

**Strawsonian Variations:
Folk Morality and the Search for a Unified Theory**

Joshua Knobe

*University of North Carolina-
Chapel Hill*

John M. Doris

*Washington University,
St. Louis*

Much of the agenda for contemporary philosophical work on moral responsibility was set by Strawson's (1962) essay 'Freedom and Resentment.' In that essay, Strawson suggests that we focus not so much on metaphysical speculation as on understanding the actual practice of moral responsibility judgment. The hope is that we will be able to resolve the apparent paradoxes surrounding moral responsibility if we can just get a better sense of how this practice works and what role it serves in people's lives.

Many of the philosophers working on moral responsibility today would disagree with some of the substantive conclusions Strawson reached in that early essay, but almost all have been influenced to some degree by his methodological proposals. Thus, almost all participants in the contemporary debate about moral responsibility make some appeal to the ordinary practice of moral responsibility judgment. Each side tries to devise cases in which the other side's theory yields a conclusion that diverges from people's ordinary judgments, and to the extent that a given theory actually is shown to conflict with ordinary judgments, it is widely supposed that we have strong reason to reject the theory itself.

It seems to us that this philosophical effort to understand the ordinary practice of moral responsibility judgment has in some ways been a great success and in other ways a dismal failure. We have been extremely impressed with the ingenuity philosophers have shown in constructing counterexamples to each other's theories, and we think that a number of participants in the debate have been successful in coming up with cases in which their opponents' theories yield conclusions that conflict with ordinary judgments. But we have been less impressed with attempts to actually develop theories that accord

with ordinary judgments. It seems that each side has managed to show that the other falls prey to counterexamples. The result is a kind of mutual annihilation or, as Fischer (1994: 83-5) calls it, a ‘dialectical stalemate.’

We want to offer a diagnosis for this persistent difficulty. We suggest that the problem can be traced back to a basic assumption that has guided almost all philosophical discussions of moral responsibility. The assumption is that people should apply the *same* criteria in *all* of their moral responsibility judgments. In other words, it is supposed to be possible to come up with a single basic set of criteria that can account for all moral responsibility judgments in all cases — judgments about both abstract questions and concrete questions, about morally good behaviors and morally bad behaviors, about the behaviors of one’s close friends and the behaviors of complete strangers. It is supposed to be completely obvious, and hence in need of no justification or argument, that we ought to apply the same criteria in all cases rather than applying different criteria in different cases. This assumption is so basic that it has never even been given a name. We will refer to it as the assumption of *invariance*.

Of course, the best way to test an assumption about people’s ordinary judgments is to go out and actually run systematic experimental studies. Recent years have seen a number of such studies, and the results appear to be converging on a surprising new hypothesis. It seems that people simply do not make moral responsibility judgments by applying invariant principles. Instead, it appears that people tend to apply quite different criteria in different kinds of cases. Thus, if we want to understand why people have the judgments they do, it is no use looking for a single basic set of criteria that fits all of people’s ordinary judgments. A more promising approach would be to look at how a variety of factors can lead a person to adopt different criteria in different cases, depending on how the question is phrased, whether the agent is a friend or a stranger, and so on.

This discovery in empirical psychology leaves us with a stark choice in moral philosophy. One option would be to hold on to the goal of fitting with people’s ordinary judgments and thereby abandon the assumption of invariance. The other would be to hold on to the assumption of invariance and thereby abandon the goal of fitting with people’s ordinary judgments. But it seems that one cannot have it both ways. As we shall see, a

growing body of experimental results points to the view that it is not possible to capture all of people's ordinary judgments with a theory that applies the very same criteria in every case.

Invariantist Theories in Philosophy

We begin by briefly reviewing some of the major invariantist theories of moral responsibility. Our aim here is not to represent individual theories in detail, but simply to introduce some of the themes that we will be discussing in more depth in the sections to come.¹

Before reviewing the major theories, however, a few words are in order about what we mean by 'invariantism.' A theory counts as invariantist if it applies the same basic criteria in all cases where people are making moral responsibility judgments. These criteria can be complex; they can be vague; they can be based on prototypes or paradigms rather than lists of necessary conditions. All that matters is that they be the *same* criteria in every case. Thus, an invariantist theory might say:

- (1) 'No matter who we are judging, no matter what the circumstances are, always make moral responsibility judgments by checking to see whether the agent meets the following criteria...'

By contrast, it would be a rejection of invariantism to say:

- (2) 'If the agent is a friend, use the following criteria..., but if the agent is a stranger, use these other, slightly different criteria...'

Of course, when we say that a rule is not invariantist, we do not mean to imply that it does not involve following any principle at all. Rule (2) does give us a definite principle; it's just that this principle does not have the property of being invariantist. It does not tell us to apply the very same criteria in all cases. Rather, it tells us to apply one set of criteria to friends and another, slightly different set of criteria to strangers.

Now, to really explain what it is for a rule to be invariantist, we would have to say more precisely what it means to apply the very same criteria in all cases. Needless to say, this would be a tricky task. It seems that there will be occasions on which we can see

¹ For more comprehensive reviews, see Eshleman (2004) and Fischer (1999).

clearly that a given approach either applies the same criteria to all cases or different criteria in different cases, but there will probably also be occasions on which an approach falls somewhere in the unclear borderline between these two categories and we don't end up knowing quite what to say. Moreover, there is the problem that one can use certain cheap logical tricks to make just about any rule look invariantist.² These are difficult problems, and philosophers of science have been wrestling with them for decades. But here, as so often, we think it is possible to make important philosophical progress without first stepping into the swamp of technicalities necessary to 'define one's terms.' The best way to make it clear what counts as an invariantist theory is just to take a look at a few of the major theories from the existing philosophical literature. None of these theories make any use of cheap logical tricks. All of them really do apply the same conditions in all cases. Our approach, then, will be to describe a few of the most influential philosophical theories and then ask whether it is really possible to capture ordinary responsibility judgments using the invariantist approach they all share.

Of the theories we will be discussing, the oldest and most well known is *incompatibilism* (e.g., Kane 1996; Pereboom 2001; van Inwagen 1983). The essence of incompatibilism is the claim that moral responsibility is incompatible with determinism. In other words, incompatibilists say that an agent can never be morally responsible for a behavior if that behavior was brought about deterministically consequence of certain initial conditions and physical laws. For example, incompatibilists endorsing the Principle of Alternate Possibilities insist that the agent is responsible *only if* she could have done otherwise, a condition allegedly incompatible with causal determinism, and these incompatibilists are committed to the view that in *no case* where the Principle does not hold can the agent be legitimately attributed responsibility (see Doris and Stich 2005). Recent years have seen an increasingly sophisticated debate about whether the

² Initially, one might think that it is possible to explain what makes a rule like (2) turn out not to be invariantist just by adverting to its logical form. But things are not quite that simple. Thus, suppose we define the predicate *franger* by saying that a person is 'franger' if she is either a friend who meets criterion *x* or a stranger who meets criterion *y*. Then someone could say: 'I apply the same basic criterion to all cases. No matter who a person is, I always determine whether or not she is morally responsible by checking to see whether she is franger.' In such a case, it seems clear that the person is *not* really applying the same criterion in each case — he is applying one criterion to friends, another to strangers — but it has proved extraordinarily difficult to say precisely how definitions like this one differ from definitions that do legitimately apply the same criteria to every case (see, e.g., Goodman 1954).

incompatibilist thesis is truly warranted. We will not be discussing the nuances of that debate here. Instead, we simply want to emphasize that incompatibilism is an invariantist view. Its chief claim is that moral responsibility is *always* incompatible with determinism. No incompatibilist we know of suggests that the incompatibilist thesis might only apply to very close friends or that it might only apply to certain particular types of behaviors. Rather, the thesis is that, for all possible behaviors and all possible contexts, moral responsibility is incompatible with determinism.

Those who reject the incompatibilist thesis are known as *compatibilists*. Thus, compatibilists say that it is possible for a person to be morally responsible even in a deterministic universe. But compatibilists are no less invariantist than the incompatibilists they argue against. Compatibilists typically say that determinism is *never* relevant to moral responsibility in any way. They then put forward some other invariant principle that is supposed to serve as a criterion for moral responsibility judgments in all possible contexts. A wide variety of such criteria have been proposed. Here we will be concerned with two of the most popular.

First, the *real self view*.³ The key claim behind this view is that people are only morally responsible for behaviors that stem from a particular part of the self. Hence, it may be suggested that we are only morally responsible for behaviors that stem from the part of the self with which we are ‘identified’ (Frankfurt: 1988 53-4) or that we are only responsible for behaviors that stem from our values (Watson 1975). Proponents of this view sometimes suggest that there may be a number of different senses of the word ‘responsible’ (Watson 1996) — which may mean that conversational context is used to determine which sense is most relevant. However, it is usually assumed that each sense of the word is governed by a single invariant standard.

The other major compatibilist position might be called the *normative competence theory*. This theory says that people are only morally responsible for behaviors that were produced by a process that is appropriately sensitive to reasons (Fischer & Ravizza 1998; Wolf 1998). Proponents of this theory have been especially explicit in claiming that a single basic criterion can be applied to all possible behaviors. Indeed, much of the

³ The two approaches need not be incompatible. Doris (2002: Ch. 7) considers a position that appears to combine elements of both the normative competence and real self views.

intellectual excitement surrounding the normative competence theory stems from the ingenious ways in which researchers have been able to derive an apparently diverse array of moral responsibility judgments from an underlying principle that is extremely simple and unified.

Much of the debate between these rival views relies on appeals to ordinary judgments. Each side tries to come up with cases in which people's judgments conflict with the conclusions that follow from the other side's theory. So, for example, incompatibilists try to devise cases in which people would ordinarily say that an agent is not morally responsible for her behavior but in which the major compatibilist positions (real self, normative competence, etc.) all yield the conclusion that she actually is responsible (e.g., Pereboom 2001). Conversely, compatibilists try to find cases in which people would ordinarily say that an agent is morally responsible but in which all of the major incompatibilist positions yield the conclusion that she is not (e.g., Frankfurt 1969).

Our claim is that this is a conflict in which both sides endure unacceptable casualties. That is, each side can show that the others conflict with ordinary judgments in certain kinds of cases. The problem, we suggest, is that people simply do not have invariant criteria for making moral responsibility judgments. Thus, whenever a theory offers an invariant criterion, it will be possible to come up with a case in which people's judgments conflict with conclusions that can be derived from the theory. If we really want to understand people's ordinary judgments, we need to abandon the search for invariant criteria and try instead to examine the ways in which people end up using different criteria in different cases.

Here we will be concerned with three kinds of factors that appear to affect the criteria people use — the abstractness or concreteness of the question, the moral status of the behavior itself, and the relationship between the person making the judgment and the agent being judged.

Abstract vs. Concrete

It is essential to distinguish between different ways of checking to see whether a given principle is in accord with ordinary people's judgments. One approach would be to

present people with an explicit statement of the principle itself and ask them whether or not they agree with it. Another would be to look at people's judgments regarding particular cases and see whether these judgments fit the criteria derived from the principle. The usual view seems to be that both of these approaches are relevant when we are evaluating proposed criteria for moral responsibility.

One of the chief lessons of contemporary cognitive science, however, is that these two approaches quite often lead to different conclusions. It can easily happen that the principles people put forward in abstract conversations have almost nothing to do with the criteria they actually use when considering concrete cases. Thus, most linguists agree that people's grammatical intuitions are based on a complex competence that is almost entirely unrelated to the principles people are able to cite when asked explicitly. Similarly, social psychologists have uncovered numerous factors that appear to influence people's judgments in concrete cases but which people regard as irrelevant when asked in the abstract (e.g., Wicker 1969). There is every reason to expect that judgments about moral responsibility will show a similar pattern. That is, one should expect to find that people's judgments in concrete cases do not match up perfectly with the principles they espouse in more abstract discussions.

One particularly striking example arises in the debate over whether ordinary people are compatibilists or incompatibilists. The most common view among philosophers has always been that most people have strongly incompatibilist inclinations:

Beginning students typically recoil at the compatibilist response to the problem of moral responsibility. (Pereboom 2001: xvi)

... we come to the table, nearly all of us, as pretheoretic incompatibilists. (Ekstrom 2002: 310)

In my experience, most ordinary persons start out as natural incompatibilists... Ordinary persons have to be talked out of this natural incompatibilism by the clever arguments of philosophers (Kane 1999).

When ordinary people come to consciously recognize and understand that some action is contingent upon circumstances in an agent's past that are beyond that agent's control, they quickly lose a propensity to impute moral responsibility to the agent for that action. (Cover & O'Leary-Hawthorne 1996: 50)

Clearly, these are empirical claims, but it has traditionally been assumed that there is no need to test them using systematic experimental techniques. After all, philosophers are continually engaged in a kind of informal polling.⁴ They present material in classes and listen to how their students respond. What they typically find, it seems, is that students lean strongly toward incompatibilism, and this is customarily taken to indicate that folk morality, as practiced outside the confines of philosophy classrooms is itself incompatibilist. But, strange as it may seem, when researchers began examining these questions more systematically, their results did not confirm the claims that philosophy professors had been making about their students. In fact, the results pointed strongly in the opposite direction. People's ordinary judgments appeared to be strongly *compatibilist*.

The first study to arrive at this surprising conclusion was conducted by Viney and colleagues (1982; 1988). The researchers used an initial questionnaire to distinguish between subjects who believed that the universe was deterministic and those who did not. All subjects were then given questions in which they were given an opportunity to provide justifications for acts of punishment. The key finding was that determinists were no less likely than indeterminists to offer retributivist justifications. This finding provided some initial evidence that most determinists were predominately compatibilists.

Woolfolk, Doris and Darley (forthcoming) arrived at a similar conclusion using a radically different methodology. They ran a series of experiments in which subjects were given short vignettes about agents who operated under high levels of constraint. In one such vignette, a character named Bill is captured by terrorists and given a 'compliance drug' to induce him to murder his friend:

Its effects are similar to the impact of expertly administered hypnosis; it results in total compliance. To test the effects of the drug, the leader of the kidnappers shouted at Bill to slap himself. To his amazement, Bill observed his own right hand administering an open-handed blow to his own left cheek, although he had no sense of having willed his hand to move. The leader then handed Bill a pistol with one bullet in it. Bill was ordered to shoot Frank in the head...

The researchers then manipulated the degree to which the agent was portrayed as *identifying* with the behavior he has been ordered to perform. Subjects in one

⁴ Jackson (1998) is a rare philosopher who makes this methodology explicit. For the difficulties Jackson faces, see Doris and Stich (2005)

condition were told that Bill did not want to kill Frank; those in the other condition were told that Bill was happy to have the chance to kill Frank. The results showed that subjects were more inclined to hold Bill morally responsible when he identified with the behavior than when he did not. In other words, people assigned more responsibility when there were higher levels of identification *even though the agent's behavior was entirely constrained*. The study therefore provides strong evidence for the view that people are willing to hold an agent morally responsible for a behavior even when that agent could not possibly have done otherwise.⁵

Of course, these results do not bear *directly* on questions about moral responsibility and determinism — it takes a theoretical *inference* to go from experimental data on responsibility attributions to the conclusion that compatibilist leanings underlie such judgments. However, when a similar method is used to ask subjects for their reactions to cases that directly involve determinism, the results come out much the same way. Nahmias, Morris, Nadelhoffer and Turner ran a series of experiments in which subjects were given stories about agents who performed immoral behaviors in deterministic worlds. Subjects were then asked to say whether these agents were morally responsible for what they had done. In one such experiment, subjects were given the following case:

Imagine that in the next century we discover all the laws of nature, and we build a supercomputer which can deduce from these laws of nature and from the current state of everything in the world exactly what will be happening in the world at any future time. It can look at everything about the way the world is and predict everything about how it will be with 100% accuracy. Suppose that such a supercomputer existed, and it looks at the state of the universe at a certain time on March 25th, 2150 A.D., twenty years before Jeremy Hall is born. The computer then deduces from this information and the laws of nature that Jeremy will definitely rob Fidelity Bank at 6:00 PM on January 26th, 2195. As always, the supercomputer's prediction is correct; Jeremy robs Fidelity Bank at 6:00 PM on January 26th, 2195.

Subjects were then asked whether Jeremy was morally blameworthy. The vast majority (83%) said yes, indicating that they thought an agent could be morally blameworthy even

⁵ A manipulation check by Woolfolk et al (forthcoming) indicated that subjects, quite sensibly, recognized the strength of the “compliance drug” constraint.

if all of his behaviors were determined by natural laws. The researchers conducted three experiments — using three quite different ways of explaining determinism — and always found a similar pattern of responses.

Looking at these results, it may seem mysterious that any philosophers could have thought that ordinary people were incompatibilists. How could philosophers have concluded that people were incompatibilists when they so consistently give compatibilist answers in systematic psychological studies? Are philosophers just completely out of touch with what their undergraduates really think? We suspect that something more complex is going on: perhaps people tend to give compatibilist answers to *concrete* questions about particular cases but incompatibilist answers to *abstract* questions about general moral principles. Then the divergence between the findings from psychological studies and the conclusions of philosophers teaching classes might simply be due to a difference between two ways of framing the relevant question.

To test this hypothesis, Nichols and Knobe (2005) ran a simple questionnaire study. All subjects were given a story about a universe ('Universe A') in which events always unfold according to deterministic laws. Subjects in the 'abstract condition' were then given the question:

In Universe A, is it possible for a person to be fully morally responsible for their actions?

Subjects in the 'concrete condition' were given the question:

In Universe A, a man named Bill has become attracted to his secretary, and he decides that the only way to be with her is to kill his wife and 3 children. He knows that it is impossible to escape from his house in the event of a fire. Before he leaves on a business trip, he sets up a device in his basement that burns down the house and kills his family.

Is Bill fully morally responsible for killing his wife and children?

The results were dramatic. A full 72% of subjects in the concrete condition said that the agent was fully morally responsible, but less than 5% of subjects in the abstract condition said that it was possible to be fully morally responsible in a deterministic universe.

If this pattern of results is replicated in further experiments, we will have good reason to believe that no invariantist theory of moral responsibility can capture all of

people's ordinary judgments. Traditional incompatibilist theories diverge from people's judgments in more concrete cases, whereas traditional compatibilist theories diverge from people's judgments about more abstract principles. The only kind of theory that could consistently accord with people's judgments would be a theory that generated different conclusions depending on whether the question at hand was abstract or concrete.

Variance due to differences in moral status

As we explained above, the ambition of invariantist accounts of moral responsibility is to find a single set of criteria that can be used to assess moral responsibility for all possible behaviors. We are not supposed to end up with one criterion for morally good behaviors and another, slightly different criterion for morally bad behaviors. The aim is rather to find a single system of underlying principles from which all moral responsibility judgments can be derived.

This approach certainly has a strong intuitive appeal. It seems that moral responsibility is one thing and the goodness or badness of the behavior is something else. If we put together a judgment of moral responsibility with a judgment about whether the behavior itself is good or bad, we can determine whether or not the agent deserves praise or blame. But — it might be claimed — we do not need to assess the goodness or badness of the behavior itself before determining whether or not the agent is responsible for performing it.

Adherents of normative competence theories have argued that it is indeed possible to find a single set of criteria that can be applied to all behaviors, regardless of their moral status. The claim is that we can use this same set of criteria to make moral responsibility judgments for morally good behaviors, morally bad behaviors, and even behaviors that are morally neutral (Fischer & Ravizza 1998; Wolf 1990). This is a claim at the level of philosophical theory, but one can always ask how well it accords with ordinary judgments about particular cases. Our task now is to figure out whether or not it is possible to capture people's ordinary judgments by setting forth a single basic set of criteria that apply to all kinds of behavior.

A growing body of evidence suggests that it is not. In fact, there appear to be at least four distinct asymmetries whereby the criteria for moral responsibility depend in part on the moral status of the behavior itself.

1. *The side-effect asymmetry*. It is widely agreed that an agent can only be morally responsible for a behavior if she stands to that behavior in a certain kind of psychological relation. Still, there has been considerable disagreement about precisely which sort of psychological relation is necessary here. Some authors suggest that the agent only needs to have certain *beliefs* about what she is doing (e.g., Fischer & Ravizza 1998); others say that the agent needs to *identify* herself with the behavior (e.g., Doris 2002); and a number of researchers have suggested that there is an important link between moral responsibility and the notion of acting *intentionally* (e.g., Wallace 1994). In general, participants in this debate have tried to find a single type of psychological relation that would be necessary for moral responsibility in all cases. What we want to suggest here is that things may not be quite so simple. Perhaps different psychological relations prove relevant depending on whether the behavior itself is good or bad.

For a simple example, consider people's judgments of moral responsibility in cases of *foreseen side-effects*. These are cases in which an agent performs a behavior because she wants to bring about one effect (the desired effect) but is aware that she will also be bringing about some other effect that she does not specifically desire (the foreseen side-effect). The question is whether people will feel that the agent is responsible for bringing about the foreseen side-effects of her behaviors. As you may have guessed, the answer appears to be that it depends on whether the side-effects themselves are morally good or morally bad.

Knobe (2003a) ran a simple experiment that addresses this issue. Each subject was randomly assigned to one of two conditions. Subjects in the 'harm condition' received 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.

Subjects in the 'help condition' received a vignette that was almost exactly the same, except that the word 'harm' was replaced with 'help.' The vignette thus became:

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, and it will also help the environment.'

The chairman of the board answered, 'I don't care at all about helping 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 helped.

As expected, people's moral judgments showed a marked asymmetry. Most subjects in the harm condition said that the chairman deserved blame, but very few subjects in the help condition said that the chairman deserved praise. Subsequent research has found similar effects using other vignettes (Hauser et al. unpublished data; Knobe 2003a, forthcoming; Sverdlik forthcoming). There appears to be a general principle according to which people are given blame for bad side-effects but are not given praise for good side-effects. Thus, if we wanted to know whether a given effect was the sort of thing for which an agent would be said to deserve praise or blame, it would not be enough merely to know about the agent's psychological relation to the effect. We would also have to know whether the effect itself was good or bad.

How is this effect to be explained? Some readers have suggested that it might be due to a quite general 'negativity bias' (e. g., Rothbart and Park 1986: 137; Richey et al. 1975, 1982; Reeder and Coovert 1986; Skowronski and Carlston 1989). Perhaps people simply have a general tendency to be stingier with praise than they are with blame. The asymmetry we see here could then be understood as just one manifestation of an across-the-board disposition to apply more stringent criteria for moral responsibility in cases where a behavior is morally good. But as we shall see, things are not so simple. There does not appear to be a general effect whereby people are always stingier with praise than

with blame. Rather, it appears that there is a complex interaction between the goodness or badness of the behavior and the criteria for moral responsibility.

2. *The emotion asymmetry.* The complexity of people's responses emerges especially clearly when we consider the role that perceived emotion plays in judgments of moral responsibility. Sometimes an agent is so overwhelmed by emotion that she cannot resist performing a particular behavior. On such occasions, will people assign less praise or blame than they would have if the agent had decided to perform the behavior after a period of calm deliberation?

Pizarro, Uhlmann and Salovey (2003) set out to determine whether the impact of emotion might depend on the moral status of the behavior itself. They began by constructing a series of vignettes about agents who perform behaviors as a result of overwhelming emotion. Some of the vignettes featured morally good behaviors; others featured morally bad behaviors. Here is an example of a vignette with a morally good behavior:

Because of his overwhelming and uncontrollable sympathy,
Jack impulsively gave the homeless man his only jacket even
though it was freezing outside.

And here is one with a morally bad behavior:

Because of his overwhelming and uncontrollable anger, Jack
impulsively smashed the window of the car parked in front of
him because it was parked too close to his.

For each of these vignettes, the researchers then constructed a contrast case in which the agent acted calmly and deliberately. So, for example, the contrast case for our morally bad behavior was:

Jack calmly and deliberately smashed the window of the car
parked in front of him because it was parked too close to his.

When the researchers gave these vignettes to subjects, they found a striking asymmetry. Subjects gave the agent considerably less blame for morally bad behaviors when those behaviors were the result of overwhelming emotion than when they were the result of calm deliberation. But for morally good behaviors, there was no corresponding effect.

Subjects assigned just as much praise when the agent acted on overwhelming emotion as when the agent acted after calm deliberation.

Putting the side-effect asymmetry together with the emotion asymmetry, we see the emergence of a complex pattern. It seems that the fact that an outcome was merely a foreseen side-effect reduces the responsibility attributed for morally good behaviors but not for morally bad ones, whereas the fact that a behavior was the product of overwhelming emotion reduces the responsibility attributed for morally bad behaviors but not for morally good ones.

3. *The intention/action asymmetry.* As if this pattern were not complicated enough, we now turn to people's attributions of responsibility for unfulfilled intentions. Suppose an agent forms an intention but never actually gets a chance to perform the corresponding action. Will people assign praise or blame for the mere forming of the intention? Here again, the answer appears to depend on whether the behavior in question is good or bad.

To confirm this hypothesis, Malle and Bennett (2004) constructed pairs of sentences — with each pair consisting of one sentence that described an action and one that described the corresponding intention. Some of the pairs described actions and intentions that were morally good; others described actions and intentions that were morally bad. One of the morally good pairs was:

[*action*] helped a neighbor fix his roof.

[*intention*] intends to help a neighbor fix his roof.

One of the morally bad pairs was:

[*action*] sold cocaine to his teenage cousin.

[*intention*] intends to sell cocaine to his teenage cousin.

Subjects were given a list of such sentences and asked to make moral judgments. Some subjects were given sentences about actions; others were given sentences about mere intentions. In either case, subjects were asked how much praise or blame the agent deserved. This methodology allowed the researchers to measure the *difference* between the amount of praise or blame given for an action and the amount given for the corresponding intention.

As expected, there was a significant asymmetry between judgments about morally good intentions and judgments about morally bad intentions. In essence, there was a larger difference between intention and action for morally good pairs than for morally bad pairs. (Indeed, the effect size of the difference for morally good pairs was *twice* as high as the effect size for morally bad pairs.) Overall, the result was a substantial difference in the degree to which different types of intentions elicited reactive attitudes. Subjects were considerably more willing to assign blame for morally bad intentions than they were to assign praise for morally good intentions. Apparently, while good intentions may not pave the road to hell, they don't do much to grease the rails to heaven.

4. *The severity asymmetry.* Finally, consider cases in which the harm is due entirely to an accident. For any given accident, it will almost always be possible to find some way in which an agent could have taken precautions that would have prevented it, but we do not always conclude that such agents are responsible for the harm that results. Quite often, we feel that the agents did all that they could reasonably be expected to do and that they are therefore not responsible for the accidents that eventually arose. So it seems that people are able to establish a vague sort of threshold, such that one can say: 'As long as the agent's degree of care does not fall below this threshold, she is not responsible for the harm that results.' The key question now is whether people always draw that threshold at the same point or whether the precise location of the threshold actually depends on the goodness or badness of the outcome that ends up occurring.

Four decades of research on this topic points unequivocally to the conclusion that the location of the threshold actually depends on the nature of the outcome itself. People are willing to say that an agent is responsible for a severe harm even when that agent's behavior was only very slightly negligent whereas they refused to say that an agent is responsible for mild harms unless the agent was very negligent indeed. Note that this asymmetry is different in form from the three asymmetries discussed above. The asymmetry here is not between good outcomes and bad outcomes but rather between mildly bad outcomes and severely bad outcomes.

The severity asymmetry was first detected in a classic study by Walster (1966) and has since been replicated in a broad array of additional experiments. In Walster's

original study, all subjects were given a story about a man who parks his car at the top of a hill. They were told that the man remembered to put on his brakes but that he neglected to have his brake cables checked. The car rolls down the hill and ends up creating an accident. In one condition, subjects were told that the resulting harm was *mild* (a damaged fender); in the other condition, subjects were told that the resulting harm was *severe* (serious injury to an innocent child). All subjects were then asked whether the agent had acted carelessly and whether he was responsible for the accident. There was no difference between conditions in subjects' judgments as to whether the agent acted carelessly, but subjects were significantly more likely to say that the agent was responsible for the accident in the condition where the harm was described as severe than they were in the condition where the harm was described as mild.

This result was regarded as surprising, since it was initially assumed that people's responsibility judgments would depend only on the agent's level of negligence and not on the level of harm that ended up resulting. But the effect obtained in Walster's original study has subsequently been replicated in a number of additional studies using quite different methodologies, and a recent meta-analysis of 75 studies on the topic leaves little doubt that the effect is real (Robbennolt 2000).

The key remaining questions are about the psychological processes underlying the effect. Walster (1966) originally speculated that the effect was due to a motivational bias. People do not want to believe that their behaviors might lead to severe harms, and they are therefore motivated to believe that behavior can only lead to severe harm in cases where the agent is culpably negligent. But it can also be argued that the effect reflects a legitimate moral principle. Shaver (1970) points out that we often use the concept of responsibility when we are trying to determine whether anyone should pay restitution to the victim of an accident. But the more severe the accident, the more need there is for restitution. It therefore makes sense that people are less stringent in their standards for responsibility in cases where the harm is severe than they would be in cases where the harm is mild.

Summing Up

Thus far, we have presented data from four studies on people's assignment of praise and blame. These four studies all used the same basic structure. People were presented with behaviors that differed in their moral status but seemed almost exactly the same in every other relevant respect. It was then shown that people ascribed a lot of praise or blame to one of the behaviors but not to the other. The key question now is whether these results indicate that the folk practice of responsibility attribution is not invariantist or whether there is some way to explain the results even on the assumption that the folk practice is entirely invariantist.

It might be argued, for example, that the asymmetries obtained in the studies are not really showing us anything about people's attributions of moral responsibility. Maybe people regard the agents in both conditions as morally responsible; it's just that they don't assign praise or blame in one of the conditions because they do not feel that the agent in that condition truly did anything good or bad. This explanation might be plausible for the intention/action asymmetry on our list, but it does not seem plausible for any of the others. Subjects can be reasonably supposed to believe that helping the environment is something good and smashing a car window is something bad. To the extent that people do not assign praise or blame for these behaviors, it is presumably because they do not take the agent to be morally responsible.

A second possible strategy would be to argue that there is a single, more or less unified criterion for moral responsibility and that all of the apparent asymmetries we have discussed can be derived in some way from this one criterion. So, for example, Wolf (1990) has argued that what we have called the 'emotion asymmetry' can actually be derived in a straightforward way from the normative competence theory. Recall that the normative competence theory says that an agent is morally responsible if and only if she is capable of doing the right thing for the right reasons. But this unified theory seems to lead immediately to different judgments in cases of different emotions. After all, it does seem that overwhelming and uncontrollable anger might render a person incapable of doing the right thing for the right reasons but that overwhelming and uncontrollable sympathy does not have the same effect. We will not be concerned here with the question as to whether or not Wolf's account actually does give us a good explanation of the results obtained in the study. Instead, we observe that the account does not even begin to

explain the various other asymmetries (e.g., the side-effect asymmetry). To show that the folk practice is invariantist in the relevant sense, it would be necessary to find a single set of criteria that can explain all of the asymmetries described in this section.

It certainly would be an instructive and valuable effort to look for a single invariant set of criteria from which all of these apparent asymmetries can be derived. We wish future researchers the best of luck in this effort, but to be frank, we don't think it is very likely that they will have much success.

Variance in the Antecedents of Responsibility Judgments

Here it might be objected that, although certain aspects of the criteria for moral responsibility differ depending on whether the behavior itself is morally good or morally bad, the most central and important aspects of these criteria nonetheless remain entirely invariant. Hence, someone might say: 'Look, you've convinced me that at least some aspects of the criteria vary, and I guess that is enough to prove your basic thesis. But let's not get carried away. It certainly does seem that many central aspects of the criteria always remain exactly the same. No matter whether the outcome is good or bad, we will always be concerned with questions about whether the agent *caused* that outcome and whether she brought it about *intentionally*. These aspects of the criteria, at least, appear to be perfectly invariant.'

Our response to this objection will be a somewhat surprising one. We are happy to admit that people may always base responsibility judgments on antecedent judgments about causation and intentional action, but we want to suggest that these antecedent judgments themselves are not invariant. It may be, e.g., that people always ask themselves whether the agent *caused* the outcome and whether the agent acted *intentionally*, but there is evidence to suggest that people do not have invariant criteria for assigning causation and intentional action. Instead, it appears that people use different criteria depending on whether the behavior itself is good or bad.

Consider first the claim that responsibility judgments are always sensitive to judgments as to whether or not the agent acted intentionally. How can this claim be reconciled with the hypothesis (presented above) that the relevant psychological relation

depends on whether the behavior itself is good or bad? The answer is simple. People always ask whether the agent acted intentionally, but their judgments as to whether or not the agent acted intentionally sometimes depend on whether the behavior itself was good or bad.

This point comes out clearly in the experiment described above (Knobe 2003a). Recall that all subjects were given a vignette about an agent who brings about a foreseen side-effect, but some subjects received a vignette in which the side-effect was morally bad and others received a vignette in which the side-effect was morally good. The surprising result was that subjects in these different conditions had different intuitions about whether or not the agent acted intentionally. Most subjects in the harm condition said that the corporate executive *intentionally* harmed the environment, but most subjects in the help condition said that the agent *unintentionally* helped the environment.⁶ (Try reading the vignettes for yourself. We suspect that you will have the same intuitions.) Here we find an asymmetry in people's views about whether the agent acted intentionally even though all of the relevant psychological features appear to be the same in the two conditions. The chief difference seems to lie in the moral status of the behavior performed.

These results are somewhat puzzling (to say the least!), and a number of competing hypotheses have been proposed to explain them (Adams & Steadman 2003, 2004; Knobe 2004; Mele 2003; Malle forthcoming b; Nadelhoffer forthcoming a). We will not be defending a specific hypothesis here. Instead, we simply want to point to the surprising mesh between the criteria used for assessing moral responsibility and the criteria used for determining whether or not an agent acted intentionally. We noted above that moral responsibility judgments rely on different psychological states depending on whether the behavior itself is good or bad. In particular, it seems that foresight is often sufficient when the behavior is morally bad but that trying is usually necessary when the behavior is morally good. We now see exactly the same pattern in people's judgments as

⁶ This effect appears to be remarkably robust. It continues to emerge when the vignettes are translated into Hindi and run on Hindi-speaking subjects (Knobe & Burra forthcoming), when subjects are only four years old (Leslie, Knobe & Cohen forthcoming), and even when subjects have deficits in emotional processing due to frontal lobe damage (Hauser, et al. unpublished data). For further replications and extensions, see Adams & Steadman (2005), Knobe (2003b, 2004), Knobe & Mendlow (forthcoming), Nadelhoffer (2004, forthcoming a, forthcoming b), Malle (forthcoming a), McCann (forthcoming), and Nichols (unpublished data).

to whether the agent acted intentionally. Here again, it seems that foresight is sufficient when the behavior is morally bad but that trying is necessary when the behavior is morally good.

Similar remarks apply to people's judgments of causation. It may well be that judgments of causation play an important role whenever people are trying to assess moral responsibility, but the evidence suggests that causal judgments themselves are not derived using invariant criteria. Instead, it appears that the criteria used in causal judgments vary depending on the moral status of the behavior itself.

This point comes out especially clearly in a well-known study by Alicke and colleagues (1992). All subjects were given a story about an agent who is driving home 10 miles per hour above the speed limit. In one condition, the agent needs to get home swiftly so that he can hide the present he just bought for his parents. In the other, he needs to get home so that he can hide his stash of cocaine. Either way, the agent then ends up getting into an accident. The key dependent variable was subjects' judgments about the degree to which the agent *caused* the accident by driving too fast. By now, you have probably guessed the result. Subjects were significantly more inclined to say that the agent caused the accident when he was driving home to hide his cocaine than when he was driving home to hide the present (Alicke 1992). Similar results have been obtained in a variety of other studies (Alicke 2000; Alicke et al. 1994; Solan and Darley 2001). The consensus among social psychologists appears to be that, collectively, these studies provide strong evidence for the view that moral considerations have a powerful impact on people's causal judgments.

Nonetheless, most social psychologists assume that moral considerations do not actually play any role in people's underlying *concept* of causation. Instead, the usual view distinguishes between multiple levels. First, there is the fundamental competence that people use to assess causation. This competence is taken to be a purely descriptive mechanism (perhaps something along the lines suggested by Kelly's, 1967, theory of 'the person as scientist'), and it is assumed that moral considerations play no role in it. Then, second, there is some additional process by which moral considerations can 'bias' or 'distort' people's causal judgments, leading them away from what the underlying competence itself would have proposed.

In recent years, however, a number of philosophers have proposed more radical views according to which moral considerations truly do play a role in people's concept of causation (Hitchcock unpublished; Knobe & Fraser forthcoming; McGrath forthcoming; Thomson 2003). These philosophers argue that the connection between moral judgments and causal judgments is not, in fact, due to a performance error. Rather, people's moral judgments influence their causal judgments because moral features actually figure in people's concept of causation.

Perhaps the best way to get a sense for what these philosophers are suggesting is to consider the kinds of cases they typically discuss. Here is one representative case:

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.

Faced with this case, most subjects say that Professor Smith *did* cause the problem but that the administrative assistant *did not* cause the problem (Knobe & Fraser forthcoming). And yet, the two agents seem to have performed almost exactly the same behavior in almost exactly the same circumstances; the principal difference between the two behaviors appears to lie in their differing *moral* statuses. In cases like these, it seems plausible to suppose that moral considerations could really be playing some fundamental role in the basic competence by which we assess causation.

Of course, it might be true that causal judgments always have the same impact on judgments of moral responsibility, regardless of whether the behavior itself is morally good or morally bad. But the moral goodness or badness of the behavior still ends up influencing moral responsibility judgments in an indirect way. It influences people's causal judgments, which in turn play a role in their judgments of moral responsibility.

Variance due to relationships

Philosophical discussions of moral responsibility are often concerned in an essential way with our ordinary practice of responsibility attribution, and philosophers therefore frequently appeal to ordinary people's judgments about particular cases. But the cases described in these philosophical discussions almost always take a somewhat unusual form. They are almost always hypothetical cases involving entirely *imaginary* characters. So, for example, Frankfurt's famous argument about alternate possibilities relies on a story about a man named Jones being controlled by a nefarious neurosurgeon named Black (Frankfurt 1969). When one sees that most people agree about whether or not the characters in these stories are morally responsible, it is easy to get the sense that there must be some invariant criterion for moral responsibility that almost everyone is using.

But, clearly, ordinary attributions of moral responsibility do not usually work like these philosophical examples. Most ordinary attributions of responsibility are not about complete strangers; they are about people to whom we stand in certain *relationships* (friends, spouses, coworkers, etc.). If we want to know whether there really is an invariant criterion for responsibility judgments, we need to look at cases involving a wide variety of relationships and see whether it is possible to identify a single criterion underlying them all.

The best way to address this question is to look at the psychological literature on moral responsibility. Unfortunately, though, most of this literature uses the very same methodology that the philosophical literature does. (A recent review of the psychological literature on blame found that 77% of studies used hypothetical scenarios, and 65% used scenarios in which the transgressor was an entirely fictional character; Pearce 2003.) So although the empirical literature does provide a few fascinating insights into the connection between personal relationships and attributions of moral responsibility, it also leaves a number of important questions unanswered. Here we discuss a few highlights of the existing literature and then pose a number of questions that still remain unaddressed.

We began with Arriaga and Rusbult's (1998) study of perspective taking and blame. The phrase 'perspective taking' refers here to a disposition to try to imagine how a situation might appear from another person's position. The researchers wanted to know whether this disposition would be associated with low levels of blame attribution. So they proceeded in the obvious way. They gave all subjects a questionnaire designed to

assess an overall disposition for perspective taking. Then they presented subjects with hypothetical stories and asked them how much blame the agents in these stories deserved. The key research question was whether or not there would be a correlation between level of perspective taking and level of blame. As it happened, there was no significant correlation.

But the researchers also introduced an interesting variation on the usual experimental paradigm. Subjects were asked questions designed to assess the degree to which they showed perspective taking *specifically in relation to their spouses*. Instead of being asked general questions about their personalities, subjects were asked to specify their level of agreement with sentences about how they normally thought of their partner (e.g., ‘When I’m upset or irritated by my partner, I try to imagine how I would feel if I were in his/her shoes.’). After answering these initial questions, each subject was presented with a hypothetical scenario concerning his or her spouse. For example:

You feel neglected by your partner, who has been very busy lately. You nevertheless make dinner plans for an approaching evening, to which the partner reluctantly agrees. Your partner arrives for dinner half an hour late, not ready to dine, explaining that he or she must cancel dinner because of a course project that is due the next day.

Subjects were then asked how much blame the spouse would deserve if this scenario had actually taken place. As predicted, there was a significant correlation whereby subjects who were high in perspective taking in relation to their spouses tended to assign lower levels of blame.

Similar results were obtained in a study by Fincham, Beach and Baucom (1987). The aim of the study was to compare blame attributions among ordinary couples with attributions from ‘distressed’ couples who had chosen to come in for counseling. Members of each of these groups received two kinds of questions:

- (1) They were told to imagine that their *spouses* had performed particular behaviors and then asked how much praise or blame the spouses would deserve.
- (2) They were told to imagine that *they themselves* had performed certain behaviors and were asked how much praise or blame they themselves would deserve.

The key question was whether subjects in each group would assign different levels of blame depending on whether the agent was the self or the spouse.

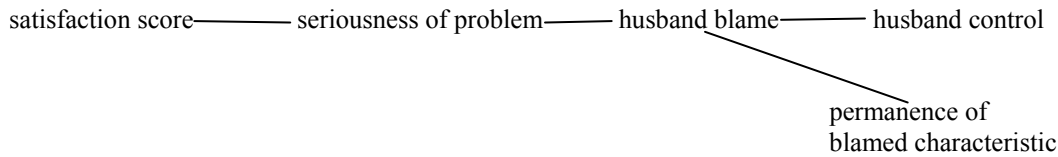
Members of distressed couples showed an asymmetry between judgments about the self and judgments about the spouse. They assigned more credit to themselves than to their spouses and more blame to their spouses than to themselves. This result is hardly surprising. The surprising results came from the normal couples (couples who had not specifically come in for counseling). Members of these couples also showed an asymmetry — but in the opposite direction. They assigned more credit to their spouses than they did to themselves. In other words, members of normal couples were in a state of systematic disagreement with each other. Each of them thought that the other was the one who deserved more praise.

Of course, a number of questions arise about how to interpret these results. It is possible (at least in principle) that the results could be obtained even if people's relationships had no effect at all on their attributions of blame and praise. For example, it could be that some factor that has nothing to do with people's relationships is causing certain couples to show higher levels of blame. This high blame might then cause the relationship to become distressed, thereby producing the correlation found in the study. But most researchers do not think that the process works like this. The most common view appears to be that attributions and marital satisfaction affect each other in a cyclical fashion, with high levels of satisfaction leading to low levels of blame and low levels of satisfaction leading to high levels of blame (e.g., Harvey et al.).

Still, a question arises about precisely how marital satisfaction impacts attributions of blame. In particular, one wants to know whether marital satisfaction is actually having any impact on the fundamental criteria underlying people's moral judgments or whether it only affects moral judgments indirectly by first affecting people's judgments regarding particular matters of fact (what their spouses are trying to do, how much control they have over certain outcomes, etc.).

The studies of Madden and Janoff-Bulman (1981) help us to address this question. Wives at varying levels of marital satisfaction were presented with hypothetical stories involving their husbands. They were then asked to make judgments about both (a) the amount of blame the husband would deserve if the fictitious events had actually

happened and (b) the degree to which the husband fulfilled certain traditional criteria for blameworthiness. Using exploratory path analysis, the researchers arrived at the following model:



This analysis does not permit us to determine the *direction* of causality. (Thus, it does not allow us to determine whether low satisfaction leads people to feel that the problem is serious or whether the feeling that the problem is serious leads to low satisfaction.) However, the analysis does allow us to figure out which variables are influencing each other directly and which only influence each other through the mediation of some other variable. In particular, the analysis suggests that marital satisfaction is not influencing blame through the mediation of judgments about control or permanence. To the extent that marital satisfaction influences blame at all, it appears to do so in a way that is independent of these other judgments.

We began this section by noting that philosophical discussions of moral responsibility often try to make contact with our ordinary practice of responsibility attribution. The most common method for investigating this practice is to look at people's judgments concerning particular cases. However, the studies we have presented seem to indicate that people's judgments about particular cases can vary dramatically depending on their relationship to the agent. Thus, people may arrive at different attributions depending on whether the agent is a beloved partner, a bitter enemy, or a complete stranger. This conclusion sheds new light on the methods usually used in philosophical discussions. It seems that these discussions have been concerned with attributions to one particular type of agent — namely, attributions to agents with whom one has no prior relationship.

Here it might be argued that the absence of any prior relationship gives us an especially pure glimpse into the nature of responsibility attributions. It might be thought, e.g., that people's relationships to the agent serve as a kind of 'distortion,' leading them

to violate norms that they themselves accept. This is an interesting hypothesis, which needs to be investigated empirically. One wants to know whether information about personal relationships actually figures in the fundamental competence underlying ascriptions of moral responsibility or whether it only impacts the process by increasing the probability of various performance errors.

On a related note, it would be helpful to know how people who stand in different relationships to the agent conceive of the difference between their perspectives. Take the case of a woman who cheats on her husband. Here we might find that her husband regards her as blameworthy but her friends do not. What remains to be seen is how people make sense of the divergence of attitudes in cases like this one. One possibility would be that both sides agree that there is a single objective answer to the question as to whether the woman is blameworthy or not and that they simply disagree about what this answer actually is. Another possibility would be that neither side believes that there is any real disagreement. That is, both sides might feel that the woman is worthy of being blamed by her husband but not by her friends. If we did obtain this result, we would have evidence that people explicitly reject the idea of invariant criteria for the ascription of moral responsibility.

Conclusion

Thus far, we have been trying to show that the approach philosophers have taken when constructing theories of moral responsibility diverges in certain ways from the approach people ordinarily take when making moral responsibility judgments. Philosophers have searched for a single invariant system of principles that can be used in all cases. But ordinary people do not appear to make use of invariant criteria. Instead, it appears that they apply different criteria in different cases.

This divergence between philosophical theory and ordinary practice leaves us with a difficult question. One wants to know *which side is actually right*. One view would be that ordinary people are making some kind of mistake and that they really ought to be applying the very same criteria in every case. Another view would be that philosophers are making a mistake and that they ought to give up their search for

invariant criteria. And, of course, it would be possible to adopt an intermediary position according to which each side ought to move a little bit toward the middle.⁷

We cannot hope to resolve this issue here. Instead, our aim is to provide a few preliminary suggestions about how one might approach the question. In particular, we want to suggest that it is possible to distinguish two basic traditions in contemporary philosophical thought about moral responsibility and that these two traditions lead to two very different ways of thinking about the questions surrounding invariantism.

The first tradition takes the questions surrounding the nature of moral responsibility to be closely tied to questions in analytic metaphysics. It is assumed that there is a certain relation between agents and events — the relation that makes a given agent responsible for a given event — and that the aim of philosophical work on moral responsibility should be to analyze the conditions under which this relation obtains. Research in this tradition is usually quite technical in nature, even to the point of occasionally including formal proofs in modal logic.

Of course, such research does sometimes appeal to the intuitions of ordinary people, but it also attaches great significance to theoretical simplicity and elegance. Researchers in this first tradition typically assume that there must be a single underlying set of criteria that can be used to make responsibility judgments in all cases. It would be seen as a mistake to propose that one ought to apply different criteria in different cases, and it would be regarded as absurd even to suggest that the very same person could rightly be classified in one way by a friend and in another by a stranger.

There is, however, a second, very different tradition in philosophical work on moral responsibility. This second tradition — coming down to us from Strawson's (1962) 'Freedom and Resentment' — focuses more on the social and psychological aspects of moral responsibility judgments. The emphasis is not so much on the relation of *being* responsible as on the social practice of *holding* people responsible. The key questions for this second tradition are about how this practice works, what role it serves in people's lives, and whether it might be able to serve this role better if it worked somewhat

⁷ Here we come into contact with difficult meta-philosophical questions about the very idea of constructing philosophical theories that diverge from ordinary judgments. For an excellent recent discussion of these questions, see Vargas (2005).

differently. Clearly, there is no way to decide ahead of time how these Strawsonian questions ought to be answered.

Our aim here is not so much to provide definitive answers as to suggest a new approach to the questions. We have suggested that the best way to arrive at an accurate understanding of the existing practice is to engage in a serious way with the results of empirical studies. These results point in a surprising new direction. Indeed, they appear to go against one of the key assumptions underlying most of the theories in the philosophical literature. Specifically, they suggest that people simply do not have invariant criteria for making moral responsibility judgments. This empirical finding leaves us with important normative questions about whether the ordinary practice of responsibility attribution is in good shape as it is or whether it needs to be revised in some way, but before we even begin to address these normative questions, we need to grapple with the evidence suggesting that the practice we have now is not an invariantist one.

References

- Alicke, M.D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, 126, 556-574.
- Alicke, M.D. (1992). Culpable causation. *Journal of Personality and Social Psychology*, 63, 368-378.
- Alicke, M.D., Davis, T.L., & Pezzo, M.V. (1994). A posteriori adjustment of a priori decision criteria. *Social Cognition*, 12, 281-308.
- Arriaga, X. & Rusbult, C. (1998). Standing in my partner's shoes: Partner perspective taking and reactions to accommodative dilemmas. *Personality and Social Psychology Bulletin*, 9, 927-948.

- Cover, J.A. and J. O'Leary-Hawthorne. 1996. Free agency and materialism. In *Faith, Freedom and Rationality*, ed. J. Jordan and D. Howard-Snyder, 47-71. Lanham, MD: Roman and Littlefield.
- Doris, John M. (2002). *Lack of Character: Personality and Moral Behavior*. Cambridge: Cambridge University Press.
- Ekstrom, L. 2002. Libertarianism and Frankfurt-style cases. In R. Kane (ed.), *The Oxford Handbook of Free Will* (New York: Oxford University Press).
- Eshleman, A. Moral responsibility. *The Stanford Encyclopedia of Philosophy* (Fall 2004 Edition), E. N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2004/entries/moral-responsibility/>.
- Fincham, F., Beach, S. & Baucom, D. (1987). Attribution processes in distressed and nondistressed couples: 4. Self-partner attribution differences. *Journal of Personality and Social Psychology*, 52, 739-748.
- Fischer, J. M. (1999). Recent work on moral responsibility. *Ethics* 110: 93-139.
- Fischer, J. & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. New York: Cambridge University Press.
- Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy*, 66, 829-39.
- Goodman, N. (1954). *Fact, Fiction, and Forecast*. University of London: Athlone Press.
- Kane, R. (1996). *The Significance of Free Will*. New York: Oxford University Press.

- Kane, R. (1999). Responsibility, luck, and chance: reflections on free will and indeterminism. *Journal of Philosophy* 96: 217-240.
- Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*. 63, 190-193.
- Knobe, J. & Burra, A. (forthcoming). Intention and intentional action: A cross-cultural study. *Journal of Culture and Cognition*.
- Knobe, J. & Fraser, B. (forthcoming). Causal judgment and moral judgment: Two experiments. In Walter Sinnott-Armstrong (ed.).
- Knobe, J. & Mendlow, G. (forthcoming). The good, the bad, and the blameworthy: Understanding the role of evaluative considerations in folk psychology. *Journal of Theoretical and Philosophical Psychology*.
- Leslie, A., Knobe, J. & Cohen, A. (forthcoming). Acting intentionally and the side-effect effect: 'Theory of mind' and moral judgment. *Psychological Science*.
- Madden, M. & Janoff-Bulman, R. (1981). Blame, control, and marital satisfaction: Wives' attributions for conflict in marriage. *Journal of Marriage and the Family*. 43, 663-674.
- Malle, B. F. (forthcoming). The moral dimension of people's intentionality judgments. *Journal of Culture and Cognition*.
- Malle, B. & Bennett, R. (2004). People's praise and blame for intentions and actions: Implications of the folk concept of intentionality. Unpublished manuscript. University of Oregon.
- McGrath, S. (forthcoming). Causation by omission. *Philosophical Studies*.

- Nisbett, R., & Wilson, T. (1977). Telling more than we know: Verbal reports on mental processes. *Psychological Review*, 84, 231-259.
- Pearce, G. (2003). The psychology of everyday blame. Doctoral dissertation. University of Oregon.
- Pereboom, D. (2001). *Living without free will*. Cambridge: Cambridge University Press.
- Pizarro, D., Uhlmann, E. & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science*, 14, 267-272.
- Reeder, G.D., and Coovert, M. D. (1986) "Revising an Impression of Morality." *Social Cognition* 4: 1-17.
- Richey, M. H., Koenigs, R. J., Richey, H. W., and Fortin, R. (1975) "Negative Salience in Impressions of Character: Effects of Unequal Proportions of Positive and Negative Information." *Journal of Social Psychology* 97: 233-41.
- Richey, M. H., Bono, F. S., Lewis, H. V., and Richey, H. W. (1982) "Selectivity of Negative Bias in Impression Formation." *Journal of Social Psychology* 116: 107-18.
- Robbennolt, J. (2000). Outcome severity and judgments of 'responsibility': A meta-analytic review. *Journal of Applied Social Psychology*, 30, 2575-2609.
- Rothbart, M., and Park, B. (1986) "On the Confirmability and Disconfirmability of Trait Concepts." *Journal of Personality and Social Psychology* 50: 131-42.
- Shaver, S. (1970). Redress and conscientiousness in the attribution of responsibility for accidents. *Journal of Experimental Social Psychology*, 6, 100-110.

- Skowronski, J. J., and Carlston, D. E. (1989) "Negativity and Extremity Biases in Impression Formation: A Review of Explanations." *Psychological Bulletin* 105: 131-42.
- Solan, L.M. & Darley, J.M. (2001). Causation, contribution, and legal liability: An empirical study. *Law and Contemporary Problems*, 64, 265-298.
- Strawson, Peter. (1962). Freedom and resentment. *Proceedings of the British Academy* 48, 187-211.
- Sverdlik, S. (forthcoming). Intentionality and moral judgments in commonsense thought about action. *Journal of Theoretical and Philosophical Psychology*.
- Thomson, J. (2003). Causation: Omissions, *Philosophy and Phenomenological Research*, 66, 81.
- van Inwagen, P. (1983). *An essay on free will*. Oxford: Oxford University Press.
- Vargas, M. (2005). The revisionist's guide to responsibility. *Philosophical Studies* 125:399-429.
- Viney, W., Waldman, D., & Barchilon, J. (1982). "Attitudes toward Punishment in Relation to Beliefs in Free Will and Determinism" *Human Relations*, 35, 939-49.
- Viney, W.; Parker-Martin, P.; Dotten, S. D.H. (1988). "Beliefs in Free Will and Determinism and Lack of Relation to Punishment Rationale and Magnitude." *Journal of General Psychology* 115, 15-23.
- Wallace, R. J. (1994). *Responsibility and the moral sentiments*. Cambridge, Mass.: Harvard University Press.
- Walster, E. (1966). Assignment of responsibility for an accident. *Journal of Personality and Social Psychology*, 3, 73-79.

Watson, G. (1975). Free agency. *Journal of Philosophy*, 205-20.

Watson, G. (1996). Two faces of responsibility. *Philosophical Topics* 24:
227-248.

Wicker, A. W. (1969). Attitudes vs. actions: The relationship of verbal and overt
behavioural responses to attitude objects. *Journal of Social Issues*, 22, 41-78.

Wolf, S. (1990). *Freedom within reason*. Oxford: Oxford University Press.