Philosophia (2011) 39:179–200
DOI 10.1007/s11406-010-9262-9

Moral Responsibility Invariantism

Brandon Warmke

Received: 1 March 2010 / Revised: 7 May 2010 / Accepted: 26 July 2010 /
Published online: 17 September 2010

Abstract Moral responsibility invariantism is the view that there is a single set of conditions for being morally responsible for an action (or omission or consequence of an act or omission) that applies in all cases. I defend this view against some recent arguments by Joshua Knobe and John Doris.

Keywords Moral responsibility · Doris · Knobe · Experimental philosophy

Introduction

Suppose I punch you in the face. Philosophers writing on moral responsibility have typically (but by no means universally) thought that there are three central kinds of conditions that must be met in order for an agent to be morally responsible for doing something like punching you in the face. First, an agent must, in so acting, meet a *control* (or *freedom*) condition. Some philosophers think this means that we just must be able to guide our actions in certain ways—that we are not compelled or forced to act by coercion, for example. Some other philosophers think that in order to meet the control condition, agents need to have so-called "alternative possibilities" available to them—the future must be open. But whatever specific kind of control condition is necessary for being morally responsible for an action (or omission, or consequence of an action or omission[1]), there is widespread agreement that *some* control condition is necessary.

Second, many philosophers have thought that in order to be morally responsible for an action, an agent must meet an *epistemic* condition. Suppose I punch you in the face because I am practicing my boxing while wearing a blind-fold in what I have

[1]For ease of expression I will henceforth suppress these second two disjuncts and speak in terms of responsibility for actions.

B. Warmke (✉)
Department of Philosophy, University of Arizona, Social Science Bldg. Rm 213, PO Box 210027, Tucson, Arizona 85721-0027, USA
e-mail: bwarmke@email.arizona.edu

good reason to believe is an empty room, and that you happen to walk accidentally into my flying fist. Many philosophers think that in a case like this, I am not morally responsible for punching you because I was non-culpably ignorant of certain relevant facts about the context of my action (e.g. the facts that you were in the way of my flying fist and that my punching you would harm you).

Third, many philosophers argue that agents must also meet some kind of *authenticity* condition. Suppose a mad scientist manipulates my brain or brainwashes me in such a way so as to give me strong desires to punch others in the face, desires I did not have before her manipulation or brainwashing. Such philosophers hold that even if I acted *freely* and in *full awareness* of what I was doing, I may not be morally responsible for punching you if the strong desires on which I acted were implanted in me by a process circumventing typical processes of desire-acquisition.[2]

Not surprisingly, not all philosophers writing on moral responsibility agree that these three conditions, broadly construed and as I have outlined them, are individually necessary and jointly sufficient for being morally responsible for an action.[3] Even less surprising, few philosophers agree as to the *specific content* of these conditions. (Consider the cottage industry devoted to contentious "Frankfurt-style cases"—cases putatively showing that alternative possibilities are not necessary for being morally responsible.) Aside from all of these differences, however, Joshua Knobe and John Doris have recently argued[4] that philosophers who are working to find a set of conditions for being morally responsible for an action are united in utilizing the same research program, one that is guided by two assumptions:

> *Invariantist Assumption*: There is a single set of conditions for being morally responsible for an action that applies in all cases.
>
> *Conservativist Assumption*: The conditions for being morally responsible for an action should accord with all (or most) of our ordinary judgments about the conditions under which an agent is morally responsible and we can discover these conditions by considering these ordinary judgments.

The Invariantist Assumption is a *metaphysical claim* about the existence of a single set of necessary and sufficient conditions for being morally responsible for an action. The Conservativist Assumption is a *methodological* claim about how we are to go about discovering that set of conditions. The assumption is that the conditions of moral responsibility can be discovered by considering our ordinary judgments and that were we to discover the conditions for being morally responsible, that discovery should leave those ordinary judgments largely unchanged.

Given a large and growing body of literature on the psychology of responsibility attribution, however, Knobe and Doris argue that ordinary judgments of moral

[2] Haji (1998, ch. 1) organizes these three broad conditions in roughly this way also.

[3] For example, Fischer and Ravizza (1998) fold an authenticity condition into their control condition. Others, like Sher (2009), have strong doubts about the epistemic condition.

[4] Knobe and Doris (2010). All references to Knobe and Doris in this paper are to their paper "Strawsonian Variations: Folk Morality and the Search for a Unified Theory," to appear in J.M. Doris et al. (eds.), *Oxford Handbook of Moral Psychology*, Oxford, Oxford University Press.

responsibility do *not* reveal a single set of criteria that people use to attribute responsibility in all cases. Therefore:

> *Empirical Conclusion*: Empirical studies of ordinary judgments of responsibility attribution reveal that there is no single set of conditions under which the folk attribute responsibility.

Knobe and Doris then argue that given the Empirical Conclusion, we are stuck with a dilemma. We can continue to hold that there is a single set of conditions for moral responsibility, but if we were to do so, we would have to give up the conservativist methodology. On the other hand, we can continue to use a methodology that consults our ordinary judgments, but if we were to do so, we would have to abandon the assumption that there is a *single* set of conditions for moral responsibility. More succinctly: given the Empirical Conclusion, we can retain invariantism or conservativism, but not both. And for Strawsonian reasons, Knobe and Doris conclude that we would do best to reject invariantism.

In this paper I defend moral responsibility invariantism. Contrary to what Knobe and Doris claim, those philosophers who are committed to both Invariance and Conservatism (call them 'Standard Theorists') need not revise their research program in light of the current empirical literature. In Sections II and III, I will explain both the Invariantist and Conservativist Assumptions and show how Standard Theorists are alleged to employ them. In Section IV, I will briefly review some of the relevant psychological literature that Knobe and Doris cite in support of the Empirical Conclusion, and explain why they think this poses a problem for Standard Theorists. My discussion turns critical in Sections V–VII where I argue that Knobe and Doris have yet to show why the Standard Theorist is committed to any kind of inconsistency and that therefore Standard Theorists are currently under no burden to give up moral responsibility invariantism.

The Invariantist Assumption

Knobe and Doris take as their foil a group of philosophers of moral responsibility they claim to be committed to two assumptions. The first of these is the assumption that there is a single set of invariantist criteria for being morally responsible for an action. An invariantist theory of responsibility is a theory that says that the conditions under which an agent is morally responsible for an action are universal and exceptionless; the conditions apply to everyone regardless of context. Invariantist theories therefore demand that when making judgments of moral responsibility, we should always use the same criteria. Knobe and Doris claim that at least since Strawson's landmark essay, "Freedom and Resentment," philosophers of moral responsibility have simply *assumed* that our account of the conditions under which an agent is morally responsible for an action ought to be invariantist.[5] They describe this assumption this way:

> 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

[5] (Knobe and Doris 2010)

> 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*. (Knobe and Doris 2010)

In contrast to an invariantist theory of moral responsibility for actions, a variantist theory would claim that the conditions under which an agent is morally responsible are context-sensitive. Therefore, such a theory would demand that when making judgments of moral responsibility, we should *not* always use the same criteria. Knobe and Doris illustrate the distinction this way:

> [A]n 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…'[6]

Whereas an invariantist theory gives us a rule that says we should apply the same criteria to everyone in every case regardless of context, a *variantist* theory of responsibility would also give us a rule for responsibility attribution, but one that says that different criteria are relevant depending on the certain kinds of contextual features. The following heuristic is therefore helpful to distinguish invariantist from variantist theories of moral responsibility: invariantist theories give us conditions for being morally responsible that are context-independent, whereas variantist theories give us conditions for being morally responsible that are context-dependent.

This, of course, is just a heuristic. On reflection, however, one can see that distinguishing variantist theories from invariantist theories turns out to be a sticky wicket. This is because for any variantist theory, we could in principle construct a conjunction of conditionals of some definite length that would give us the rules for attributing moral responsibility in the various morally relevant contexts. That conjunction would give us, it would seem, an *invariantist* criterion for attributing moral responsibility in every morally relevant context, for in every possible morally relevant context in which an agent is morally responsible for an action, that conjunction of conditionals will be satisfied. For example, suppose we had a criterion like the following:

> [(If agent S is in context C1, then S must meet criteria set R1 in order to be morally responsible for x) and (If agent S is in context C2 (where C2 ≠ C1, then S must meet criteria set R2 (where R2 ≠ R1) in order to be morally responsible for x) and...]

[6] Knobe and Doris (2010)

The worry here is that once we discovered what the operative criteria are in each morally relevant context, we could in principle construct an invariantist criterion for moral responsibility. This suggests that the distinction between invariantist and variantist theories of moral responsibility (or of anything for that matter) is one without difference.

Knobe and Doris disagree with this assessment, as do I.[7] It seems that we can make sense of the difference between variantist and invariantist criteria without incurring the burden of having to provide necessary and sufficient conditions for what distinguishes invariantist theories from variantist theories. Ostensibly, there *are* important differences between these two kinds of criteria. For instance: for an invariantist theory, there might be, let's say, a *single* control condition, a *single* epistemic condition, and a *single* authenticity condition, and that these conditions apply universally and are exceptionless. Ostensibly, this would not be true of a variantist theory, for there would *not* be a single control condition, a single epistemic condition, and a single authenticity condition that *apply universally* and that are *exceptionless*. At least for our present purposes, then, our heuristic seems good enough: invariantist theories are context-independent, while variantist theories are context-dependent.[8]

These differences should become clearer after we discuss the empirical literature below. But until then, it might be helpful to give some examples of an invariantist theory of moral responsibility. Knobe and Doris provide their own examples of what they take to be invariantist theories of moral responsibility. For instance, they point out that *incompatibilist* theories claim that moral responsibility is *always* incompatible with determinism.[9] There are no circumstances, so the theory goes, in which a person who is determined to act might be morally responsible for that action. It does not matter whether the agent under question is an authority figure, in a high emotional state, or a close relative. The same necessary condition for being morally responsible applies to everyone in every case. Compatibilist theories, on the other hand, claim that it is possible to be morally responsible for one's actions even if one's actions are causally determined by the conjunction of laws of nature and the distant past. *Real self* compatibilist views claim that agents are responsible only if[10]

[7] They write: [T]here is the problem that one can use certain cheap logical tricks to make just about any rule look invariantist. [fn. removed] 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'" (Knobe and Doris 2010).

[8] We should be careful to note what an invariantist theory of moral responsibility does *not* entail. First, invariantist theories are not committed to the claim that there cannot be *different kinds of responsiblility* See, for example, Watson (1996), on the two "faces" of responsibility: attributability and accountability. One face being invariantist would not entail that the other is. Second, an invariantist theory of moral responsibility does not exclude the possibility of an agent having an excuse that exculpates her for some wrongdoing. An excusing condition is a condition revealing that while an agent has done something morally wrong or bad, she is not morally blameworthy (and perhaps not even morally responsible) for it. So long as the conditions under which a person is excused apply to everyone in every case, an invariantist theory can account for this. For more on excuses see Austin (1956–7), Zimmerman (1988) and Wallace (1994).

[9] This is true setting aside cases of derivative responsibility. I may have been casually determined to crash my car but morally responsible for having done so if I freely chose to go on a bender, and my choosing was not causally determined by the remote past and the laws of nature.

[10] On a stronger real-self view, acting in accordance with one's values or real self is both necessary and sufficient for moral responsibility.

their actions stem from, for example, either the part of the self with which they identify,[11] or their values.[12] *Reasons-responsiveness* compatibilist views claim that agents are responsible only if their actions are the result of a process that is sensitive to moral reasons in the proper ways.[13] According to such compatibilist views, Knobe and Doris claim, a single invariant standard is operative: regardless of the context, the same condition must be met in order to be morally responsible.

It is important to note that not all of these putative invariantist theories of moral responsibility that Knobe and Doris cite are accounts of the conditions for moral responsibility *as such*. Rather, they are (among other things) competing theories about the nature of the control (or freedom) condition for moral responsibility. It is true that philosophers generally focus on the control condition for moral responsibility. But unless one had the view that meeting an invariantist control condition is both necessary and sufficient for being morally responsible, adhering to an invariantist control condition on moral responsibility does not all by itself entail adherence to an invariantist theory of moral responsibility. For example: one might have a variantist epistemic or authenticity condition, which would yield a variantist account of moral responsibility. Furthermore, contrary to what Knobe and Doris claim, incompatibilism as such is not an invariantist criterion. Incompatibilism as such only tells us that one cannot be free (and therefore morally responsible) if one is determined—it does not tell us what *is* necessary for being free. One could, I think, construct an incompatibilist theory of moral responsibility that had a variantist control condition. It would tell us that the freedom necessary to meet the control condition varies by context, but that it is *always* a necessary condition that one is not determined. Knobe and Doris's examples, therefore, are not good ones.

That being said, I think we can abstract from these specific invariantist control-condition theories that Knobe and Doris provide and imagine what an invariantist theory of moral responsibility would look like: simply add to an invariantist control condition an invariantist epistemic condition and an invariantist authenticity condition. Regardless, their main point is that it has been standard practice for philosophers of moral responsibility to assume that a criterion for moral responsibility for actions ought to be invariantist. This much seems right.

The Conservativist Assumption

According to Knobe and Doris, philosophers of moral responsibility have not only been committed to the metaphysical assumption that there is an invariantist criteria set for being morally responsible for an action, they have also been committed to a methodological assumption about how we ought to go about discovering those criteria. That methodology is guided by the constraint of conservatism: that whatever the criteria for being morally responsible turn out to be, those criteria will accord with our ordinary judgments about when agents are morally responsible for what

[11] See Frankfurt (1969)
[12] See Watson (1975)
[13] See Fischer and Ravizza (1998)

they do. When we discover the right criteria, that discovery will leave our ordinary judgments largely unchanged. This methodological assumption is what we are calling the Conservativist Assumption. Because it will be helpful to have a name for them, we can call those philosophers that Knobe and Doris believe to be committed to both the Conservativist and the Invariantist Assumption 'Standard Theorists'.

The way Standard Theorists go about building and defending their theories is in large part through the method of cases. It is counted as a virtue of a Standard Theorist's account of moral responsibility if the verdicts given by a theory on a wide array of cases accord with our ordinary judgments about whether a person is morally responsible.[14] If, for example, an account of the conditions for being morally responsible gave the verdict that people are *always* responsible for actions they commit in their sleep, this would be to the detriment of the theory, for such a verdict would radically be at odds with a pre-theoretical belief about moral responsibility that says that *surely* we are not *always* morally responsible for things we do when we are unconscious.[15] Such a theory would be rejected by Standard Theorists because it violates our ordinary belief that we are not always responsible for things we do in our sleep. Knobe and Doris describe this method of cases like this:

> 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). (Knobe and Doris 2010)

The Conservativist Assumption is therefore operative in the Standard Theorist's methodology in roughly the following way. First, a Standard Theorist reflects on her (and others') ordinary beliefs about moral responsibility, considering her judgments about specific cases. For example, Ishtiyaque Haji opens his book on moral responsibility by saying:

> On the deep-seated presumption that people are morally responsible agents, we frequently *hold* each other morally accountable for some of our actions or omissions or the consequences of our actions or omissions. In turn, when we

[14] As Knobe and Doris (2010) note, *hard incompatibilism* may be a notable exception but suggest that the view's relative unpopularity may be due to its violation of the Conservativist Assumption. In fact, Pereboom (2001, esp. pp. 199–213) devotes much energy to showing how such a view is not in as much of a blatant violation of the Conservativist Assumption as many people think.

[15] We may be responsible for *some* things we do when we are asleep if, for example, we have a known past of committing violent acts in our sleep and have not taken any precautions to prevent it from happening again. At any rate, we are not *always* responsible for things we do when we sleep.

> do, we do so presumably because we believe that we *are* in fact morally responsible for these things. Furthermore, *our ordinary dealings with people in everyday life reveal that we have a rudimentary understanding of the conditions that have to be met in order for someone to be blameworthy or praiseworthy for the particular events she brings about* [italics added]. [fn. removed] We standardly believe, for instance, that a person's accessibility to praise or blame can be undermined by ignorance or force. (1998, pp. 3–4)

By reflecting on our ordinary beliefs and ordinary moral practices, Haji suggests, we can stumble upon some raw data for a theory of moral responsibility—this is the grist for the Standard Theorist's mill. Second, once we have reflected on our ordinary beliefs about when we are responsible, the Standard Theorist develops a theory of the conditions of moral responsibility that she thinks, among other things, accords with these ordinary beliefs. That is, she will develop some specific set of conditions that must be met in order to be morally responsible for an action. Third, her theory is tested by other philosophers using the method of cases: some ingenious philosopher will devise a masterfully constructed case and if her theory gives the "wrong answer" about whether someone is responsible in that case, then that is a strike against her theory. On the other hand, if her theory accords with our ordinary judgments about that case, that counts as a virtue to her theory. Fourth, if possible, the Standard Theorist revises her theory in order to accommodate a wider range of previously-unaccounted-for ordinary beliefs about moral responsibility. If the theory is ultimately judged unable to accommodate a certain case, the theory is often given up on, and the search for a better theory continues.

At both ends of this process of theory development are our two assumptions: (1) the Invariantist Assumption that there is a single set of invariantist conditions for being morally responsible for which we are all looking; and (2) the Conservativist Assumption that this single set of invariantist conditions will accord with most of our ordinary beliefs about the conditions under which agents are morally responsible for what they do.

The Empirical Conclusion

Given the research program of the Standard Theorists, if a good theory of the conditions for moral responsibility is to accord with our ordinary judgments, then the most expedient way to arrive at a good theory is to figure out what people's ordinary judgments about moral responsibility are and develop a theory based on those judgments. These ordinary judgments, Knobe and Doris claim, are the sorts of things that can be discovered and systematized by an empirical psychology. The conclusion they draw having examined the empirical literature, however, is that people *do not* draw on a single, unified theory when they attribute responsibility. Rather, the criteria people use when making these judgments are affected by at least three kinds of factors: how the relevant case is framed when presented to the judge, the moral status of the behavior, and the relationship between the person being judged and the person doing the judging. I will briefly rehearse some of the relevant literature on the first two factors. Then we'll look at the conclusions Knobe and Doris draw from it.

Abstract and Concrete Framing

In a 2005 study[16] subjects were presented with the following vignette and asked whether the agent in the story is blameworthy:

> 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 was 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 6:00PM on January 26th, 2195.

The overwhelming majority of subjects (83%) stated that Jeremy is blameworthy for the robbery, and similar results were obtained by three other studies. To the Standard Theory Compatibilist, this might be seen as vindication—being responsible in a deterministic world is not a violation of our ordinary beliefs after all!

But in another study[17] subjects were told about a Universe A, which unfolds deterministically, and were presented with only one of the following:

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

(2) 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 three 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 the family. Is Bill fully morally responsible for killing his wife and children?

Subjects' responses in the two cases reveal a startling asymmetry. Of those given Case (1), only 5% of the subjects said Bill was fully morally responsible, whereas of those who were given Case (2), 72% of the subjects said that Bill was fully morally responsible even though he is living in a deterministic universe.

What appears to be going on here, claim Knobe and Doris, is that the manner in which a case is framed can determine the set of criteria people use in making judgments of responsibility. In abstractly-framed cases, subjects appear to be utilizing a broadly incompatibilist set of conditions for moral responsibility, where in the concretely-framed cases, subjects appear to be utilizing a broadly compatibilist set of conditions. But if this is the case, then ordinary judgments of responsibility, so the

[16] Namias et al. (2005)
[17] Nichols and Knobe (2007)

story goes, are not invariantist—the same set of criteria is evidently not being applied in each and every case.

The Emotion Asymmetry

In their 2003 study, Pizzarro, Uhlmann and Salovey presented subjects with various vignettes about agents who engage in morally bad acts. In one case, subjects are told about an agent who commits a bad act in a high emotional state.

> 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.

Another class of subjects was presented with a vignette about an agent who commits the same bad act but in a low emotional state.

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

The results of the study show that subjects attributed considerably less blame to the agent acting out of a high emotional state than to the agent acting out of a low emotional state, even though in both cases the agent commits the same bad act. Another class of subjects was given another vignette about an agent who commits a morally *good* act by giving a homeless man his jacket in the freezing weather either "impulsively" or "calmly and deliberately". In *this* case, there was only a negligible difference between the praise the agent received in the low-emotion state and the high-emotion state. Therefore it appears that high-emotion states mitigate blame, but not praise. What this suggests, Knobe and Doris conclude, is that people use one set of criteria to assess good acts and another set to assess bad acts.[18]

But according to the Invariantist Assumption, it seems natural to suppose that whether a person is morally responsible for an action or not should not depend on whether the action is morally (or metaphysically) good or bad. A theory would be variantist if it gave one set of criteria of responsibility for good acts and another set of criteria for bad acts and therefore if our ordinary judgments about responsibility are invariantist, we should expect to find that people use the same criteria in assessing responsibility for both good and bad acts. But as Knobe and Doris argue, this is not what the research suggests.

[18] Knobe and Doris also cite studies that reveal intention and action asymmetries and side-effect asymmetries. The intention/action asymmetry reveals that bad actions receive more blame than good actions receive praise, but bad intentions receive twice as much blame as good intentions receive praise. See Malle and Bennett (2004). The side-effect asymmetry reveals that subjects attribute to persons who engage in acts that have bad unforeseen side-effects a high degree of blame, whereas they give persons who engage in acts that have good unforeseen side-effects a low degree of praise. See Knobe (2003). As Knobe and Doris see things, what these asymmetries show is that people use different criteria, depending on whether they are assessing acts or intentions in the first case, or bad unforeseen side-effects or good unforeseen side-effects, in the second case.

The Effect of Consequences

In what has become a classic study, Walster (1966) presented each of two classes of subjects a story about a man who parks his car atop a hill and puts on his emergency brake. He had previously known that he needed his brake cables to be serviced, but he neglected to do so. The car rolls down the hill, causing an accident. One group of subjects was told that the accident caused mild harm: the fender of a bystander's car was damaged. The second group of subjects was told that the accident caused severe harm: a young child was seriously injured. Subjects were then asked to determine whether the man had acted negligently and whether he was to blame for the accident. Both test groups believed the man to be *equally negligent*. However, subjects in the mild harm case attributed less blame than subjects in the severe harm case, even though in both cases the harm was due to an accident. If correct, then it looks as if people use a different set of criteria in assessing responsibility when the consequences of an act are severe than they do when the consequences are mild.

The upshot of all this, claim Knobe and Doris, is that there does not appear to be a single set of invariantist criteria that the folk use in attributing responsibility. Rather, what we find is that people's ordinary judgments are highly contextualized and are largely determined by factors that no moral responsibility theorists have yet thought to be relevant in determining whether an agent is morally responsible for an action. They write:

> Philosophers have searched for a single invariantist 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. (Knobe and Doris 2010)

One may be curious as to why Knobe and Doris suggest that in light of the tension between Invariance, Conservatism, and the Empirical Conclusion, Standard Theorists ought to give up Invariance instead of Conservatism. Their reasoning has to do with what they see as a more fruitful way of theorizing about moral responsibility. They write:

> 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 differently. (Knobe and Doris 2010)

It is indeed true that some have read Strawson as claiming that *holding* morally responsible is metaphysically prior to *being* morally responsible.[19] And so, the thought goes, if we are interested in the conditions under which agents are morally responsible, we should investigate the conditions under which we hold agents responsible. Such an investigation can help us understand the purposes of these moral practices in human societies, and shed light on how we could alter them in order to

[19] See, for example, Sher (2006, p. 81 ff.).

better suit those purposes. Hence, we ought to give up the metaphysical commitment to invariantism and retain the methodological commitment to conservatism.

Conservativism Revisited

In this section my discussion of Knobe and Doris's work turns critical.[20] To begin, I point out that their argument against Standard Theorists is premised on the claim that there is a significant range of philosophers who have implicitly (or explicitly) assumed both Invariantism and Conservativism, and whose methodology ought to lead them to seriously doubt their metaphysics in light of the Empirical Conclusion. Are Knobe and Doris right about this?

First, consider the claim that philosophers have been committed to the Conservatist Assumption. Knobe and Doris have provided evidence that, at least on the face of things, suggests that people's ordinary beliefs do not evince invariantism. But have philosophers really thought that their theories of moral responsibility are conservative with respect to the ordinary judgments about cases by the folk? Is *this* what philosophers have really been assuming—that the judgments that undergraduates make about cases are a good guide to a theory of moral responsibility? Maybe some philosophers think that. But others have explicitly advocated a different methodology. For example, Fischer and Ravizza explicitly state theirs:

> [W]e shall be trying to articulate the inchoate, shared views about moral responsibility in (roughly speaking) a modern, Western democratic society [...] our method will then be similar to the Rawlsian method of seeking a "reflective equilibrium" in the relevant domain. Here we shall be identifying and evaluating considered judgments about particular cases—actual and hypothetical—in which an agent's moral responsibility is at issue. (1998, p. 10–11)

But the subjects in these studies, it seems, are not engaging in any reflective equilibrium—they have simply been asked to make a judgment about a particular case or set of cases. It is difficult to see, therefore, how Knobe and Doris's challenge cuts much ice against philosophers like Fischer and Ravizza. This is because Knobe and Doris have not shown that according to Fischer and Ravizza's *own* methodology—the methodology I think Standard Theorists *actually* adopt—that we would discover a variantist criteria for responsibility, even if we *could* arrive at a variantist criteria using some *other* methodology. We could also arrive at a set of variantist criteria, I suspect, by throwing darts at a specially-made moral responsibility dartboard.

Am I presuming that philosophers who use the method of reflective equilibrium have special insight into the standards for being morally responsible? No. My present point is just that Knobe and Doris have foisted upon the Standard Theorist a methodology that many of them would reject. For while they have claimed that the Standard Theorist's methodology leads them to variantism, they have used some *other* methodology to evince putative variantist conclusions. But this is just a bait

[20] I am grateful to both Lewis Powell and Stephen White, who independently suggested that I develop the arguments in the section.

and switch. In light of this point, hardcore Standard Theorists might want to rest content with this criticism and read no further. They are invited to do so.

Invariantism Revisited

Turn now to the Invariantist Assumption. If you recall, Knobe and Doris claim that we ought to give up invariantism and embrace a Strawsonian research program, giving metaphysical priority to our practices of *holding* responsible over the "panicky metaphysics" involved in searching for the conditions of *being* responsible. There are two points to make here.

First, it is important to note that a careful reading of Strawson's "Freedom and Resentment" should raise doubts about whether Strawson was really committed to the metaphysical priority of holding responsible over being responsible. As many of his commentators have pointed out, in that same paper Strawson famously argued that there are both *excusing* and *exempting* conditions for being morally responsible and morally blameworthy for an action. According to Strawson, these conditions are facts about an agent that make it inappropriate to react to her—either situationally or globally—with the so-called reactive attitudes (e.g. resentment or indignation). But if this is true, then it is unclear how Strawson can maintain the view that holding responsible is metaphysically prior to being responsible. Just as an exegetical note, then, I am dubious that Strawson held the view that Knobe and Doris attribute to him.

And second, even if Strawson did hold the view that being morally responsible is metaphysically dependent upon our practices of holding responsible, he shouldn't have. That Strawson provided an account of excusing and exempting conditions suggests that even if he ascribed to this metaphysical dependence view, he was not fully aware of its untoward implications. As others have argued, there are serious problems with this metaphysical dependence view, the view that the facts about being responsible can be accounted for fully by the facts concerning our practices of holding responsible.[21] Indeed, even philosophers who count themselves as wholly committed to the Strawsonian program have rejected this putative tenet, and, I think, rightfully so.[22]

Recall that Knobe and Doris concluded that if forced to choose between conservativism and invariantism, we should adopt conservativism. Their reason was that there is another way of thinking about moral responsibility—a Strawsonian way—that does not focus on the conditions for moral responsibility, but rather focuses on the moral responsibility practices in which people actually engage. And so given the Empirical Conclusion, people appear to apply variantist criteria in judging agents morally responsible. My goal in this section has simply been to push back against this line of argument. For even if Strawson did hold this metaphysical dependence view, he ought not have. And further, we can

[21] See, for example, Sher (2006) and McKenna (*Conversation and Agent Meaning: A Theory of Moral Responsibility*, unpublished)

[22] See McKenna (unpublished), who denies that holding moral responsible is metaphysically prior to being morally responsible, but instead argues that they are conceptually connected in a way such that one cannot fully understand the one without making reference to the other. Nevertheless, he accepts two other Strawsonian tenets: the claim that the morally reactive attitudes are central to understanding the nature of being morally responsible, and that what is most fundamental with respect to whether an agent is blameworthy or praiseworthy is the nature of the agent's "quality of will" with which she acted.

embrace a broadly Strawsonian project without thinking that the practices of holding responsible enjoy metaphysical priority. But setting aside that worry, we can still ask whether the psychological literature that Knobe and Doris cite provide evidence for the Empirical Conclusion at all. To that question we now turn.

The Empirical Conclusion Revisited

I begin by making what I hope to be an uncontroversial claim: in devising an experiment, if the goal is to assess and systematize the ordinary judgments of people about *the conditions under which agents are morally responsible,* then we had better make sure that the test subjects are actually issuing judgments about *that* and not something else. This might sound like an obvious point. Sometimes, however, we need to be reminded of the obvious. For as the attentive reader might have already noticed, the studies that Knobe and Doris discuss are not all of a piece. In fact, we can separate these studies into three broad groups. Call the group of studies in which subjects are asked to assign a certain degree of blame or praise "Group A" studies. This group would include, for example, the 2003 Pizarro, Uhlmann, and Salovey "emotion asymmetry" studies as well as the 2004 Malle and Bennett "intention/action asymmetry" studies. Next, call the group of studies in which subjects are asked to ascribe blameworthiness or praiseworthiness "Group B" studies. This group includes the 2005 Nahmias, Morris, Nadelhoffer, and Turner "Fidelity Bank" study and the 2003 Knobe "side effect" study. Third, call the group of studies in which subjcts are asked to ascribe moral responsibility "Group C" studies. The 2007 Nichols and Knobe "Universe A" studies would be included in this group.

The first thing to notice here is that these studies are fragmented—they are not all asking the same question of subjects, and therefore on the face of things appear to be testing different kinds of judgments. The natural worry, then, is that this body of evidence simply does not provide unified evidential support for the Empirical Conclusion. Is this worry justified? I think it is, and to show why, let's address each of these groups of studies severally and see whether any of them are in a position to support the Empirical Conclusion.

Group A

We begin with the studies that reveal asymmetries in how subjects assign degrees of blame and praise.[23] Because there are asymmetries in the *degree* of blame or praise

[23] What might we mean by "degrees of blame and praise?" Take, for example, the distinction that Michael Zimmerman makes between what he calls *weak overt blame* and *strong overt blame* (e.g., Zimmerman 1988, p. 149). A case of weak overt blame might involve the blamer merely uttering a blaming judgment. In a case of strong overt blame, a blamer brings it about that the blameworthy person suffer as she deserves. But even if one disagreed with this assessment, there is good reason to think that there is a spectrum across which one can assign blame. Overt blame can be as weak as a knowing scowl. It can be as strong as stern censure, or perhaps sanctions carried about by a special authority. One could give similar examples of degrees of moral praise. We can also assign degrees of private blame and praise: consider the difference between the experience of "cool" moral protest and "hot" resentment or ill-will. When we assign degrees of praise and blame, then, we might be understood as assigning to an agent some appropriate (or deserved) response that falls along some relevant spectrum.

that subjects attribute to agents, Knobe and Doris claim that this supports the view that ordinary folk use a variantist criteria for attributing moral responsibility. But this does not follow. There is nothing in the claim that people use a variantist standard in determining *how much* blame or praise they think an agent deserves that supports the claim that people use different standards in determining whether someone is morally responsible *at all.*

Here's why. It is a widely held view that being blameworthy or praiseworthy for an action is sufficient for being morally responsible for that action. This means that whatever the criteria happen to be that people use in judging that an agent is blameworthy or praiseworthy for some action, once *these* criteria are met, then the criteria for being morally responsible are met *also*.[24] If this is right, then—and this is important—the *degree* of blame or praise a subject assigns to an actor is irrelevant to whether a person is judged responsible *at all*.[25] Whether I assign a low degree of blame (say, .1) or a high degree (say, .9), this is irrelevant to whether I judge you to be *morally responsible* full stop. An invariantist theory *of moral responsibility* need not entail the further claim that there are invariantist criteria for correctly attributing degrees of deserved blame or praise. Indeed, we have two very good reasons for thinking that the criteria for correctly attributing degrees of deserved blame or praise *ought* to depend on highly specified contextual features.

First, consider two agents who commit the same kind of bad action, say robbing a convenience store. It is reasonable to hold the following: (a) that both agents are morally responsible for their respective bad actions, but (b) that because one agent had a worse upbringing, she is *worthy of less blame* than the agent who had a better upbringing. Historical and situational features can affect the degree of blame one deserves without changing the fact that both agents meet the conditions for being morally responsible for their respective acts. While the criteria for being morally responsible for that bad action are not variantist, it is quite plausible to think that contextual features can play a role in determining *how much blame an agent deserves*.

Second, consider a case in which an acquaintance, let's call her Jill, tells an inappropriate joke during a dinner party. Suppose that everyone at the table correctly judges that Jill is morally responsible for telling that joke and everyone at the table also correctly judges that she is morally blameworthy for having told the joke. We can distinguish between: (a) S judging the degree to which Jill is blameworthy or praiseworthy, and (2) S judging the degree to which it would be appropriate *for S* to blame or praise Jill.[26] These may diverge. For while I may judge correctly that Jill deserves a certain degree of blame, it may not be appropriate for *me* to blame her to that degree (or even at all), if, for example, I am also disposed to tell similarly inappropriate jokes and therefore I do not have the kind of moral standing to blame

[24] There are exceptions to this generalization: George Sher, in his 2006 discussion of moral responsibility for one's character, argues that one can be morally blameworthy for one's *character* without being morally responsible for it. He thinks this because he holds that moral responsibility is a causal notion, and we can be blameworthy for our characters without having caused them. Our discussion here centers on responsibility for actions, however, and Sher would agree, I think, that at least in the case of actions, being morally blameworthy entails being morally responsible. Cf. Scanlon (2008).

[25] Randolph Clarke raised a similar point in conversation.

[26] Thanks to Michael McKenna for pressing me to make this distinction clear.

her that others at the table have. Therefore the degree of blame or praise that it would be appropriate for you to direct at Jill might be different than the degree of blame or praise it would be appropriate for me to direct at Jill. These differences are putatively explained by contextual features.

The point here is that the invariantist about moral responsibility for actions can embrace a certain kind of contextualism about degrees of blame and praise, both in the criteria that determine to what extent an agent deserves blame or praise, and what degree of blame or praise is appropriate for any given agent to blame or praise her. Should we call these criteria variantist? I don't know. But Knobe and Doris have given us no reason to think that even if the folk use a variantist criteria for judging degrees of blame or praise that this reveals something about the conditions that people use to determine whether an agent is morally responsible full stop. Certainly, one could hold that situational and relational features can affect the degree of praise and blame that might be appropriate in a given case. But if this is right, then those studies that ask subjects to attribute a certain *degree* of blame or praise along a scale cannot, even in principle, show that there is not a single invariantist set of criteria of moral responsibility that is used in ordinary judgments. If a judgment of an agent's blameworthiness entails that one is judged to be morally responsible (though such a judgment need not be made consciously), then any studies that trade on folk judgments about *degrees* of blame are moot. Therefore, the asymmetry studies about degrees of blame and praise that Knobe and Doris cite in their defense of variantism about responsibility cannot be used to support the Empirical Conclusion.

Group B

In this second group of studies, experimenters asked subjects whether actors in the vignettes were worthy of blame (or praise) *full stop*. Given the ways that the subjects responded, Knobe and Doris claim that this group of studies provides evidence for the Empirical Conclusion—that the folk are not using an invariantist criterion for attributing moral responsibility. Does this follow? I think that there are at least two good reasons for thinking that it does not.

The first reason is simply the fact that a judgment that someone is worthy of blame (or praise) is not logically equivalent to a judgment that someone is morally responsible. The reason for this is that, as some philosophers have noted, there is an important conceptual gap between moral responsibility on the one hand, and blameworthiness and praiseworthiness, on the other. For example, John Martin Fischer writes:

> Moral responsibility [...] is more abstract than praiseworthiness or blameworthiness: moral responsibility is, as it were, the "gateway" to moral praiseworthiness, blameworthiness, resentment, indignation, respect, gratitude, and so forth. Someone who is morally responsible is an *apt candidate* for moral judgments and ascriptions of moral properties; similarly, a morally responsible agent is an *apt target* for such attitudes as resentment, indignation, respect, gratitude, and so forth. Someone becomes an apt target—someone is "in the ballpark" for such ascriptions and attitudes—in virtue of exercising a distinctive kind of control ("guidance control"). (2006, p. 233)

He continues:

> In my view, further conditions need to be added to mere guidance control to get to blameworthiness; these conditions may have to do with the *circumstances under which one's values, beliefs, desires, and dispositions were created and sustained, one's physical and economic status, and so on.* (2006, p. 233, italics added)

If this "gateway view" of moral responsibility for actions is right, then it is illicit to draw conclusions about the conditions under which agents attribute moral responsibility from the conditions under which agents attribute or blame- (or praise-)worthiness. There is an important conceptual space between these two sets of judgments, and therefore a judgment of blameworthiness or praiseworthiness is different than a judgment of moral responsibility.[27]

One might object to this criticism by pointing out that being blame- (or praise-) worthy entails being morally responsible. It then follows that the conditions for being responsible are *already* met if one judges an agent to be blameworthy. Therefore, if we discover that subjects have a variantist criterion for attributing blameworthiness, this entails that they have a variantist criterion for attributing moral responsibility.

This first part is right. But this conclusion would only follow if blameworthiness *just is* being morally responsible for a bad action. If *that* were true, then we could infer from a variantist criterion for blameworthiness a variantist criterion for moral responsibility. But the view that one is blameworthy for an action just in case one is morally responsible for a bad action is a substantive philosophical thesis, one in need of argument and one for which they have offered no argument.

Indeed some philosophers explicitly deny that this is what blameworthiness amounts to.[28] If Fischer is right, then exercising guidance control is a sufficient condition for being morally responsible for an action, but there are extra conditions that must be met in order for an agent to be morally responsible *and* blameworthy. So suppose that Fischer is right that those extra conditions that need to be met are determined by the specific context of the agent's action. This would leave open the possibility that while the criteria for being morally responsible for an action are independent of context, the criteria for being blameworthy *are* context dependent. But while this is consistent with the Group B studies, it is not consistent with the Empirical Conclusion. This should lead us to think that the Group B studies do not provide the kind of evidence that Knobe and Doris claim.

But is this distinction between conditions for being morally responsible and conditions for being blameworthy (or praiseworthy) simply a philosopher's fiction? Why think this conceptual point tracks anything in our actual moral practice? This is a good question. But as it turns out, there does seem to be a modicum of empirical

[27] This is perhaps most clearly seen in cases in which one can be morally responsible for morally neutral actions, actions for which an agent is neither praiseworthy nor blameworthy. For more on moral responsibility for morally neutral actions, see Zimmerman (1988), Haji (1998), Fischer (2006), and McKenna (unpublished).

[28] See, for example, the passage from Fischer (2006) cited above and McKenna (unpublished).

evidence supporting a "gateway view" of moral responsibility. For example, Critchlow concluded that "judgments of cause, responsibility, blame, and punishment, although related to each other [with correlations ranging from .20 to.70], should not be taken as measures of the same thing."[29] One reason for thinking that they should not be taken as measures of the same thing is because Harvey and Rule (1978) have shown that when subjects are asked to rate an actor who caused a harm along a number of responsibility-related dimensions, their responses revealed that judgments were made in two sets of distinct clusters, one associated with *responsibility*, and another associated with *blame* or *moral evaluation*. In other words, the kind of responsibility-related concept one applies to a case plays a role in the kinds of judgments that subjects make about the case. This suggests that the criteria people use in attributing blameworthiness (or praiseworthiness) are different than the criteria people use in attributing moral responsibility. But if this is right, then we ought not draw conclusions about the conditions under which people attribute moral responsibility from studies that ask subjects to attribute blameworthiness (or praiseworthiness). This undermines the support that Knobe and Doris find within the Group B studies for the Empirical Conclusion.

Group C

Thus far I have argued that neither the studies in which subjects assigned degrees of blame or praise (Group A), nor those in which subjects ascribe blameworthiness (or praiseworthiness) (Group B) support the Empirical Conclusion. Those studies fail to evince the Empirical Conclusion because they were asking subjects to attribute things other than moral responsibility, which is what the Empirical Conclusion is a claim about. But if we want to know whether the folk are variantists about moral responsibility, then it is moral responsibility that we should be asking them to attribute in our surveys. One study that Knobe and Doris discuss does indeed ask subjects to attribute moral responsibility to actors: the 2007 Nichols and Knobe study in which Bill kills his wife and family suggests that subjects are less likely to attribute to him moral responsibility in the "abstract" scenario than they are in the "concrete" scenario. Does this support the Empirical Conclusion?

First, consider what Nichols and Knobe themselves concluded about their own study:

> Our hypothesis is that when people are confronted with a story about an agent who performs morally bad behavior, this can trigger an immediate emotional response, and this emotional response can play a crucial (distorting) role in their intuitions about whether an agent was morally responsible. In fact, people may sometimes declare such an agent to be morally responsible despite the fact that they embrace a theory of responsibility on which the agent is not responsible. (2007, p. 664)

On Nichols and Knobe's preferred explanation for what is going on in these studies, those subjects who are presented with the concrete cases in which more details about

[29] See Critchlow (1985, p. 271). For a meta-analysis of 75 responsibility attribution studies (some of which are those cited by Knobe and Doris) see Robbennolt (2000).

the bad behavior are given are committing a *performance error*. Very roughly, a psychological performance error is the result of the malfunctioning of some psychological competence, that is, a psychological mechanism. When there is interference or foul play, the mechanism issues a performance error. Linguistic examples like spoonerisms are helpful. If the department chair asks "Is the bean dizzy?" when she means to ask "Is the dean busy?" this does not mean that she doesn't understand the meanings of the relevant English words (i.e. a competence error). Rather, she just had a slip of the tongue, perhaps due to some slight anxiety. The gears of the relevant psychological mechanism were just temporarily gunked up.

In the present case, Nichols and Knobe argue that while subjects in these cases possess a moral competence that issues in judgments about when agents are responsible, when they are given vignettes that trigger an emotional response, the resulting affect gets in the way of the moral competence issuing its judgment. This is not to claim that these performance errors are incorrect judgments. Rather it is to claim that such judgments do not reflect the competence of an agent in attributing moral responsibility. Or another way to put it: these judgments do not reflect the implicit theories that subjects have about the conditions for being moral responsible; rather they reflect what happens when their theories run into interference.

Now this "performance error" model is only one plausible explanation for what is going on in these studies, and Nichols and Knobe discuss others before rejecting them.[30] The issues here are complex and to treat them fully would demand another paper. Fortunately, however, we need not enter into those issues here. What is relevant here is that on Nichols and Knobe's *preferred* interpretation of their own studies, the fact that subjects ascribe moral responsibility asymmetrically is attributed to a *performance error*. This raises problems for Knobe and Doris.

The first is that if the Empirical Conclusion is intended to be a claim about moral competence (i.e. a claim about the implicit theories that the folk have about the conditions for being morally responsible for an action), then the studies Knobe and Doris have cited in support of that conclusion fail to do so if some of the subjects' responses are performance errors. If the performance error model is correct, then these studies *do not* provide evidence for the claim that the folk have a variantist theory of moral responsibility. Instead of observing subjects using a variantist criterion for moral responsibility attribution, we would be observing subjects committing performance errors that give the *appearance* of subjects applying a variantist criterion. If this model is correct we would be no more justified in claiming that these subjects have a variantist account of moral responsibility than we would be in claiming that the department chair doesn't understand English vocabulary.

Of course, another option would be to claim that the Empirical Conclusion is the kind of claim that can be evinced by performance errors. But if this is all that the Empirical Conclusion is claiming, the invariantist should be unmoved—surely the invariantist will concede that people can make performance errors even if they possess an invariantist theory of moral responsibility.

What Knobe and Doris would need to do, it seems, is to reject the performance error model of Nichols and Knobe. They could then supplement these extant studies with new studies that evince the Empirical Conclusion and show that their results

[30] But see Kimbrough (2009) for a criticism of this performance error model interpretation.

trade on no performance errors. This option seems to me to be exactly what Knobe and Doris need to do to defend their conclusions about these Group C studies.[31] Until then, however, we should be skeptical about the Empirical Conclusion.

If I may be so bold, I will conclude this section with a short story and a bit of friendly advice. First the story: the philosophical literature on moral responsibility and cognate topics over the past fifty years or so is a messy and at times confused literature. The primary culprit for this messiness and confusion is that philosophers have often not been careful to explain clearly what they have in mind when they discuss "moral responsibility." The result has been that a lot of philosophers have spent a lot of time talking past each other. It has only been fairly recently, however, that philosophers have begun to untangle and delineate all of the various ways of understanding "moral responsibility."

Now the friendly advice: Given the fact that even people who get paid to think about responsibility have been at times confused about exactly what is at issue, experimenters who design studies testing for folk beliefs about "moral responsibility" ought to be careful about what their subjects think they are being asked to do in these studies. For example, if experimenters are interested in discovering the conditions under which a subject attributes "moral responiblity" for an action, they should be clear about what it is they are trying to find. Are they interested in discovering the conditions under which the folk determine whether an actor is morally responsible in the sense of being *accountable* (as in Watson's 1996 sense) for her action? If so, then we should do our best to make sure that the subjects understand that this is what is being asked, so that we can prevent them from inadvertently making a judgment about some other, related moral responsibility concept, for instance: (1) role responsibility (the fulfillment of certain role-specific social duties), (2) causal responsibility (being the cause of a certain act), (3) legal liability (having broken the law and being liable to punishment for it),[32] (4) blameworthiness, (5) appropriateness of the experimental subject's blame, (6) praiseworthiness, (7) appropriateness of the experimental subject's praise, (8) guilt for wrongdoing, (9) desert of sanction or punishment, (10) moral responsibility as "attributability" (in Watson's 1996 sense), and (11) degree of deserved praise or blame. To the extent that social psychologists and experimental philosophers can improve and clarify their experimental designs along these lines, their results will carry a greater force.

Conclusion

Knobe and Doris have argued that a significant range of philosophers working on moral responsibility are committed to an inconsistent triad: their consersativist methodology and their invariantist metaphysics are in conflict with the alleged empirical discovery that the folk's ordinary judgments about moral responsibility are

[31] It is important to note that a performance error model could be used to explain the results of *all* of the studies cited by Knobe and Doris, not just those from Group C. As we have seen, however, those studies from Groups A and B have other serious flaws that should prohibit us from using them as evidence for the Empirical Conclusion.

[32] For further explanation of these first three responsibility-related concepts see Hart (1968, ch. 9).

variantist. Therefore, they conclude, in a Strawsonian spirit we should seriously consider abandoning invariantism and adopt a variantist outlook on the conditions for being morally responsible for actions, an outlook that opens up new avenues for understanding our moral responsibility practices.

I have claimed that there are a number of things wrong with this line of argument. First, I have argued that even if the empirical literature that they cite does indeed evince the Empirical Conclusion, it does so using a methodology that most philosophers do not endorse, and therefore cannot be used to show that *those* philosophers are committed to any inconsistency. You cannot show that a philosopher's methodology leads her into troubles on the grounds that she uses a methodology that she doesn't actually use. Second, I have argued that the Strawsonian convictions that Knobe and Doris present in favor of variantism fail to hit their target. Third, I have argued that the psychological literature that Knobe and Doris cite in favor of the Empirical Conclusion is just not good evidence for the claim that the folk use a variantist theory for attributing moral responsibility: some studies are asking the wrong kinds of questions, and all of them are plausibly interpreted as eliciting performance errors, not a variantist folk theory. For the moment at least, moral responsibility invariantism is safe, if not fashionable.

Acknowledgements I am grateful to audiences at UCLA and Florida State University and to Michael McKenna, Randy Clarke, Stephen White, and Lewis Powell for their comments on previous drafts.

References

Austin, J. L. (1956–7). A plea for excuses. *Proceedings of the Aristotelian Society, 57*, 1–30.

Critchlow, B. (1985). The blame in the bottle: attributions about Drunken Behavior. *Personality and Social Psychology Bulletin, 11*, 258–274.

Fischer, J. (2006). *My way: Essays on moral responsibility.* New York: Oxford University Press.

Fischer, J., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility.* Cambridge: Cambridge University Press.

Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy*, *66*(3), 829–839.

Haji, I. (1998). *Moral appraisability: Puzzles, proposals, and perplexities*. New York: Oxford University Press.

Hart, H. L. A. (1968). *Punishment and responsibility: Essays in philosophy of law.* Oxford: Clarendon.

Harvey, M. D., & Rule, B. G. (1978). Moral evaluations and judgments of responsibility. *Personality and Social Psychology Bulletin, 4*, 583–588.

Kimbrough, S. (2009) Explaining compatibilist intuitions about moral responsibility: a critique of Nichols and Knobe's performance error model. *Florida Philosophical Review.* IX(2)

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63*, 190–193.

Knobe, J., & Doris, J. (2010). Strawsonian variations: Folk morality and the search for a unified theory. In J. Doris (Ed.), *The handbook of moral psychology.* Oxford: Oxford University Press (in press).

Malle, B., & Bennett, R. (2004) People's praise and blame for intentions and actions: Implications of the folk concept of intentionality. Technical Reports of the Institute of Cognitive and Decision Sciences, No. 02-2, Eugene, Oregon.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom: folk intuitions about free will and moral responsibility. *Philosophical Psychology, 18*, 561–84.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: the cognitive science of folk intuitions. *Nous, 41*(4), 663–685.

Pereboom, D. (2001). *Living without free will*. Cambridge: Cambridge University Press.

Pizarro, D., Uhlman, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: the role of perceived metadesires. *Psychological Science, 14*, 267–72.
Robbennolt, J. (2000). "Outcome severity and judgments of 'responsibility': a meta-analytic review. *Journal of Applied Social Psychology, 30*, 2575–2609.
Scanlon, T. M. (2008). *Moral dimensions: Permissibility, meaning, blame*. Cambridge: Harvard University Press.
Sher, G. (2006). *In praise of blame*. Oxford: Oxford University Press.
Sher, G. (2009). *Who knew?: Responsibility without awareness*. Oxford: Oxford university Press.
Strawson, P. F. (1962) Freedom and resentment. *Proceedings of the British Academy*, *48*.
Wallace, R. J. (1994). *Responsibility and the moral sentiments*. Cambridge: Harvard University Press.
Walster, E. (1966). Assignment of responsibility for an accident. *Journal of Personality and Social Pscyhology, 3*, 73–79.
Watson, G. (1975). Free agency. *Journal of Philosophy, 72*, 205–20.
Watson, G. (1996). Two faces of responsibility. *Philosophical Topics, 24*, 227–248.
Watson, G. (2004). *Agency and answerability: selected essays*. Oxford: Clarendon.
Zimmerman, M. (1988). *An essay on moral responsibility.* Totowa: Rowman and Littlefield.