Effects of Moral Cognition on Judgments of Intentionality

Jennifer Nado
[draft of a paper to appear in The British Journal for the Philosophy of Science]

Abstract

Several recent articles on the concept of intentional action center on experimental findings suggesting that intentionality ascription can be affected by moral factors. I argue that the explanation for these phenomena lies in the workings of a tacit moral judgment mechanism, capable under certain circumstances of altering normal intentionality ascriptions. This view contrasts with that of Knobe ([2006]), who argues that the findings show that the concept of intentional action invokes evaluative notions. I discuss and reject possible objections to the moral mechanism view, and offer arguments supporting the model over Knobe's account on grounds of simplicity and plausibility.

1 Introduction

There is a growing view among the philosophical and psychological communities that humans are endowed with an innate, largely tacit ability to attribute mental states to other members of their species. Though the specific mechanisms underlying this ability

are often debated, it is generally agreed that the primary function of this innate commonsense psychology is to aid in the prediction and explanation of behavior. As such, one might assume that the only information employed in folk psychological judgment would be of an objective, non-normative nature. However, a number of recent studies by Joshua Knobe and others have demonstrated that attributions of intentionality can sometimes be influenced by the moral status of the relevant action. In particular, people seem much more willing to say that an action was intentional if it was morally bad.

Two general approaches to explaining this phenomenon have dominated the literature. The first, Knobe's own view, claims that the commonsense concept of intentionality invokes evaluative notions at its very core. This might be called the *competence* view; its claim is that the tendency to employ moral information in judgments of intentionality is part of an underlying conceptual competence. The second general view may be called the *performance* view. According to adherents of this approach, the commonsense concept of intentionality does not essentially involve evaluative notions. Rather, some other factor is interfering with intentionality judgment, causing individuals to misapply their own concepts.

Several versions of the performance view have been put forth, notable examples of which include Malle and Nelson ([2003]), Adams and Steadman ([2004b]), and Nadelhoffer ([2004]). Much of Knobe's recent work has centered on offering experimental data and philosophical arguments undermining such accounts, and I believe that no performance account currently on offer adequately deals with Knobe's objections. In this paper, I propose a new version of the performance view which I believe stands up

to all of Knobe's arguments. This account, which I call the Moral Mechanism Interference account, proposes a tacit moral mechanism which can distort morally relevant information when it conflicts with the mechanism's blameworthiness judgments.

The first section of this paper introduces the relevant experimental findings, as well as Knobe's own explanation of the data. The second section surveys the major performance accounts on offer and Knobe's objections to them, introducing the Moral Mechanism Interference account as a modification of these basic approaches. In the third section, a more detailed model of the Moral Mechanism Interference account is offered. In the final section, I discuss positive reasons to prefer the Moral Mechanism Interference account over Knobe's own competence account.

2 The competence hypothesis

Philosophers have long debated the conditions under which an action counts as intentional (see for example Bratman [1984] , Harman [1976]). Only recently, however, has interest arisen in the conditions under which ordinary people deem an action intentional. In an influential study, Malle and Knobe ([1997]) found that participants showed high interpersonal agreement in attributions of intentionality, providing evidence for a robust, shared concept. Malle and Knobe isolated five central features of this shared folk concept of intentionality—belief, desire, intent, awareness, and skill. In order to *A* intentionally, an agent must have desires that would be satisfied by *A*-ing, she must believe that *A*-ing would help fulfill those desires, she must intend to *A*, she must be aware that she is *A*-ing, and finally, she must have the skill or ability to successfully *A*.

All five factors must be present before participants will deem an action intentional. Or so it seemed—as Knobe would soon show, one or more of these prerequisites may apparently be waived if the action under discussion is morally significant.

Though observations of moral effects on intentionality had appeared in the literature prior to Knobe's studies, Knobe was the first to demonstrate how pervasive and varied such effects really are. Two main effects have been detected. The first, which had already seen occasional mention in the philosophical literature, involves the effect of skill on intentionality—we will call it the 'skill effect'. The second is Knobe's own discovery, and is often referred to as the 'side-effect effect'.

Knobe ([2003b]) offers an experimental demonstration of the skill effect. Participants were presented with one of four vignettes about a man named Jake—these will hereafter be referred to as the 'Jake' cases. In two vignettes, Jake is attempting to win a contest by shooting a bulls-eye (the 'achievement' condition). In the other two vignettes, Jake is attempting to kill his aunt in order to get inheritance money (the 'immoral' condition). Within each of these conditions, Jake is depicted either as being an excellent shot, or as being an inept marksman who happens to get lucky. The immoral/no-skill version, for example, runs as follows.

> 'Jake desperately wants to have more money. He knows that he will inherit a lot of money when his aunt dies. One day, he sees his aunt walking by the window. He raises his rifle, gets her in his sights, and presses the trigger. But Jake isn't very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes

wild… nonetheless, the bullet hits her directly in the heart. She dies instantly' (Knobe [2003b], p. 313).

Participants in this case (and in the immoral/skill case) were then asked whether Jake intentionally killed his aunt. In the achievement cases, participants were asked whether Jake intentionally hit the bulls-eye.

The results showed a strong contrast between the achievement and the immoral conditions. 79% judged Jake's behavior to be intentional in the achievement/skill case, while only 26% judged the achievement/no-skill case intentional. Thus, in the achievement condition, participants behaved as predicted by Malle and Knobe ([1997]). In the immoral condition, however, participants seemed to ignore the skill prerequisite. As expected, almost everyone (95%) judged Jake's skillful killing of his aunt to be intentional; however, a large majority (76%) judged the killing intentional even when Jake made a lucky shot.

In a follow-up experiment, Knobe demonstrated that this effect is not limited to immoral actions. Participants are willing to forgo the skill requirement when the action in question is morally good, as well. Knobe offered participants vignettes describing a WWII soldier named Klaus (hereafter, the 'Klaus' cases). In both vignettes, Klaus objects to the deeply immoral mission his regiment has been sent on. In one vignette, Klaus is a skilled marksman and expertly shoots a crucial communications device, thereby foiling the mission and saving innocent lives. In another vignette, Klaus is a poor marksman, but manages to hit the communication device by luck. Strikingly, 92% of participants given the latter, morally-good/no-skill case deemed Klaus's action

intentional. Thus, it seems that Malle and Knobe's proposed skill prerequisite only applies in morally neutral cases—if an action is morally good or morally bad, ordinary people are willing to ascribe intentionality even in the absence of sufficient skill.

Knobe's side-effect effect serves as an even more unexpected demonstration of the effects of morality on judgments of intentionality. Knobe ([2003a]) presented subjects with what I will call 'Chairman' cases, the first of which runs as follows:

> '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], p. 191).

He presented another group with the same vignette, only with 'harm' replaced by 'help'. Participants were then asked whether the chairman intentionally harmed/helped the environment. The results were quite remarkable—82% judged that the chairman intentionally harmed the environment, while only 23% claimed that the chairman intentionally helped the environment. Since the chairman did not have an intention to harm the environment (indeed, he ignores the environment entirely), the immoral case is once again a departure from Malle and Knobe ([1997]). In this case, however, the

asymmetry is not between morally loaded and morally neutral cases, but between morally good and morally bad cases.

A good overview of Knobe's explanation of these findings can be found in Knobe ([2006]). The primary claim of the account is that evaluative notions are central to the very concept of intentional action. If Knobe is correct, then in order to give a definition of the folk concept 'intentional' we would have to say something like 'an action is intentional if and only if it is *a* and *b* and *c* and *good*, *d* and *e* and *f* and *bad*, or *x* and *y* and *z* and *neutral*'. This, Knobe readily admits, goes against the very strong intuition that commonsense psychology is a 'quasi-scientific, purely naturalistic' way of understanding our fellow humans. A quasi-scientific commonsense psychology would have no use for concepts whose conditions of application changed according to whether or not we liked what we saw. Thus, Knobe claims, we should conclude that commonsense psychology is tailored not only to prediction and explanation, but also to the evaluation of action.

It is crucial to note that Knobe proposes that it is the *goodness* or *badness* of the outcomes which causes the observed asymmetries; for instance, people more frequently judge side-effects to be intentional when those side-effects are bad, and less frequently when the side-effects are good. Goodness and badness, on Knobe's account, are to be contrasted with moral praise and blame. Non-morally blameworthy outcomes can nonetheless be bad in Knobe's sense. Knobe offers the following example:

> 'Consider the agent who hurts his wife's feelings. Here we might say that the agent's behavior itself is bad. That is to say, when we ignore every other aspect of the situation, we might classify the hurting of the wife's feelings as a bad

> thing. Still, we will be unlikely to blame the agent if he has a good excuse (ignorance, mental illness, provocation, etc.)' (Knobe [2006], p. 217).

According to Knobe, then, the apparent moral effects are not really moral at all. The effects are caused solely by the *valence* of the relevant action, where an action's valence is a measure of its goodness or badness evaluated entirely *independently* from judgments about the blameworthiness of the agent. The primary motivation for this view (as opposed to the view that the effects are due to moral features) is the quite plausible assumption that information about whether an action was intentional plays a crucial role in assessments of the praiseworthiness or blameworthiness of that action.[i] This intuitive assumption would be undermined if information about praiseworthiness and blameworthiness played a crucial role in the application of the concept 'intentional'. Knobe's account takes non-moral valence to be an input to the assessment of intentionality, and intentionality to be an input to the assessment of praise and blame, thereby avoiding a looming circularity.[ii]

3 The performance response

Others have not been as eager as Knobe to abandon the notion of a purely objective commonsense psychology. As mentioned in the introduction, several accounts have been offered which attempt to explain away the effects as due to interference or distortion—i.e., to recast the effects as performance errors. I, too, find it difficult to

believe that commonsense psychology is fundamentally evaluative, and I believe that the performance approach is the correct one.

There are two types of account that I find particularly plausible; the first proposes interference from explicit, conscious beliefs, and the second proposes interference from emotional response.[iii] Both argue that blame plays a central role in distorting participants' intentionality judgments; this contrasts with Knobe's account, which proposes that the non-moral valence of the action is the crucial factor. Unfortunately, extensive discussion of these two performance accounts by Knobe has shown that, as stated, the accounts simply fail to fit all the available data. However, I am not yet convinced that this warrants abandoning the notion of an objective folk psychology, nor the quite intuitive idea that blame is, well, to blame.

As such, I propose a modified version of the performance explanations, according to which the interfering factor is a tacit moral mechanism. This mechanism uses a variety of inputs, many from the theory of mind mechanism, in order to produce moral judgments—and the mechanism is sometimes capable of going back and overwriting those inputs after its judgment has been made. This ability would be used to correct inputs which strongly conflict with the conclusion the moral mechanism has reached, thereby serving as a sort of 'error-catching' device as well as providing increased justification and reinforcement for the mechanism's judgment. In this section I will show that this moral mechanism interference hypothesis (hereafter, MMI) retains the insights of previous performance accounts while avoiding their empirical difficulties. The following section will focus on a detailed model of the overwriting process I propose.

The first of the two performance accounts I will discuss hypothesizes a direct influence of moral belief on ascriptions of intentionality. Mele ([2001]) suggests that explicit beliefs about moral responsibility could distort intentionality intuitions in the Jake cases. Participants' underlying intuitions might tell them that Jake's killing is unintentional, but this might conflict with an explicit belief that all blameworthy actions are intentional. Participants might then 'correct' their intuition to resolve the inconsistency. This, of course, presupposes that ordinary people have some fairly theoretically sophisticated beliefs about morality; even assuming that this is plausible, however, the suggestion runs into trouble.

If the asymmetry in Knobe's cases is being caused by the interference of an explicit belief, then it is reasonable to suppose that the effect would dissolve if participants come to change their belief. Knobe ([2003b]) used this assumption to put Mele's view to the test, presenting participants with a scenario describing an unintentional yet blameworthy action suggested by Mele himself. I will refer to this as the 'Drunk Driver' case.

> 'Bob got rip-roaring drunk at a party after work. When the party ended, he stumbled to his car and started driving home. He was very drunk at the time—so drunk that he eventually lost control of his car, swerved into oncoming traffic, and killed a family of five' (Mele [2001], p. 41).

Knobe then presented the same participants with one of the four Jake vignettes. Most participants judged Bob's behavior to be unintentional but nonetheless blameworthy;

Knobe therefore assumes that any participants who believed that all blameworthy actions are intentional were led to revise that belief in light of the Drunk Driver case. Responses to the Jake cases, however, continued to demonstrate a strong asymmetry between immoral and non-moral unskilled actions. On this basis, Knobe concludes that explicit interfering beliefs do not cause the asymmetry phenomenon.

It is of course an idealization to suppose that people *always* revise their beliefs in light of such counterexamples. Several exceptions spring to mind. First, the relevant belief and its counterexample may not be consciously entertained within a short enough time interval for the inconsistency to be noticed. This sort of case does not seem relevant here, however, given the proximity of the counterexample (the Drunk Driver vignette) to the judgment that Jake's killing was intentional—which, recall, is supposed to be due to interference of an explicit tokening of the 'all blameworthy actions are intentional' belief.

A counterexample may also go unnoticed if it is presented in a form in which it does not obviously contradict with the relevant belief—in cases involving complicated mathematical expressions, or abstract philosophical assertions.[iv] Again, this does not seem to be the case here; the proposed belief 'all blameworthy actions are intentional' and the participants' assertions that Bob's action in the Drunk Driver case was blameworthy yet unintentional form a quite glaring contradiction. There are of course other factors which could prevent the contradiction from being recognized—lack of attention, for instance. But there is also a further possibility—the relevant belief is never *consciously* entertained, and is for that reason unavailable for conscious revision.

If Mele's proposed interfering belief were *tacit*, there would be no reason to think that presentation with a counterexample would lead the participants to revise it.

Speculating further, if the belief formed part of an innate, tacit body of knowledge employed for reasoning within a specific domain—similar to the widely hypothesized theory of mind mechanism—there would be reason to suspect that the belief might be quite resistant to conscious revision. A plausible proposal, then, is that interference comes not from an explicit, conscious belief that all blameworthy actions are intentional, but from a tacit belief to similar effect which forms part of a relatively encapsulated innate moral reasoning system. This is the core hypothesis of the MMI account.

In fact, such a hypothesis is not particularly novel. We have seen that many researchers believe that humans possess an innate 'theory of mind'. Several researchers are now coming to believe that humans possess some kind of innate *moral* mechanism, as well. As with the theory of mind literature, the specific architecture of this mechanism is widely debated. Haidt ([2001]) proposes that moral judgment is primarily the result of quick, affect-driven intuitions, and that most moral reasoning serves as post-hoc justification. Greene ([2001]) holds a similar view. Sripada and Stich ([2006]) propose that humans possess a 'norm acquisition mechanism', which uses behavioral cues in the environment to infer the social and moral norms present in the person's culture, and an 'implementation mechanism', which uses these norms to produce judgments and to motivate compliance and punitive action. Rawls ([1971]), Dwyer ([1999]), Harman ([1999]), and Hauser ([2006]) each suggest that moral cognition might operate in a similar way to linguistic processing. Perhaps we possess an innate moral grammar, which when combined with input from the environment results in the ability to generate quick, automatic moral intuitions. Discussion of the potential suitability of these accounts for the proposed interfering moral mechanism will be deferred to the next

section; for now, it is enough to note that my proposed emendation of Mele's account is solidly in line with the current literature on moral judgment.

The second type of performance account to propose some effect of blame has been suggested both by Malle and Nelson ([2003]) and by Nadelhoffer ([2004]). The basic proposal is that negative affect towards the agent produces the interference. This negative affect causes a desire to blame the agent, and that causes participants to deem Jake's and the Chairman's actions intentional. Malle and Nelson write:

> 'Because intentional behaviors elicit more blame than unintentional behaviors, negative affect toward the agent can easily bias judgments of intentionality, because characterizing a behavior as intentional warrants more blame, anger, and perhaps aggression' (Malle & Nelson [2003], p. 575).

Nadelhoffer offers a similar analysis. On his view, the lack of concern for the environment evidenced in *both* the harm and help conditions of the Chairman case leads us to form negative evaluations of the chairman in *both* cases. The asymmetry in intentionality ascription is not due to the asymmetry in the goodness/badness of the outcomes, but rather to an asymmetry in how we treat blameworthy individuals.

> 'Insofar as subjects judge that an *agent* is blameworthy, they are more inclined to say that any negative side effects brought about by the agent are intentional and any positive side effects brought about by the agent are not intentional' (Nadelhoffer [2004], p. 209).

Thus, our feelings of blame explain both cases.

Knobe has two main arguments against this type of position, both expressed in Knobe and Mendlow ([2004]). Both center on the question of whether it is blameworthiness or non-moral valence that causes the observed asymmetries. As such, these arguments equally affect other proposals based on blame processing, such as MMI. In truth, the blame vs. valence question is at the heart of the disagreement between the competence and performance accounts. For that reason, I wish to postpone discussion of these arguments until a more detailed proposal of my model has been presented.

Even leaving Knobe's blame-based arguments aside, however, the affect-bias account suffers from empirical shortcomings. An interesting study by Young et al. ([2006]) provides significant worries for any affect-driven account. Young et al. presented the Chairman vignettes to participants with emotional processing deficits due to damage to the ventromedial prefrontal cortex. If the affect-bias account is correct, the patients' emotional deficits should prevent emotional response from biasing their intentionality judgments. However, these subjects demonstrated the side-effect effect just as readily as normal subjects. Therefore, it seems overwhelmingly likely that emotional response has little influence on attributions of intentionality. MMI, however, need not be troubled by this finding. The proposed moral mechanism is a *reasoning* system, employing a body of facts including theory of mind information, information about local norms, etc. Moral emotions, if they play any role in such moral reasoning, play a supporting role at best. [v]

Finally, though Knobe does not mention it, his Drunk Driver vignette also presents a challenge to the affect-bias account. I see no reason to assume that the drunk driver would be less likely to elicit negative affect than the crooked businessman described in the Chairman cases. If participants do form a negative evaluation of the drunk driver, and if this causes a tendency to judge his bad actions to be intentional, why don't participants say he killed the family intentionally? There is an asymmetry between responses to the Drunk Driver case and to the Jake and Chairman cases which cannot be explained merely by negative evaluation of the agent.

In fairness, however, the bare-bones version of MMI I have sketched thus far does not yet provide an explanation for this asymmetry, either. A tacit belief to the effect of 'all blameworthy actions are intentional' should lead participants to judge Bob's killing of the family intentional, as well—after all, it is blameworthy. Though I have suggested that postulating a tacit belief might avoid Mele's problem—the problem of explaining why presentation with a counterexample does not alter response to the Jake case—this second trouble with the Drunk Driver case remains. Fortunately, MMI's approach provides flexibility not available to the affect-bias account. MMI proposes a tacit interfering *mechanism*, with a vast body of information and reasoning strategies at its disposal; complex interactions between these tacit beliefs and strategies may explain the asymmetry observed in the Drunk Driver case. Indeed, the elaborated interference model I propose in the next section would predict just such an asymmetry.

4 Moral mechanism interference

A full defense of MMI requires a detailed explanation of the process by which interference occurs. However, this does not necessarily require adherence to any of the existing proposals for the structure of the moral mechanism *itself*. Although the Haidt and Greene accounts will not likely provide a good basis for an explanation of the intentionality data due to the Young et al. data challenging the role of emotion in producing the effects, it seems likely to me that either the Sripada and Stich account or a linguistic analogy account could easily fit with the interference model I wish to propose. I will, therefore, only offer the following three constraints on the structure of the moral mechanism. First, the operation of the mechanism must be largely tacit. Second, the mechanism must employ fairly sophisticated reasoning operations over a wide variety of situational inputs, many of which must come from a theory of mind mechanism. Third, the moral mechanism must not only provide blameworthiness judgments, but also praiseworthiness judgments. The Sripada and Stich account and the linguistic analogy account both meet the first two conditions; though neither discusses praise judgments, I see no in principle reason why either account could not be expanded appropriately.

With this in mind, let me put forth a fuller account of the interference process. Humans possess both a tacit theory of mind mechanism and a tacit moral judgment mechanism. In order to decide whether a certain action was blameworthy or praiseworthy, the moral judgment mechanism needs a lot of input information—much of which will be output information of the theory of mind mechanism. The moral mechanism needs to know quite a bit about the beliefs and desires of the agent, and quite a bit about the beliefs and desires of the recipient of the action, if there is one. Without this information, it won't be able to distinguish a boxing match from an unprovoked

violent attack. Presumably, one of the things the moral mechanism needs to know in order to dispense judgment is whether the agent brought about the particular consequences of his action *intentionally*.

Now, imagine a situation like Jake's unskilled shooting of his aunt in the immoral/no skill condition of the Jake cases. The input information the moral mechanism receives includes facts like the following—he wanted to shoot her, he tried to shoot her, she got shot, he shot her *unintentionally*.[vi] Since my claim is that the 'intentional' judgment is produced by interference, the *initial* output judgment of the theory of mind mechanism will not take moral information into account. Rather, intentionality ascription will likely resemble the model put forth by Malle and Knobe. Jake lacked the skill requirement, so the theory of mind mechanism judges his shooting to be unintentional. Once it has received all the inputs relevant to assessing the Jake case, the moral mechanism calculates the degree to which each input increases or decreases Jake's blameworthiness—we might imagine it assigning 'blameworthiness scores'. The moral judgment mechanism then applies various weights to each of these scores based on the importance (and perhaps certainty) of its corresponding input and computes a weighted average which serves as the total 'blameworthiness score' for the case. In the current case, most of the factors of the situation point to Jake's being blameworthy (i.e., they have high blameworthiness scores), except for one—the unintentional nature of the act, which is normally a highly mitigating circumstance.

This sole highly mitigating factor doesn't seem to fit with the rest of the data points, which are primarily strong exacerbating factors and which include Jake's *intent* and *desire* to kill his aunt.[vii] The 'unintentional' input is abnormally removed from the

rest of the data, in the sense that its 'blameworthiness value' as computed by the mechanism lies a large distance away from the values of the other inputs. As such, it is an 'outlier'. I propose that the moral mechanism, in the process of calculating the blameworthiness of this case, detects this outlier and overwrites the intentionality judgment, thereby bringing it in line with the rest of the data.[viii]

It is crucial to note that the moral judgment mechanism does not overwrite every piece of conflicting information; only highly removed outliers. This correction of suspicious inputs that do not fit the overall pattern is fairly analogous to the removal of outliers in statistical analysis, which potentially increases the accuracy of models constructed from the data.[ix] As such, it can be seen as an error-catching device; much as outlying data points in the results of a psychological experiment may indicate errors in procedure, outlying data points in information received by the moral mechanism may indicate some kind of 'misfire' in an earlier analysis.[x]

This proposed rewriting would also be useful for increasing justification for the moral judgment the mechanism has made. Assuming with Haidt ([2001]) that some post-hoc, conscious moral reasoning occurs whenever a person's moral judgment comes into question, uncorrected, strongly conflicting beliefs could cause cognitive dissonance or conscious reversals of the moral mechanism's judgments. The moral mechanism's overwriting feature works to prevent this. Similar methods for avoiding cognitive dissonance are commonplace in the psychological literature—for instance, 'choice-supportive bias' causes individuals reflecting on a choice to selectively remember the positive features of the option they chose while remembering only the negative features of the options they rejected.

By contrast, the purpose of a fundamentally evaluative theory of mind concept like the one Knobe proposes is less clear. The moral mechanism in the MMI account still gets to perform an objective evaluation of the data given to it by the theory of mind mechanism before determining whether or not to bias its inputs; Knobe's account, however, holds that moral judgment deals with non-objective information from the start. It is not clear how this strategy would aid evaluative tasks or increase the usefulness of moral judgments. Presumably information about the valence of the outcome is independently available in moral judgment—what is the advantage of building such information into a psychological concept?

We have seen how the proposed overwriting process explains the asymmetry in the Jake cases. In the achievement/no-skill condition, no morally relevant action occurs, and the theory of mind mechanism's 'unintentional' judgment stands. In the immoral/no-skill condition, however, the 'unintentional' judgment is an outlier in an otherwise strongly incriminating data set, and is overwritten. The same overwriting occurs in the inept but praiseworthy shooting of the communications device in the Klaus cases—recall that one of my constraints on a moral mechanism was that it be able to produce praise judgments as well as blame judgments.

The Chairman cases are similar. In the 'help' version, the chairman isn't the nicest individual, but he doesn't *do* anything particularly blameworthy. Presumably, the moral mechanism does evaluate the chairman's helping the environment, and judges that it is neither praiseworthy nor blameworthy. Since there is no conflict with the theory of mind mechanism's judgment that the chairman's helping was unintentional, no interference occurs. The case is judged unintentional. In the 'harm' version, the moral

mechanism evaluates the chairman's harming the environment. The act is blameworthy, but the 'unintentional' input is abnormally removed—so the moral mechanism alters the theory of mind mechanism's judgment to 'intentional'. All of this happens within a split second, because the operation of both of these systems is automatic, tacit, and *fast*—just like, for instance, linguistic reasoning. By the time the participant is ready to answer the test question, the belief he is left with is that the action was intentional.

MMI seems to fit well with the original data, and we have seen that it avoids many objections leveled against other performance accounts. As mentioned at the end of the previous section, however, MMI must still explain the results of the Drunk Driving case. Recall the vignette:

> 'Bob got rip-roaring drunk at a party after work. When the party ended, he stumbled to his car and started driving home. He was very drunk at the time—so drunk that he eventually lost control of his car, swerved into oncoming traffic, and killed a family of five' (Mele [2001], p. 41).

The problem here is that Bob's killing the family is blameworthy, but not intentional. Why doesn't interference occur in this case? First, it should immediately be clear that this case diverges sharply from both the Jake cases and the Chairman cases. In the immoral/no-skill version of the Jake case, Jake wanted to shoot his aunt, and had an intention to shoot his aunt. Here, Bob has neither a desire nor an intention to kill the family—in fact, he presumably has a desire to get home safely and *not* kill anyone. In the 'harm' version of the Chairman case, the chairman foresees that he will harm the

environment. Bob does not really *foresee* killing the family of five; he may understand that it is a possibility, but he certainly does not believe it to be more likely to occur than not.

So Bob does not have a desire or an intention to kill the family, nor does he foresee killing the family. Unlike the actions in Knobe's other cases, Bob's killing of the family of five does not in fact meet *any* of the prerequisites for intentionality put forth by Malle and Knobe ([1997]). In this case, then, the 'unintentional' judgment of the theory of mind mechanism is not really an outlier—it's not removed enough from the other inputs to prompt overwriting. Bob's lack of desire, intent, and foresight all count as mitigating factors, as well. In this case, the set of data as a *whole* is more ambiguous (though the weight of the evidence still points to guilt), but there is no single point that fails to fit the pattern. The eventual result is that the moral mechanism judges Bob to be blameworthy, but less so than a cold-hearted, calculating murderer.[xi] The outlier-based approach to MMI, then, predicts not only the original asymmetries, but also the unintentional-yet-blameworthy responses in the Drunk Driver case.

5 Blame or Valence?

The outlier model for the MMI account has been shown to avoid a number of specific objections leveled at previous performance accounts; however, there remains a final line of attack. As mentioned earlier, Knobe and Mendlow ([2004]) present arguments purporting to show that negative valence, not blame, is the factor which produces the asymmetries. The valence vs. blame question is in fact at the heart of the

debate between the performance and competence approaches. Performance accounts have almost invariably stressed the blameworthiness of the actions in the Jake and Chairman cases; the original discussions of skill-based asymmetries in the philosophical literature tended to emphasize moral features, as well. Knobe's competence account, however, denies this intuitive take on the phenomena.[xii]

Recall that Knobe takes it for granted that information about whether an action is intentional plays a central role in the assignment of praise and blame—indeed, Knobe's fundamental claim is that theory of mind concepts are tailored to evaluative tasks. For this very reason, a central feature of Knobe's account must be that the factor which causes asymmetry—i.e., which causes the concept to apply in one case but not the other - is the *non-moral* goodness or badness of the outcome. The valence approach is essential to Knobe's view—as we saw earlier, claiming that praiseworthiness or blameworthiness is doing the work leads to circularity. Praise/blame assessments would be an input to intentionality assessments which would be an input to praise/blame assessments.

Knobe and Mendlow's first argument, in fact, is that the affect-bias view and other blame-based views fall prey to this circularity—praise and blame must be assigned *before* determining whether an action is intentional, in conflict with the intuition that the concept of intentional action provides important input to moral judgment. However, it seems to me that any blame-based view which genuinely attributes the asymmetries to *performance* factors is not likely to encounter this circularity. According to MMI, for instance, a tacit theory of mind mechanism first makes intentionality judgments *without* any information from the moral mechanism. That information is then used by the moral mechanism in order to make moral judgments, and occasionally the moral mechanism

decides the information must be overwritten to maintain consistency. Such overwriting will be quite rare, occurring only when intentionality is a suspicious outlier in the moral mechanism's data. Unbiased intentionality information will most likely be used by the moral mechanism in the vast majority of cases.

Since Knobe offers a competence account, he cannot escape the circularity in the way I do, by claiming that praise and blame affect intentionality judgments only *after* moral judgments have been made on *unbiased,* purely non-evaluative information. He must instead deny that the evaluative factors relevant in application of the concept are moral. As such, if it turns out that praise and blame are the relevant features for the production of asymmetry in intentional action cases, we will have serious cause to doubt Knobe's account—performance based accounts, however, will emerge unscathed.

Knobe and Mendlow's second argument attempts to show that blame is *not* the relevant feature by providing an experimental scenario for which blame based accounts and Knobe's valence account will make diverging predictions. Participants were offered an analogue to the Chairman case, in which an agent knowingly brings about a side-effect which is bad but not blameworthy. The vignette they used, which I will call the 'New Jersey' case, runs as follows:

> 'Susan is the president of a major computer corporation. One day, her assistant comes to her and says, "We are thinking of implementing a new program. If we actually do implement it, we will be increasing sales in Massachusetts but decreasing sales in New Jersey." Susan thinks, "According to my calculations, the losses we sustain in New Jersey should be a bit smaller than the

> gains we make in Massachusetts. I guess the best course of action would be to implement the new program." "All right," she says. "Let's implement the program. So we'll be increasing the sales in Massachusetts and decreasing the sales in New Jersey."' (Knobe & Mendlow [2004], p. 257).

Participants were asked both whether Susan intentionally decreased sales in New Jersey and whether she deserved any praise or blame for doing so. Most participants judged that Susan deserved neither praise nor blame for her action, and yet 75% claimed she decreased sales in New Jersey intentionally. The results are thus in line with Knobe's predictions—the side effect was bad (but not blameworthy), and it was judged intentional. Blame-based accounts, on the other hand, only predict that side-effects will be judged intentional when they are blameworthy.

Despite the strength of many of Knobe's other arguments, I believe that this objection has missed the mark. For nowhere in Knobe and Mendlow's paper do they report running the analogue to the 'help' condition of the Chairman case. They do not present participants with a vignette in which Susan increases sales in *both* Massachusetts and New Jersey. Such a vignette would require a bit of modification, but might run as follows:

> Susan is the president of a major computer corporation. One day, her assistant comes to her and says, 'We are thinking of implementing a new program. If we actually do implement it, we will be increasing sales in Massachusetts. As a side-effect, we will

> also increase sales in New Jersey.' Susan thinks, 'According to my calculations, the gains we make in New Jersey should be a bit smaller than the gains we make in Massachusetts. But I see no downside. The best course of action would be to implement the new program.' 'All right,' she says. 'Let's implement the program. So we'll be increasing the sales in Massachusetts as well as increasing the sales in New Jersey.

Did Susan intentionally increase sales in New Jersey? The vignette comes across as a bit awkward even with the modifications, but my intuition is that Susan *did* intentionally increase New Jersey sales.

Why doesn't the asymmetry appear here? Perhaps because there is a crucial difference between the New Jersey cases and the original Chairman case—in both versions of the Chairman case, the chairman *didn't care about the environment*. Presumably, in the New Jersey case, Susan cares about whether or not sales decrease. Here, it will be useful to discuss a view put forth by Harman ([1976]). Harman's view is that we may judge an unintended action to be intentional so long as the agent acted *in the face of a reason not to*. He offers an example of the skill effect—if an unskilled sniper nonetheless succeeds in shooting his enemy, we would judge the shooting to be intentional. Harman claims that we do so because the sniper acts in the face of a *reason*—the fact that killing is wrong. Harman also, however, offers a non-moral example:

> ‘In firing his gun, the sniper knowingly alerts the enemy to his presence. He does this intentionally, thinking that the gain is worth the possible cost. But he certainly does not intend to alert the enemy to his presence’ (Harman [1976], p. 433).

Also, note an example from Bratman ([1984]):

> ‘Suppose I intend to run a marathon and believe that I will thereby wear down my sneakers. Now it seems to me that it does not follow that I intend to wear down my sneakers, and in a normal case I will not so intend… Even so, if I proceed to run the marathon and actually do wear down my sneakers then I might well do so intentionally’ (Bratman [1984], p. 400).

It seems that, if a side-effect is something the agent would prefer not to bring about, we regard bringing it about as intentional. My intuition is that the same applies to side-effects that the agent *would* prefer to bring about. Consider a variant of the sniper case in which the sniper’s shot will alert his ally to his presence. This is not the sniper’s immediate goal, but he is happy about the side-effect nonetheless. It seems reasonable to say that the sniper intentionally alerted his ally to his presence.[xiii]

It is not the case, however, that we ascribe intentionality to all side-effects by default. If the sniper knows that shooting will have the side-effect of heating the barrel of his gun, but he really doesn’t care one way or another about that, it seems odd to claim that he heats the barrel of his gun intentionally. Thus, my suggestion is that the default

strategy of the theory of mind mechanism is to judge side-effects to be unintentional, unless the agent cares about the side-effect—if the agent cares, the action is judged intentional.[xiv] Then, if the agent is deemed morally blameworthy for the side-effect, yet the theory of mind mechanism judged it unintentional, interference from the moral mechanism occurs. Thus, 'caring' asymmetries are competence-based, and 'blame' asymmetries are performance-based. Since caring is an objective mental state of an agent, this proposal is in harmony with the general hypothesis that commonsense psychology deals only with non-evaluative, objective information.

There is in fact some empirical evidence to suggest that caring is a crucial variable in Knobe's side-effect cases. Leslie, Knobe and Cohen ([2006]) ran a version of the side-effect effect cases on preschoolers in order to determine the age at which the effect emerges. In their version, a boy name Andy brings a frog over to Janine's house, which he knows will make her sad/happy. Andy is described as not caring whether Janine gets sad/happy. The children are then asked whether Andy made Janine sad/happy on purpose. As a control question, children were also asked whether Andy cared that Janine will get sad/happy. Results indicated that most four and five year olds exhibited the side-effect effect; more interestingly, however, failing the 'caring' control question was highly correlated with failing to exhibit the side-effect effect. 'As soon as preschoolers effectively process the theory-of-mind concept 'not care that P', children show the side-effect effect' (Leslie et al. [2006], p. 421). Prior to acquiring the not-care concept, there is a general tendency to say that the action was 'on purpose' in both cases.

Given these considerations, it seems clear that Knobe and Mendlow's New Jersey vignette does not provide a decisive test of blame-based accounts. Since the experiment

lacked an adequate control condition, we cannot rule out the hypothesis that the caring feature is doing the work. The claim that we normally judge all side-effects that the agent cares about to be intentional explains the New Jersey case in a manner which is perfectly compatible with the hypothesis that blame processing, not valence, produces the asymmetries of the Jake and Chairman cases. And, given the intuitiveness of the idea that the blameworthiness of the action is the relevant factor, it seems that the burden of proof is still on Knobe to show that valence is producing the effects.

The only other attempt I am aware of to test a side-effect case with bad but not blameworthy consequences is Knobe ([2004a]). Here, Knobe attempts to construct an aesthetic variant of the side-effect cases.

> 'The Vice-President of a movie studio was talking with the CEO. The Vice-President said: "We are thinking of implementing a new policy. If we implement the policy, it will definitely increase profits for our corporation, but it will also make our movies worse from an artistic standpoint." The CEO said: "Look, I know that we'll be making the movies worse from an artistic standpoint, but I don't care one bit about that. All I care about is making as much profit as I can. Let's implement the new policy!" They implemented the policy. As expected, the policy made the movies worse from an artistic standpoint.' (Knobe [2004a], p. 274).

In the second condition, 'worse' was replaced with 'better', such that the CEO didn't care about making the movies better artistically. Responses to the vignettes showed the

classic asymmetry, from which Knobe concludes that goodness/badness, not blameworthiness, is the relevant factor.

It is not clear to me, however, that Knobe's case is not subtly moral. There certainly seems to be something immoral about attempting to dupe the public into buying a shoddy product. In addition, we might imagine that the CEO in this case is shutting down the artistic ambitions of the film-makers, forcing them to produce substandard films. At the very least, the act seems selfish—I certainly would not trust this CEO further than I could throw him.

Existing attempts to show that valence causes the skill and side-effect asymmetries are inconclusive at best. This is good news for the blame-based approach. What the defender of a blame-based view really needs, however, is a case in which Knobe's valence account must predict an effect, yet one is absent. Such a case, fortunately, exists. First, consider Knobe's own Klaus case—in particular, the morally-good/no-skill version. In that vignette, Klaus makes a lucky shot destroying a communications device and saving innocent lives. Participants readily said that Klaus intentionally destroyed the device, thus disregarding the usual skill prerequisite. This alone presents no problem; Knobe can claim that both good outcomes and bad outcomes cause subjects to disregard skill. However, making this move leads to trouble with Knobe's Jake cases. In the achievement/no-skill version of the Jake story, Jake makes a lucky shot, which wins a rifle contest. Winning a rifle contest is a good outcome. But the vast majority of participants do not disregard skill and deem his winning to be intentional. The features in the Klaus and Jake cases are identical—desire, intent, lack of skill, good outcome, etc.—except for the praiseworthiness of Klaus's act. Knobe's

valence account can not explain why we judge the one intentional and the other unintentional. Blame-based accounts like MMI can.

Though the evidence is not yet decisive, I believe it warrants the assumption that the feature producing the asymmetry phenomena is in fact blameworthiness (or praiseworthiness). As previously noted, this leaves the competence account in an awkward situation. The claim of the competence account is that the asymmetry appears because some relevant feature of the case is a fundamental factor in the application of the concept 'intentional'—if this fundamental factor is praiseworthiness or blameworthiness, it becomes difficult to understand how the concept 'intentional' could serve as an input to the assessment of that same factor. Performance accounts, by their very nature, have the resources to avoid this problem.

It is possible that Knobe's account could be altered in some other way to avoid the troublesome circularity. It is worth noting, therefore, that this is not the only issue on which performance accounts enjoy a clear advantage. More trouble arises for the competence approach when one considers data showing that similar moral asymmetry effects occur with judgments of *causation*.[xv] Knobe and Fraser ([forthcoming]) offered participants the following vignette:

> 'The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take the pens, but faculty members are supposed to buy their own. The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist has repeatedly emailed them reminders that only administrative

> assistants are allowed to take the pens. On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later that day, the receptionist needs to take an important message… but she has a problem. There are no pens left on her desk' (Knobe & Fraser [forthcoming]).

Participants are then asked how much they agreed with two statements: 'Professor Smith caused the problem' and 'the administrative assistant caused the problem'. Participants tended to judge that Professor Smith caused the problem, and that the administrative assistant did not. Of course, in purely objective terms, both caused the problem equally. Thus, the moral features of the vignette seem to be biasing the participants' responses.

The above findings could plausibly be due to a quirk in how participants interpret 'caused the problem'—maybe this carries some implication that the question is about who is to blame. However, further studies in this vein have shown similar biasing effects. Alicke ([1992]) presented participants with vignettes about a man named John who was driving over the speed limit and hit another car at an intersection. The vignettes cited one of two reasons for John's speeding—in one condition, he was speeding home to hide an anniversary present for his parents. In the other condition, he was speeding home to hide a vial of cocaine. The vignettes each also specified another possible cause of the accident—either an oil spill, a tree branch in the road, or the other car running a stop sign. Participants were asked to give the primary cause of the accident. Participants cited John

as the cause much more frequently when his motive was hiding the vial of cocaine; thus, John's blameworthiness seems to be biasing causation intuitions.[xvi]

Knobe ([2005]) explains the phenomenon by arguing that *both* the concept of intentional action *and* the concept of causation were fundamentally shaped by the need to answer blame questions. This approach does indeed fit with Knobe's overall view, but it is not particularly parsimonious—Knobe is required to claim that two separate mechanisms (theory of mind and folk causation) happened to evolve similar sensitivities to valence information in order to help out with blame processing. Such parallel evolution is of course possible, but would perhaps be somewhat surprising. On the other hand, performance accounts can approach the causation phenomena in a simpler way—the very same biasing factor simply operates in two different domains. In the case of MMI, this is particularly plausible. Causation information might well be part of the input evaluated by the moral mechanism, and could thus produce outlying data points as well. No further emendation to MMI is required. The very same process can explain interference effects in a variety of morally relevant fields.

I have no argument to conclusively disprove Knobe's assertion that the concept 'intentional' invokes evaluative notions. However, I believe the MMI account explains the available data at least as well, if not better—particularly in that it easily allows for blame to be the underlying factor causing the asymmetries, and in that its explanation for the intentionality phenomena can be applied equally well to the causation phenomena. Further, the MMI account fits neatly with current views in the psychological and philosophical literature; views hypothesizing tacit mechanisms for theory of mind or for moral cognition are commonplace, as are accounts of biasing phenomena aimed at

reducing cognitive dissonance. By contrast, I am not aware of any work other than Knobe's that argues for fundamentally evaluative folk psychological concepts, and I am not convinced that building evaluative criteria into such concepts would provide much assistance in blame processing. With these considerations in mind, I believe we are justified in preferring the MMI account and retaining our commitment to a non-evaluative folk psychology.

Acknowledgements

I am grateful to Stephen Stich, Joshua Knobe, and the referees of BJPS for invaluable help with earlier versions of this paper.

Department of Philosophy

Rutgers University

26 Nichol Avenue

New Brunswick, NJ 08901, USA

nado@philosophy.rutgers.edu

References

Adams, F., & Steadman, A. [2004a]: 'Intentional action and moral considerations: still pragmatic', *Analysis,* **64**, pp. 268-276.

Adams, F., & Steadman, A. [2004b]: 'Intentional action in ordinary language: core concept or pragmatic understanding?', *Analysis,* **64**, pp. 173-181.

Alicke, M. [forthcoming]: 'Blaming Badly', *Journal of Cognition and Culture*.

Bratman, M. [1984]: 'Two faces of intention', *The Philosophical Review,* **93**, pp. 375-405.

Dwyer, S. [1999]: 'Moral competence', In K. Murasugi & R. Stainton (eds), *Philosophy and Linguistics*, Boulder, CO: Westview Press, pp. 169-190.

Greene, J. [2001]: 'An fMRI investigation of emotional engagement in moral judgment', *Science,* **293**, pp. 2105-2108.

Haidt, J. [2001]: 'The emotional dog and its rational tail: a social intuitionist approach to moral judgment', *Psychological Review,* **108**, pp. 814-834.

Harman, G. [1976]: 'Practical reasoning', *Review of Metaphysics,* **29**, pp. 431-463.

Harman, G. [1999]: 'Moral philosophy and linguistics', In K. Brinkmann (ed.), *Proceedings of the 20th World Congress of Philosophy, Volume 1: Ethics,* Bowling Green, Ohio: Philosophy Documentation Center, pp. 107-115.

Hauser, M. [2006]: *Moral minds: the unconscious voice of right and wrong*, New York: Harper Collins.

Knobe, J. [2003a]: 'Intentional action and side-effects in ordinary language', *Analysis,* **63**, pp. 190-193.

Knobe, J. [2003b]: 'Intentional action in folk psychology: an experimental investigation', *Philosophical Psychology*, **16**, pp. 309-324.

Knobe, J. [2004a]: 'Folk psychology and folk morality: response to critics', *Journal of Theoretical and Philosophical Psychology*, **24**.

Knobe, J. [2004b]: 'Intention, intentional action, and moral considerations', *Analysis,* **64**, pp. 181-187.

Knobe, J. [2005]: 'Cognitive processes shaped by the impulse to blame', *Brooklyn Law Review,* **71**.

Knobe, J. [2006]: 'The concept of intentional action: a case study in the uses of folk psychology', *Philosophical Studies,* **130**, pp. 203-231.

Knobe, J., & Fraser, B. [forthcoming]. 'Causal judgment and moral judgment: two experiments', In W. Sinnott-Armstrong (ed.), *Moral Psychology*, Cambridge, MA: MIT Press.

Knobe, J., & Mendlow, G. [2004]: 'The good, the bad, and the blameworthy: understanding the role of evaluative reasoning in folk psychology', *Journal of Theoretical and Philosophical Psychology,* **24**, pp. 252-258.

Leslie, A., Knobe, J., & Cohen, A. [2006]: 'Acting intentionally and the side-effect effect: theory of mind and moral judgment', *Psychological Science,* **17**, pp. 421-427.

Malle, B., & Knobe, J. [1997]: 'The folk concept of intentionality', *Journal of Experimental Social Psychology,* **33**, pp. 101-121.

Malle, B., & Nelson, S. [2003]: 'Judging mens rea: the tension between folk concepts and legal concepts of intentionality', *Behavioral Sciences and the Law,* **21**, pp. 563-580.

Mele, A. [2001]: 'Acting Intentionally: Probing Folk Notions', In B. Malle, L. Moses, and D. Baldwin (eds), *Intentions and Intentionality: Foundations of Social Cognition*, Cambridge, MA: MIT Press.

Mele, A. [2003]: 'Intentional action: controversies, data, and core hypotheses', *Philosophical Psychology,* **16**, pp. 325-340.

Nadelhoffer, T. [2004]: 'On praise, side effects, and folk ascriptions of intentionality', *Journal of Theoretical and Philosophical Psychology,* **24**, pp. 196-213.

Rawls, J. [1971]: *A theory of justice*, Cambridge, MA: Harvard University Press.

Sripada, C., & Stich, S. [2006]: 'A framework for the psychology of norms', In P. Carruthers, S. Laurence & S. Stich (eds), *The Innate Mind, Vol. II, Culture and the Innate Mind*, Oxford: Oxford University Press.

Thagard, P., & Verbeurgt, K. [1998]: 'Coherence as constraint satisfaction', *Cognitive Science,* **22**, pp. 1-24.

Young, L., Cushman, F., Adolphs, R., Tranel, D., & Hauser, M. [2006]: 'Does emotion mediate the effect of an action's moral status on its intentional status? Neuropsychological evidence', *Journal of Cognition and Culture,* **6**, pp. 291-304.

[i] In fact, such an assumption is quite crucial to Knobe's account. The received view is that commonsense psychology is a tool for predicting and explaining behavior. As mentioned, this view predicts that all information relevant to such a task would be non-evaluative. In order to grant any plausibility to the proposal that the concept 'intentional' incorporates evaluative information, Knobe must claim that the concept plays a central role in evaluative tasks like blame assessment.

[ii] Goodness and badness also, of course, directly influence assessment of praise and blame.

[iii] A third important type, pragmatic accounts, will not be discussed here. I am inclined to agree with Knobe's arguments against that view, and since those arguments do not bear on my proposal, I will not restate them. For an overview see Adams and Steadman ([2004b]). For Knobe's arguments against the account, see Knobe (2004b). See Adams and Steadman ([2004a]) for a reply.

[iv] I have in mind here cases such as the following, which is due to an anonymous reviewer—a student in an introductory philosophy class is able to state on a test what is wrong with the cosmological argument, yet when prompted to justify her belief in God offers the cosmological argument in some form or other. The form in which she presents the argument may be quite different from the form in which it was discussed in class; the counterexample, therefore, may not affect the former belief.

[v] Alicke ([forthcoming]) offers an account which seems to fall somewhere between the affect-bias account and the account I propose. His main claim is that spontaneous negative evaluations prompted by the side-effect cases trigger 'blame-validation' processing, in which morally relevant information is biased in order to justify blame attribution. This would place him squarely in the affect-bias camp, except for the fact that he explicitly states that the spontaneous evaluations need not be emotional. Thus, Alicke may evade the Young et al. data. However, I am not clear as to whether the account evades Knobe and Mendlow's blame-based arguments, or how it deals with the drunk-driving example. It is possible that elaboration of the proposal would successfully avoid these objections; I suspect the necessary elaborations would in fact place Alicke's view quite close to my own.

[vi] It's possible that the last bit of information would be 'did not shoot her intentionally'. 'Unintentionally' seems to carry some extra information, implying that Jake did not want or intend to shoot her aunt. In fact, it seems wrong to say Jake unintentionally won the contest in the achievement/no-skill case.

[vii] An anonymous reviewer raised the following concern—why should intent and desire be inputs both to the assessment of intentionality and to the assessment of blameworthiness? Given that the moral mechanism already receives information about intentionality, isn't information about intent and desire redundant? This is a very valid worry, but I do believe a plausible defense of the supposed redundancy can be offered. Firstly, there are several inputs to the assessment of intentionality that are not inputs to the assessment of blameworthiness, notably Malle and Knobe's skill requirement, and perhaps their awareness and belief requirements as well. Secondly, the moral relevance of desire and intent go beyond their contributions to the assessment of intentionality, and cannot necessarily be read off of information about intentionality. For instance, in the Chairman case, the chairman intentionally harms the environment, though he has no particular desire to do so—it is merely a side-effect. If he did actively desire to harm the environment, his action would still be intentional, but he would be even more blameworthy. The situation is a bit like the following, which is an adaptation of an example by the same reviewer—you are an employer, and you wish to assess the academic achievement of a potential employee. The applicant's combined SAT test score will be quite valuable to your assessment. However, you also consider reading comprehension to be particularly relevant to your assessment. Though the SAT contains a number of reading comprehension questions, the applicant's performance on these questions cannot be determined merely by looking at his SAT score. You could request a copy of the entire test, but you don't need to see his answers on every question. What would be *most* valuable would be a report of his performance on the reading comprehension questions alone.

[viii] The details of the blameworthiness calculation process and of the outlier detection procedure are relatively inessential to the model—the specific 'weighted mean' approach to the former discussed here, for instance, is only one possible method and is only offered for illustrative purposes.

[ix] This correction process also bears interesting similarities to the 'constraint satisfaction' process discussed by Paul Thagard and others (see for instance Thagard and Verbeurgt

[1998]); however, there are significant differences as well. A full comparison would be valuable, but is beyond the scope of this paper.

[x] It is important to note that the 'error-catching' being proposed is essentially subjective error-catching. The idea is that the moral mechanism corrects inputs that suggest that the theory of mind mechanism is not operating normally. Of course, this leaves open the question of whether the 'normal' operation of the theory of mind mechanism succeeds in accurately characterizing the external world.

[xi] It is also worth noting that the risk of cognitive dissonance in this case is somewhat decreased because the killing was preceded by the uncontroversially intentional, uncontroversially blameworthy act of drunk driving. In fact, most of the blame in the killing is in a way 'inherited' from the blameworthiness of the drunk driving, in the sense that, had the fatal swerving of the car been caused by an epileptic seizure rather than drunkenness, Bob's blameworthiness would evaporate.

[xii] Knobe no longer believes that valence produces the observed asymmetries (though I am not sure he has said so in print yet). However, the exploration of the blame vs. valence question put forth in this section is still highly relevant, because it is my contention that, in order to hold a plausible competence account, Knobe *must* endorse the valence line. My further contention that the valence line is unsuccessful thereby casts doubt on competence accounts generally.

[xiii] Admittedly, it is hard to get a grasp on these cases. If the sniper knows he will alert his ally by shooting, and if he desires that his ally be alerted, and if he was going to shoot anyway, it may be that he thereby forms an intention to alert his ally (in contrast with Harman's suggestion in the previous case that the sniper does *no*t intend to alert his enemy but *does* do so intentionally). My intuitions are far from clear. However, regardless of whether the sniper forms an intention, it certainly seems wrong to say that the sniper *did not intentionally alert his ally*. It also seems wrong to say that Susan did not intentionally increase sales in New Jersey. It would of course be desirable to see these intuitions backed up by experimental results. But the burden of proof is on Knobe to show that his 'sales in New Jersey' case is a true analogue to the original chairman case—as stated, it has no control condition to demonstrate that some other feature particular to the case (e.g., caring) is not producing the effect. For the findings to be valid, a version of the 'help' condition must be run.

[xiv] Of course, a cared-about side effect will only be judged intentional if the other prerequisites for intentionality (whatever those are—I suspect Malle and Knobe's analysis is fairly accurate) are met.

[xv] This line of argument was suggested by Knobe (personal communication).

[xvi] Alicke ([1992]) also contains three other intriguing experiments which support this general trend.