut it off. Consequently, a considerable amount of gas filled the cellar and the duplex, partially asphyxiating another tenant (*Regina v. Cunningham*, Court of Criminal Appeal, 1957, 41 Crim. App. 155, [1957] 3 *Weekly L.R.* 76). The defendant was subsequently convicted of unlawfully and maliciously endangering the life of the tenant.⁸

If Alicke's CCM is correct, the ability of jurors to pass impartial judgments about the intentionality of a defendant's actions is greatly undermined in cases where they are being presented with a defendant who is charged with having committed an overtly immoral act. According to CCM, the immoral nature of the act can *spontaneously* trigger jurors to go into the default mode of blame attribution—a mode that causes them to be affected by negative and relatively unconscious reactions that prejudice both their assessment of the crime and their assessment of the structural linkages relative to establishing the defendant's guilt. This problem is compounded even further if Alicke is right that these spontaneous blame-validation biases are not 'exceptions to rational norms', but rather 'inherent aspects of blame ascription' (Alicke 2000, 558).

In this case, the mere fact that the defendant is accused of having committed a heinous crime increases the chances that jurors will view the evidence in a biased or impartial way. After all, once a juror's blame-validation mode has been triggered, she will be more likely to exaggerate the defendant's volitional or causal control and more inclined to lower the evidential standards of blame upon which the verdict is supposed to be based. Moreover, this spontaneous presumption of blame can cause the juror to *selectively look for evidence that supports blame attribution* while at the same time causing her to *overlook factors that might otherwise mitigate or exculpate blame or guilt*.⁹

This sobering possibility suggests that perhaps folk ascriptions of intentional action *should not* be affected by evaluative considerations, even if the evidence suggests that they frequently are. Minimally, to the extent that the sixth Amendment of the US Constitution guarantees that 'the accused shall enjoy the right to a speedy and public trial, by an *impartial jury*' (my emphasis), judges should consider taking more direct measures to inform jurors of the genuine risk each of them runs of allowing moral considerations to lead them to pass partial verdicts. Perhaps, if jurors were made aware of the various—and seemingly predictable—ways that their judgments can be unwittingly affected by

evaluative considerations and blame-validation biasing, they would be better able to live up to their legal duty to base their decisions solely on the material facts of the case. But as we are about to see, there is reason to suspect that not even heavy-handedness on the part of judges would help secure an impartial jury for the accused.

5. Some Possible Solutions

Some scholars have suggested that one way of minimizing the influence that moral considerations have on our ascriptions of intentionality would involve making sure that the criminal law defines mental states in a way that clearly distinguishes culpability from intentionality. According to Bertram Malle and Sara Nelson, for instance, even if legal scholars are correct in pointing out that ‘in criminal law, attributions of *mens rea* simply are (at least provisionally) attributions of culpability’ (Lacey 1993, 625), it does not follow that we cannot take steps to insure that moral judgments and judgments of intentionality remain separate in the minds of jurors. On their view, ‘to the extent that judgments of intentionality have important implications for verdicts and sentencing and do not just foreshadow them, every effort should be made to dissociate intentionality judgments from evaluative feelings or culpability assignments’ (Malle and Nelson 2003, 576). In short, Malle and Nelson suggest that we should do everything we can to separate the *mens* from the *rea* in the criminal law.

One dissociative strategy put forward by Malle and Nelson for separating the *mens* from the *rea* would involve asking jurors to make intentionality judgments while at the same time ‘exhorting them to leave their evaluative feelings aside’ (Malle and Nelson 2003, 576). Indeed, this is precisely the kind of possibility I entertained at the end of the last section when I suggested that perhaps the biasing effect that moral considerations have on jurors’ ascriptions of intentional action could be minimized if jurors were informed of the potential for bias. And while this is certainly something that we should do more of than we currently do, it is unclear whether taking these kinds of measures would be very effective.

In an interesting paper on what they call ‘mental contamination’—i.e., ‘cases whereby a judgment, emotion, or behavior is biased by unconscious or uncontrollable mental processes’—Timothy Wilson and Nancy Brekke suggest that in order for individuals to be able to avoid cognitive biases, the four following conditions would need to be met (Wilson and Brekke 1994, 118). First, they must be made aware of the unwanted mental processes in question. Second, they must be motivated to correct the error. Third, in addition to being motivated to correct for the error, they must be ‘aware of the direction and the magnitude of the bias’ (Wilson and Brekke 1994, 118). Finally, they must have sufficient control over their mental processes to be able to correct for the biases in question. For present purposes, I am going to assume that in order for Malle and Nelson’s jury instruction dissociative strategy to work, jurors would minimally need to be able to satisfy these four conditions as well. Unfortunately, the empirical data from social and cognitive psychology suggest that attempts on the part of jurors to keep judgments of intentionality separate from judgments from culpability will be unsuccessful.

First, there is gathering evidence that many (if not most) of our cognitive processes are inaccessible to conscious processing (see, e.g., Erikson and Simon 1980; Jacoby, Lindsay, and Toth 1992; Kihlstrom 1987; Nisbett and Wilson 1977; Posner and Rothbart 1989). Second, recent research has suggested that even if people are made aware of the

occurrence and magnitude of a cognitive bias, their ability to subsequently control their thoughts and feelings is often very limited (Bargh 1989; Logan 1989; Wegner 1989, 1992; Wegner and Pennebaker 1993). To get a sense for the relevance of this kind of research for our present concern, consider the research that has been carried out specifically on mental contamination and legal proceedings. For example, rules of evidence and other procedural rules have been put in place to prevent biases from affecting jurors' judgments concerning the evidence. One assumption that underlies a number of these rules is that jurors are able to discount or ignore testimony and evidence that turns out to be inadmissible. However, there is considerable evidence that people are unable to discount information very effectively (see, e.g., Sue, Smith, and Caldwell 1973; Thompson, Fong, and Rosenhan 1981, Wrightsman 1991).

Yet another problem for Malle and Nelson's suggestion concerning jury instruction is that people often underestimate their own susceptibility to mental contamination even once they are made aware of the general ubiquity of the underlying biases, while at the same time overestimating their own ability to control their mental processes. Consider, for example, the problem of prejudice and stereotyping. Even once people are made aware of the fact that stereotypes are usually learned at an early age and are often invoked automatically when we encounter members of certain groups (see, e.g., Billing 1985; Brewer 1989), people nevertheless underestimate their own tendencies to stereotype—which thereby undermines their ability to prevent prejudices and stereotypes from biasing their judgments (Devine 1989; Wegner 1994). Worse yet, it appears that under certain circumstances, the 'very act of trying to suppress stereotypic responses can increase their frequency' (Wilson and Brekke 1994, 133).

When we look at data on mental contamination and cognitive biases collectively, we have good reason to suspect that instructing jurors not to allow their culpability judgments to affect their intentionality judgments will be rather ineffectual. And for present purposes it makes little difference whether this is because (a) the biases in question are inaccessible to the jurors, (b) the jurors themselves underestimate their susceptibility to the biases, (c) the jurors overestimate their ability to control their mental processes, or (d) some combination thereof. Minimally, more studies would need to be run that specifically address people's ability to dissociate their judgments concerning the intentionality of an agent's action and their judgments concerning the culpability of the agent. And while I am doubtful that instructing jurors to dissociate intentionality and culpability will be effective—especially in cases involving serious crimes such as assault or murder—the issue is straightforwardly empirical. So, the verdict on Malle and Nelson's dissociative strategy will be out until the relevant studies are run.

In the meantime, I want to reconsider the aforementioned possibility that perhaps the entire way I have framed the issue concerning the relationship between moral judgments and ascriptions of intentional action is itself misguided or incorrect. One of the basic assumptions of my treatment of this relationship is that the former judgments often *distort* the latter ones. On this view, the folk concept of intentional action is ordinarily applied roughly along the lines of the five-component model put forward by Malle and Knobe—whereby performing an action intentionally 'requires the presence of five components: a desire for an outcome; beliefs about an action that leads to that outcome; an intention to perform the action; skill to perform the action; and awareness of fulfilling the intention while performing the action' (Malle and Knobe 1997, 12). However, once morally loaded features are built into scenarios, these features often

trump or override the standard application of the concept of intentional action—thereby distorting our judgments about intentionality. According to the moral biasing model I have put forward in this paper, affective responses often undermine our ability to apply the concept of intentional action in an unbiased way.

Indeed, the very fact that I have called this a ‘biasing effect’ indicates that I think that even though moral considerations surely do act expansively on folk ascriptions of intentional action, I nevertheless follow Mele and Sverdlik (1996) in believing that ideally they ought not have this effect—i.e., that whereas our ascriptions of intentional action should affect our judgments concerning an agent’s responsibility, the converse should not be the case. Nichols and Knobe have called the kind of model I have been developing—whereby affective or emotional responses sometimes *inappropriately* bias our otherwise rational judgments—a ‘performance error model’ (Nichols and Knobe n.d.). And while they are mainly interested in models of folk morality rather than models of folk psychology, I think the notion of a performance error is helpful in the present context—especially given that not everyone agrees that what I have been calling a ‘biasing effect’ represents a performance error at all.

Knobe, for instance, has suggested that folk psychology cannot be properly understood if we assume that its sole purpose is to predict and explain behavior—rather it is best understood as a multi-purpose tool (Knobe 2003b; Knobe and Burra forthcoming). While allowing that folk psychology plays an important role in the prediction and explanation of other people’s behavior, Knobe insists that it plays other important roles in our daily lives as well. On his view, some folk psychological concepts—such as intentional action—are ‘bound up in a fundamental way with evaluative questions—questions about good and bad, right and wrong, praise and blame’ (Knobe 2003b, 309–10). Given the intimate relationship between judgments of intentionality and moral judgments, it purportedly does not make sense to talk about moral judgments having a biasing effect on ascriptions of intentional action. After all, according to Knobe’s view, moral considerations ‘really do play a role in the very concept of intentional action’ (Knobe and Burra forthcoming).

While I wholeheartedly agree both with Knobe’s claim that folk psychology is best viewed as a multi-faceted tool as well as his claim that judgments of intentionality and moral judgments are intimately related, his view nevertheless fails to allay my present worries concerning jury partiality. After all, the problem I have been concerned with in this paper is that people are more likely to judge that a morally bad action or side effect is intentional than they are to judge that a structurally similar morally good or neutral action or side effect is intentional. Hence, while it may be true that the concept of intentional action cannot be fully understood lest we appreciate the role it plays in our moral deliberations, it nevertheless appears that people sometimes put the moral cart before the intentional horse. Consequently, even if the concept of intentional action is intimately bound up with moral considerations—which I entirely accept—there is still a question concerning the proper direction of fit.

If the concept of intentional action were not relevant to issues of moral responsibility, then it is unlikely that it would play such a central role in criminal proceedings in the first place. But the role it is supposed to play in the criminal law is as follows: judges and jurors are first supposed to determine whether the defendant is responsible for having performed the *actus reus* as well as whether she satisfied the relevant *mens rea* requirement for the crime of *x-ing*.¹⁰ Having made a decision concerning whether the defendant did

x purposely, knowingly, recklessly, etc., judges and jurors can then determine whether the defendant is legally culpable for *x-ing*. However, in cases where the crime with which the defendant has been charged is particularly bad—or the defendant is a particularly immoral or sordid individual—the empirical data suggest that there is a real risk that these moral features may distort the judgments of judges and jurors concerning whether the defendant purposely, knowingly, or recklessly committed the crime in question.

In this respect, the moral cart once again ends up ahead of the intentional horse. Surely, we don't want the very fact that a defendant is charged with having committed an immoral act to make it more likely that jurors will find her to be guilty of the crime in question. Hence, it looks like we have a performance error after all even if we accept Knobe's view of folk psychology—unless, of course, he thinks that not only should our ascriptions of intentional action inform our moral judgments, but also that the latter should sometimes inform the former. Surprisingly, Knobe suggests something along precisely these lines when he says that even though the blameworthiness or praiseworthiness of an individual cannot (or should not?) affect our ascriptions of intentional action, the intrinsic badness or goodness of an action often can (and presumably should) influence people's judgments of intentionality (Knobe and Burra forthcoming).

But even if Knobe's 'badness not blame' hypothesis were correct—and as I have already suggested, there is evidence that it is not—this would at best only solve half of the problem with juror partiality that I am presently addressing. After all, if jurors in a trial involving a gruesome death are more likely to say that the defendant intentionally brought the death about because the death is perceived to be intrinsically bad, then in most cases involving serious crimes, the cards really are stacked against the defendants from the start. This problem persists even if the blameworthiness of the defendant is not similarly affecting jurors' judgments of intentionality.

Ultimately, the issue with which I am presently concerned is not whether ascriptions of intentional action are relevant to our moral considerations—something few people would deny—but whether the gathering data on the relationship between the two give us reason to worry that *mens rea* concepts such as intentional action are likely to be used impartially in criminal proceedings. By my lights at least, Knobe's preferred account of the relationship between folk psychology and folk morality does not help allay this doubt. After all, if the main worry is that negative moral judgments concerning either the badness of the crime or the blameworthiness of the defendant are actually influencing the judgments of intentionality that jurors are supposed to rely on in determining the defendant's culpability, then it is a small consolation to be told that these two kinds of judgments are intimately related. But if, on the other hand, the influence that moral considerations have on ascriptions of intentional action really does amount to a performance error—at least as far as the criminal law is concerned—and if we have reason to doubt whether jurors can successfully avoid making the error even if they are made aware of it, then where does that leave us when it comes to intentionality and the criminal law?

If *mens rea* concepts such as knowingly, purposely, and intentionally are going to continue to play a role in legal proceedings, we need to do everything within our power to insure that they are used impartially. Figuring out how best to accomplish this goal will require more of the kind of empirical research I have already examined. Presumably, leaving judges and jurors to their own devices will continue to be inadequate. But until we have a better understanding of both the nature and depth of the problem, we will be unable to devise any viable solutions. One important step involves taking a closer

look at the relationship between folk psychology and folk morality. Another step involves a close examination of the role that *mens rea* concepts play in ordinary language and the criminal law. This is an investigation that will require philosophers, psychologists, and legal scholars to work hand in hand. If my own project serves to motivate further research along these lines, then it will have been a success even if I admittedly left a number of important questions unanswered.

ACKNOWLEDGMENTS

I would like to thank Alfred Mele, Joshua Knobe, Joel Anderson, George Rainbolt, Virginia Tice, and two anonymous referees for their helpful comments and suggestions on earlier drafts of this paper.

NOTES

1. While the main focus in this paper is on the potential partiality of jury deliberations, the problems I discuss concerning the biased application of *mens rea* concepts would presumably arise in trials that only involve judges—although more studies would need to be run that tested whether judges were immune to the sort of biases at issue in a way that jurors are not. Given the data I discuss in Section 3, I doubt that judges are any better at avoiding these biases than the folk—but it is admittedly a hunch on my part.
2. My own view on this matter has evolved. Whereas I originally agreed with Knobe that moral considerations both do *and* should influence our ascriptions of intentional action, I now think the normative claim that the former should affect the latter is incorrect for the reasons I discuss in Section 4.
3. For the purposes of this paper, whenever I discuss intentionality, I am only talking about the question of whether an agent's actions are intentional. This sort of intentionality is to be distinguished from discussions of intentionality that one finds in the literature on the philosophy of mind. When philosophers talk about intentionality in this latter context, they are usually interested in the question of how some of our mental states can be *about* things in the world.
4. These three structural links are called 'volitional behavioral control', 'volitional outcome control', and 'causal control', respectively (Alicke 2000, 560).
5. While the Model Penal Code does not define what it means for an action to be done intentionally, it gives the following guidelines for deciding whether an action is done purposely or knowingly (Section 2.02): (a) A person acts purposely with respect to a material element of an offense when: (i) if the element involves the nature of his conduct or a result thereof, it is his conscious object to engage in conduct of that nature or to cause such a result; and (ii) if the element involves the attendant circumstances, he is aware of the existence of such circumstances or he believes or hopes that they exist. (b) A person acts knowingly with respect to a material element of an offense when: (i) if the element involves the nature of his conduct or the attendant circumstances, he is aware that his conduct is of that nature or that such circumstances exist; and (ii) if the element involves a result of his conduct, he is aware that it is practically certain that his conduct will cause such a result. In cases involving the types of serious crimes we have been discussing, the requisite *mens rea* is usually either purposely or knowingly. For a more thorough discussion of *mens rea*, see Duff (1990), Hart (1968), and Kenny (1968).

6. For a complete version of Model Jury Instructions on Homicide go to: www.sociallaw.com.
7. That such biasing occurs has been shown in other psychological experiments as well. For example, Fischhoff's research on 'hindsight bias' suggests that the actual outcome of an action may alter an observer's judgment concerning how foreseeable the risks were to the agent before the action was performed (Fischhoff 1975). As a result of hindsight bias, events that have *already occurred*—which is the case in *all criminal trials*—are judged to have been more likely to occur than they would have been judged *before their occurrence*. In these studies, two groups of subjects were given the very same set of antecedent conditions leading up to an accident—the only difference being that some subjects were told that the accident in question had already taken place, whereas others were not. Interestingly, the subjects who were told that the accident actually occurred were much more likely to say that the agent could have foreseen the accident than those who were not told the accident occurred, even though all of the antecedent conditions were the same in both groups. These studies also suggested that in addition to causing observers to overestimate the degree to which decision makers foresaw the accidents before the accidents occurred, hindsight bias may cause observers to distort the level of uncertainty facing the decision maker. And in cases where juries are asked to determine whether the consequences of a defendant's actions were either foreseeable, foreseen, or intentionally brought about, the potential for hindsight bias is particularly problematic.
8. The conviction was later overturned due to the way the judge had defined 'maliciously' when giving the jury their instructions.
9. And while the possibility that jurors' verdicts are consistently being affected by blame-validation biasing is problematic in and of itself, it becomes even more troublesome when jurors are further affected by other arbitrary factors such as the race of the defendant or the victims. As Alicke says, 'Racially prejudiced observers . . . who respond more negatively to a minority group member's harmful actions, require less evidence of intention, negligence, foresight, or causal influence than unbiased observers' (Alicke 2000, 566). In addition to race, psychologists have also shown that other factors can produce these kinds of negative spontaneous reactions as well, such as the appearance, personality, and demographics of the observers, perpetrators, or victims. Of course, the idea that the race, appearance, or character of the defendant might prejudice the jury is neither novel nor surprising, but when considered in light of the other factors I have already examined that may bias jurors' ascriptions of intentional action and blame, it certainly deepens the fear that getting a fair trial by jury is neither as common nor as easy as we had previously hoped.
10. It is worth highlighting the fact that before the problem of juror partiality I have been discussing could arise in a criminal proceeding, the jurors would have already decided that the defendant is responsible for committing the prohibited act in question. So, for instance, in the Smith case no one questioned the fact that his actions ultimately led to the officer's death. The issue was whether he was guilty of homicide or some lesser offense such as manslaughter—an issue that can only be resolved by making judgments about Smith's mental states.

REFERENCES

- ALICKE, M. D. 1992. Culpable causation. *Journal of Personality and Social Psychology* 63: 368–78.
- . 2000. Culpable control and the psychology of blame. *Psychological Bulletin* 126: 556–74.

- ALICKE, M. D., and T. L. DAVIS. 1989. The role of a posteriori victim information in judgments of blame and sanction. *Journal of Experimental Social Psychology* 25: 362–77.
- ALICKE, M. D., T. L. DAVIS, and M. V. PEZZO. 1994. A posteriori adjustment of a priori decision criteria. *Social Cognition* 8: 286–305.
- BARGH, J. A. 1989. Conditional automaticity: Varieties of automatic influence in social perception and cognition. In *Unintended thought: Limits of awareness, intention, and control*, edited by J. S. Uleman and J. A. Bargh. New York: Guilford Press.
- BILLING, M. 1985. Prejudice, categorization, and particularization: From a perceptual to a rhetorical approach. *European Journal of Social Psychology* 15: 79–103.
- BRATMAN, M. 1987. *Intention, plans, and practical reason*. Cambridge, Mass.: Harvard University Press.
- BREWER, M. B. 1989. A dual process model of impression formation. In *Advances in social cognition*, edited by R. S. Wyer and T. K. Srull. Hillsdale, N.J.: Erlbaum.
- BUTLER, R. 1978. Report on Analysis 'problem' no. 6. *Analysis* 38: 113–14.
- DEVINE, P. G. 1989. Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology* 56: 5–18.
- DUFF, R. A. 1982. Intention, responsibility, and double effect. *The Philosophical Quarterly* 32 (126): 1–16.
- . 1990. *Intention, agency, and criminal liability*. Oxford: Basil Blackwell.
- ERIKSON, K. A., and H. A. SIMON. 1980. Verbal reports as data. *Psychological Review* 87: 215–51.
- FISCHOFF, B. 1975. Hindsight does not equal foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance* 1: 288–99.
- GREENE, J., and J. HAIDT. 2002. How (and where) does moral judgment work? *Trends in Cognitive Science* 6: 517–23.
- HARMAN, G. 1997. Practical reasoning. In *The philosophy of action*, edited by A. Mele. Oxford: Oxford University Press. First published in 1976 in *Review of Metaphysics* 79: 431–63.
- HART, H. L. A. 1968. *Punishment and responsibility*. Oxford: Oxford University Press.
- JACOBY, L., S. D. LINDSAY, and J. P. TOTH. 1992. Unconscious influences revealed: Attention, awareness, and control. *American Psychologist* 47: 802–9.
- KATZ, L. 1987. *Bad acts and guilty minds*. Chicago: University of Chicago Press.
- KENNY, A. 1968. Intention and purpose in law. In *Essays in legal philosophy*, edited by R. Summers. Oxford: Basil Blackwell.
- KNOBE, J. 2003a. Intentional action and side effects in ordinary language. *Analysis* 63: 190–94.
- . 2003b. Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology* 16 (2): 309–24.
- . 2004. Intention, intentional action, and moral considerations. *Analysis* 64: 181–87.
- KNOBE, J., and A. BURRA. Forthcoming. What is the relation between intention and intentional action? *The Journal of Cognition and Culture*.
- LACEY, N. 1993. A clear concept of intention: Elusive or illusory? *The Modern Law Review* 56: 621–42.
- LOGAN, G. D. 1989. Automaticity and cognitive control. In *Unintended thought: Limits of awareness, intention, and control*, edited by J. S. Uleman and J. A. Bargh. New York: Guilford Press.
- MALLE, B., and J. KNOBE. 1997. The folk concept of intentional action. *Journal of Experimental Social Psychology* 33: 101–21.
- MALLE, B., and S. NELSON. 2003. Judging *mens rea*: The tension between folk concepts and legal concepts of intentionality. *Behavioral Sciences and the Law* 21: 563–80.
- MELE, A., and S. SVERDLIK. 1996. Intention, intentional action, and moral responsibility. *Philosophical Studies* 82: 265–87.

- NADELHOFFER, T. 2004a. The Butler problem revisited. *Analysis* 64: 277–84.
- . 2004b. Praise, side effects, and intentional action. *The Journal of Theoretical and Philosophical Psychology* 24: 196–213.
- . 2004c. Blame, badness, and intentional action: A reply to Knobe and Mendlow. *The Journal of Theoretical and Philosophical Psychology* 24: 259–69.
- . 2005. Skill, luck, and intentional action. *Philosophical Psychology* 18: 343–54.
- NICHOLS, S., and J. KNOBE. n.d. Moral responsibility and determinism: The cognitive science of folk intuitions. Unpublished manuscript.
- NISBETT, R. E., and T. D. WILSON. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84: 231–59.
- POSNER, M., and M. K. ROTHBART. 1989. Intentional chapters on unintentional thoughts. In *Unintended thought: Limits of awareness, intention, and control*, edited by J. S. Uleman and J. A. Bargh. New York: Guilford Press.
- SUE, S., R. E. SMITH, and C. CALDWELL. 1973. Effects of inadmissible evidence on the decisions of simulated jurors: A moral dilemma. *Journal of Applied Social Psychology* 3: 345–53.
- THOMPSON, W. C., G. T. FONG, and D. ROSENHAN. 1981. Inadmissible evidence and juror verdicts. *Journal of Personality and Social Psychology* 40: 453–63.
- WEGNER, D. M. 1989. *White bears and other unwanted thoughts*. New York: Viking Press.
- . 1992. You can't always think what you want: Problems in the suppression of unwanted thoughts. In *Advances in experimental social psychology*. Vol. 25, edited by M. P. Zanna. San Diego, Calif.: Academic Press.
- . 1994. Ironic processes of mental control. *Psychological Review* 101: 34–52.
- WEGNER, D. M., and J. W. PENNEBAKER, eds. 1993. *The handbook of mental control*. Englewood Cliffs, N.J.: Prentice Hall.
- WILSON, T. D., and N. BREKKE. 1994. Mental contamination and mental correction of unwanted influences on judgments and evaluations. *Psychological Bulletin* 116 (1): 117–42.
- WRIGHTSMAN, L. S. 1991. *Psychology and the legal system*. Pacific Grove, Calif.: Brooks/Cole.

Thomas Nadelhoffer, Visiting Assistant Professor, Florida State University, Department of Philosophy, Tallahassee, FL 32306-1500, USA. E-mail: tnadelhoffer@gmail.com

