

Do We Have A Coherent Set of Intuitions About Moral Responsibility?¹

Introduction

For quite a while, philosophers have been struggling to answer fundamental questions concerning the nature of freedom and moral responsibility: Can we be responsible for our actions, and, if so, under what conditions? Must we have been able to do otherwise than what we actually do in order to be responsible? Must it be the case that our actions are not determined, and that the true theory of physics must be an indeterministic one in order for us to be responsible?

There is great debate about how to answer these questions, and a variety of ingenious strategies are employed in efforts to settle them. One prominent strategy includes the construction of cases that possess one or another of the features under consideration and claims about our intuitive responses to them. For example, a strategy employed by some incompatibilists—those who believe responsibility is incompatible with determinism—is to point to a variety of cases to support the conclusion that responsibility requires the ability to do otherwise. We are not responsible for failing to act when we are tied up with a rope and cannot escape, for example. If responsibility does in fact require the ability to do otherwise, then, so one line of reasoning goes, it also requires the falsity of determinism. On the other hand, compatibilists often try to counter this kind of argument by an appeal to intuitive responses to other sorts of cases. For example, in what are now known as “Frankfurt-style” cases, a person performs an action, but apparently couldn’t have done otherwise because had she seemed to waver in the intention to perform it, another individual would have intervened to ensure that she did. Nevertheless, as long as the “counterfactual intervener” does not actually play any causal

role in the agent's action, we are invited to react with the intuition that the agent was responsible for her action, despite lacking the ability to do otherwise.² A remarkable literature has developed concerning just the Frankfurt-style examples.

This literature alone provides a window on the way that the debates hinge to a great extent on claims about our judgments of responsibility in certain kinds of cases. Philosophers often go further, too, in drawing more general conclusions about what our natural theoretical views on these issues are. For example, in several recent influential books and articles, philosophers claim that we are “natural” incompatibilists—that is, we are incompatibilists before we (compatibilists) start studying philosophy and talk ourselves out of it.³

In recent years, some philosophers have decided that some of the claims about what “we” find intuitive could use some systematic testing, and they have begun to do just that by using pencil-and-paper surveys of undergraduates and people quite literally “on the street.” Eddy Nahmias and his colleagues were among the first philosophers to undertake this kind of research, and they have found some surprising and intriguing results that suggested to them that perhaps students were natural *compatibilists*, contrary to the often uncontested claims of many incompatibilists.⁴ Since then, the data has accumulated, and it is proving to be more recalcitrant to either side's theories than it seemed at first and even second glance. In fact, the data concerning the conditions under which we make judgments of responsibility is already so interestingly complex that one might naturally conclude that our concept of responsibility must be much more complicated than philosophers have assumed—if not simply completely confused. For example, Manuel Vargas suspects that the data “reflects the operation of distinct psychological mechanisms involved in the metaphysics of responsibility, and the pragmatics of holding people responsible”, and he suggests that we revise our concepts and self-conception in

light of this data.”⁵ And in a recent paper, Joshua Knobe and John Doris present much of this data (a considerable portion of which was also collected by them and Shaun Nichols), and draw a relatively radical conclusion. They argue that the assumption that there is a single set of criteria for moral responsibility that should apply in all cases is false. In their terminology, the assumption of “Invariantism” is false, and can be shown to be false by data that simply resist any single coherent explanation in terms of a unified concept of responsibility.

I believe that the data is both fascinating and instructive, but in this paper I will resist the conclusion that we must give up Invariantism, or, as I prefer to call it, Unificationism. In the process of examining the challenging data and responding to it, I will try to draw some larger lessons about how to use the kind of data being collected. First, I will provide a brief description of some influential theories of responsibility, and then explain the threat to them from the experimental results. Finally, I will set out my general approach to the data, as well as some specific suggestions about how to think about each set of experiments. I will conclude that philosophers searching for a unified theory need not give up, but that at the same time they can learn a great deal from the new data.

Unifying Theories of Responsibility and the Nature of the Threat

As noted earlier, a major dividing line among theories of responsibility is that between compatibilism and incompatibilism. Incompatibilism is generally motivated in two ways: determinism is thought to preclude responsibility because it takes away the ability of an agent to act other than she actually does, and also because it shows that the agent is not the source of her action in a fundamental sense.⁶ Incompatibilists are thus united in including the falsity of

determinism as one of their conditions of responsibility, even though they often differ about what other conditions must obtain for one to be responsible.

Compatibilists explicitly reject the requirement of indeterminism for responsibility, while offering positive conditions in its place. It will be useful to consider two groups of such views here: real self views and reasons-responsiveness views. According to real self views, one is responsible for an action if and only if one's action flows from one's "real self" ("real self" can then in turn be cashed out in a variety of ways, including one's valuational system and what one identifies with.)⁷

In contrast to real self views, reasons responsiveness views require an explicit capacity to respond to certain sorts of reasons, but like real self views, are compatibilist. According to one reasons-responsiveness view, one is responsible for an action if and only if one acts with the ability to do the (or a) right thing for the right reasons.⁸ (I will call this the "reason view", following Susan Wolf.) It is worth noting at this point that there is a built-in asymmetry to this view: one needs the ability to do otherwise in order to be responsible while acting against reasons, but not in order to act for them.

Although this is a very brief description of just a sampling of theories, I believe it will be all that we will need to proceed. For we can already see that these theories, both the compatibilist and incompatibilist ones alike, are unifying theories in the sense that they offer conditions for responsibility that do not vary among speaker or subject contexts. As Knobe and Doris put it, "It is supposed to be completely obvious [according to unifying theories], and hence in no need of justification or argument, that we ought to apply the same criteria in all cases rather than applying different criteria in different cases." (2) In my view Knobe and Doris are correct: participants in the debate have tended to assume that some such criteria can be found. And I

believe that the participants go further: they believe that such criteria exist precisely because there really is a unified set of criteria for *being responsible*.

At the same time, it is important to note that a unifying theory can be perfectly consistent with there being degrees of responsibility, and even with there being gray areas in which it is unclear whether one is responsible or not. It is also possible that a unifying theory can include at least some disjunctive conditions at some level of description, as long as there is a general unity to the theory. For example, it is possible that possessing the abilities required for responsibility to which the reason view appeals in turn require possessing some set of specific perceptual and cognitive abilities. At some levels of description, these might be manifested in any of a number of ways. For example, there might be a variety of ways that one might come to recognize a morally salient fact, any of which would satisfy the unifying conditions on moral responsibility.⁹,

¹⁰

We are now in a position to articulate the threat to unifying theories of responsibility. We have seen that unifiers accept what I am going to call the “Unity Assumption”:

(Unity Assumption) There is a single set of conditions for moral responsibility that applies in all cases.

As we have also seen, unifiers also appeal to intuitions to support their theories over competing unifying theories. So they accept another assumption that I’ll call the “Fit Assumption”:

(Fit Assumption) The criteria for moral responsibility attributions fit with all (or most) of our ordinary judgments.

The problem for unifiers now arises because empirical data show that

(Empirical Conclusion) There is no plausible single set of criteria that fit with all (or even most) of our ordinary judgments.

Given the Empirical Conclusion, it follows that either the Unity or the Fit assumption must be false. And in either case, the consequences appear quite serious—it seems that either we must give up hope of finding a single set of criteria that apply in all cases, or we give up one central argumentative strategy of appealing to intuitions in assessing theories of responsibility.¹¹

In the next section, I will explain the basis on which the Empirical Conclusion rests. But first, to anticipate, I will agree that taken in one sense of “fit”, the Empirical Conclusion is true and interesting. Yet, as I will argue, on this same sense of “fit”, the Fit assumption is false, or at least in need of further defense.

The Data and the Drawing of the Empirical Conclusion

A. The Abstract v. The Concrete

Are people really natural incompatibilists, as has been assumed by so many philosophers?

Some recently collected data strongly suggests that the answer is “no.” Nahmias, Morris,

Nadelhoffer, and Turner conducted a series of experiments in which subjects were presented

with vignettes about agents who act immorally in a deterministic world, and then asked

whether the agents were morally responsible for their actions. For example, subjects were given the following story:

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.

After being asked whether Jeremy is blameworthy for robbing the bank, 83% of the subjects answered "yes," and a natural conclusion to draw is that people are actually natural *compatibilists*.¹²

But this is not the last word. Hypothesizing that philosophers who reached the opposite conclusion had some genuine evidence on the basis of informal polling of students, Nichols and Knobe (forthcoming) wondered whether the results of Nahmias et al were due to certain particular features of the stories provided to subjects. In particular, they hypothesized that the results depended on the fact that the actions in question were described in very concrete terms that elicit high affect, and that the results would not be the same if the action descriptions were more abstract. To test the hypothesis, Nichols and Knobe presented subjects with a story about Universe A, a universe in which events always unfold according to deterministic laws. Subjects were then divided into two groups. Those in the "abstract condition" were asked whether in Universe A it is possible for "a person to be fully morally responsible for their actions". Subjects in the concrete condition were given a detailed description of a man named Bill who inhabits universe A and who, upon becoming attracted to his secretary, and deciding that the only way to

be with her is to kill his wife and three children, sets fire to his house, thereby killing them. They were then asked whether Bill is morally responsible for this particular action. The results seem to have borne out the hypothesis: 72% of subjects answered “yes” in the concrete condition, while only 5% answered “yes” in the abstract condition. Doris and Knobe conclude that “[i]f this pattern of results is replicated...we will have good reason to believe that no invariantist theory...can capture all of people’s ordinary judgments” (10-11).

B. Moral Status Asymmetries

But there is more data to consider before drawing a final conclusion. There are at least four kinds of cases in which there seem to be asymmetries in people’s judgments about responsibility, based on differences in the moral status of the relevant actions. This might not be troubling, if the asymmetries functioned in the same way, but they don’t. For example, it seems that when it comes to bad actions, people are more likely to see acting with great emotion as a responsibility (or at least blame) mitigator, while when it comes to good actions, acting with great emotion makes little difference to judgments of responsibility.¹³ On the other hand, when judging responsibility for *intentions* relative to actions, the data suggest that there is little difference in responsibility judgments for bad intentions and their corresponding actions, while there is a great deal of difference in responsibility judgments concerning the possession of good intentions and performing corresponding actions.¹⁴ Sometimes having a good moral status leads to greater responsibility, sometimes to less. Let us examine the data in a bit more detail.

Emotion Asymmetry

Pizzaro, Uhlmann, and Salovey (2003) provide some intriguing results concerning how people judge agents who act with a great deal of emotion. They offered one group of subjects a vignette about 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.” They presented another group with this vignette about 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.” Contrasting cases for each of these were presented to other subjects in which Jack either does the good or bad action, but this time “calmly and deliberately”. It turns out that subjects judged the agents considerably less blameworthy when they acted badly with emotion, as compared to when they acted badly in a calm and deliberate manner. At the same time, there was a negligible difference in the degree of praise for good actions done with great emotion and good actions performed calmly and deliberately. Thus, acting with great emotion mitigates blame in the case of bad actions, but does not affect the level of praise when it comes to good actions.

Intention and Action Asymmetry

Malle and Bennett (2004) ran a series of experiments to test the hypothesis that there is a higher degree of blameworthiness for bad intentions relative to bad actions than praiseworthiness for good intentions relative to good actions. They invited subjects to judge how responsible agents were for each of the following actions or intentions:

Positive pair:

[action] helped a neighbor fix a roof

[intention] intends to help a neighbor fix a roof

Negative pair:

[action] sold cocaine to his teenage cousin

[intention] intends to sell cocaine to his teenage cousin

It turns out that in general intentions get less praise and blame than their corresponding actions, but there is *twice* as much discounting of praise when it comes to intentions as discounting of blame. For example, forming the intention to help a neighbor fix his roof engenders half as much praise, relative to actually helping him, as the blame engendered by forming the intention to sell cocaine, relative to actually selling it.

Side Effect Asymmetry

A third asymmetry can be found when it comes to judgments of responsibility for foreseen side-effects. It turns out that people judge agents as having a high degree of blameworthiness for bad unforeseen side effects, and offer hardly any praise for good unforeseen side effects. Knobe (2003a) gave subjects two vignettes that confirmed this hypothesis:

Harm Condition

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.

Help Condition

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.

Most subjects said the vice-president deserved blame in the Harm Condition, while few said he deserved praise in the Help Condition. When it comes to side-effects, then, we have a new kind of asymmetry. Doing good things (as side-effects) does not engender praise; doing bad things (as side-effects) does.

Severity

A fourth asymmetry in people’s judgments is revealed when the severity of harms is varied.

Even when harm is done accidentally, we sometimes still judge people responsible for the harm if we think that negligence was involved. There is interesting data suggesting that the severity of the harm affects where we draw the threshold of negligence, even though the severity of the harm is sometimes simply a matter of luck and not something one would antecedently think is relevant to the attribution of responsibility. Nevertheless, Walster (1996) showed that people are more likely to say that agents are responsible for a severe harm (e.g., serious injury to an innocent child) than they are to say that agents are responsible for a mild harm (e.g., a damaged fender), even when in both cases the agent is described as performing the same negligent action (e.g., parking on a hill, having failed to have his brakes checked).¹⁵ While these results do not reveal a moral status asymmetry between praise and blame in the way that the first three kinds of

results do, they do reveal a kind of “levels-of-harm” asymmetry, and in any case present an interesting phenomenon in need of explanation.

A Defense of Unifiers

The data does present a challenge. The data demonstrates that there is no easy “derivation” of particular judgments of responsibility from any internalized unified theory of conditions for responsibility together with obvious features of the cases. So if the Fit assumption requires something like a kind of straightforward derivation of (most) judgments from a unified theory, then the data would show that either it or the Unity assumption would have to go.

However, I believe that unifying theories can meet the challenge and preserve both assumptions, at least on one reading of the Fit assumption. I do not mean to defend either assumption here, only to show that the data does not undermine them.

To begin, I think that there are two main reasons for resisting the challenge. The first is that even philosophers who rely very heavily on intuitions in support of their theories of responsibility tend to adopt an aim of reflective equilibrium.¹⁶ That is, they aim to find the best balance of intuitions about particular cases and general principles, recognizing that there will be conflicts at times among our intuitions. Reflection plays a key role in the strategy of those appealing to intuitions; it is essential to see which intuitions survive the reflective weighing of plausible principles and particular judgments. The mere fact that some (initial) intuitions do not fit neatly into any extant unifying theory is not sufficient to undermine the Unity assumption. At the same time, if the picture of our intuitions is sufficiently complex and seems to resist any and all unifying principles, then the Unity assumption would indeed be on the ropes. And some of the data described might seem to point in this direction.

A second reason for resisting the challenge is that judgments of responsibility depend on people's concept of responsibility and also on their understanding of the facts in the relevant cases, which can include both case-specific features and also general empirical assumptions. People can make mistakes about all of these things, and subjects and experimenters can differ in unanticipated ways on the facts as presented in various vignettes. Thus, one way of understanding the "fit" between intuitions and unifying principles—and in my view the preferable one—is to think of the principles as fitting a complete story of the intuitions, background empirical assumptions (whether right or wrong), and features of the case as understood by subjects. When "fit" is understood in this way, one can adopt a combination strategy of explaining how the data can fit with a unified theory. For example, in some cases, one can easily "derive" judgments from the theory and obvious facts about the relevant situation, and in others one can expose background empirical assumptions (whether true or false) that can show that judgments are consistent with the theory after all. This allows the unifier more resources with which to address the data.

In what follows, I will offer some suggestions for how to address the data and still preserve a unifying theory. In making these suggestions, I aim to provide both a kind of template for unifiers to use and also some particular hypotheses that can explain specific phenomena. But since the data continues to flood in, I offer the particular hypotheses as hypotheses to be tested, fully prepared that further data will undermine at least some of them. But even if some of them turn out to be false, I hope that they will help make plausible a general strategy for unifiers, and contribute to the ongoing research concerning judgments of responsibility.

One important preliminary point. I will offer hypotheses that support a particular unifying view of responsibility, namely, the reason view described earlier. At the same time, I believe that many of them could just as easily support a different unifying view.

Let me begin with the emotion asymmetry. This one seems tailor made for the reason view, which itself entails a certain moral status asymmetry. According to this view, one must act with the ability to recognize and act on good reasons in order to be responsible. This means that if one actually acts for good reasons, one is responsible, but if one does not, then whether one is responsible depends on whether one could have done otherwise. Thus, the view itself contains an asymmetry: responsibility requires an ability to do otherwise in cases of “bad” actions, but not in cases of “good” action.

Now consider the ways we think about emotions and actions. It is natural to suppose that in cases of “bad” actions done with, say, great anger, the emotional experience itself acts to block one’s ability to see the reasons there are or one’s ability to act on them. We tend to say things like “he couldn’t help it” or “he was overcome” by the emotion. In contrast, a rush of empathy, for example, could itself be seen as a vehicle for reasons-recognition. Rather than getting in the way, it actually shines a light on the reasons there are for acting, or so one might think. Thus, the fact that bad emotions tend to be mitigating factors in people’s judgments of responsibility, while good emotions tend not to be, looks like it could be a natural consequence of an internalized reason view of responsibility and some natural assumptions about the operation of emotions. In fact, this is one place in which the experimental results can be seen to provide previously unrecognized *support* for a particular unifying theory.

Things are not quite so simple in the other moral asymmetry cases, and to handle these, we will need to appeal to some of the other strategies mentioned earlier. For example, consider

the case of judgments about intentions and actions. The experiments show that people tend to attribute greater responsibility for bad intentions relative to bad actions than to good intentions relative to good actions. This is one place in which non-obvious empirical assumptions could explain the different responses in the case. A starting point is the familiar expression: “The road to hell is paved with good intentions.” I gather that the idea behind this expression is that people’s good intentions don’t always result in action; in important situations, people fail to follow through.¹⁷ At the very least, it might be that people assume this to be true, and if so, this could explain the asymmetry in responsibility judgments concerning bad intentions and good intentions.

Alas, initial results obtained by Malle and Bennett suggest that people do *not* make this assumption. Actually, what they seem to show is that “people consider negative intentions significantly *less* likely to be fulfilled than positive intentions...” An assumption of differential predictive diagnosticity for good and bad intentions, then, does not explain the asymmetry in responsibility judgments. However, it does not follow that the familiar expression should be rejected as a source of clues here. It is worth noting that Malle and Bennett go on to mention that “perhaps [people consider negative intentions less likely to be fulfilled] because they believe that the world generally works against the fulfillment of negative intentions.” (11). This suggests that while an assumption of differential predictive diagnosticity is not at work here, a more fundamental assumption is at play instead. Although Malle and Bennett pursue other explanations instead, one natural hypothesis is that we assume that in general the level of *commitment* associated with positive intentions is lower than the level of commitment associated with negative ones. Thus, were the world to cooperate—e.g., were there no police around to prevent the drug sale—perhaps the negative intention would have a higher chance of being

fulfilled than the positive one.¹⁸ Again, this assumption might be false, but if people often believe it (and the familiar expression provides at least some suggestion that it could be true), then that, together with any of a number of unifying theories of responsibility, would fit well with the judgments people actually tend to make. The asymmetry in the judgments in this case might come entirely from the asymmetry in common empirical background assumptions.

The side effect asymmetry can also be understood in a way that is consistent with unifying theories, and with the reason view in particular. The explanation here might be due to asymmetries embedded in morality, and to the conditions for praise and blame (which, it is important to note, are likely to be interestingly different from conditions of responsibility *per se*). In particular, on some plausible theories of our moral duties, one has a (negative) duty to avoid harmful consequences of one's actions. The person who harmed the environment as a foreseen byproduct of his actions violated this duty. On the other hand, the person who helped the environment did not even fulfil (let alone go beyond) a positive duty, because she did not help the environment intentionally. The explanation here is, fundamentally, due to an assumption about what moral duties one has. The unifying theories in question are theories of responsibility, but it is reasonable to suppose that particularly when it comes to blame and praise, judgments will depend at least in part on moral assumptions about what is right and wrong, and what we are obligated to do. In this case, the results point up an interesting and perhaps somewhat neglected question of the relation of moral theory to responsibility. But they are not a reason to reject unifying theories of responsibility, even when combined with other experimental results.

The severity cases are interesting in part because they fall into a category of cases that philosophers have already worried quite a lot about, precisely because people's initial reactions do not seem to fit with more general moral principles. The problem here is the problem of moral

luck. On the one hand, many people accept what might be called the Control Principle, according to which we are morally assessable only to the extent that we are in control of what we are assessed for.¹⁹ On the other hand, we make judgments that seem to violate this principle all the time. Even when the only difference between two situations is outside of the control of the relevant actors, we often provide different judgments of blame in the two situations, for example. The drunk driver who kills a small child receives more blame than an equally drunk driver who by chance passes no one on the way home. Such judgments appear to be reflected in the law, as well. For example, murder typically comes with a higher sentence than attempted murder, and this may reflect a general willingness to give differential judgments of desert and, hence, responsibility. In this case, however, it is interesting to note that many philosophers, upon reflection, revise their immediate responses, or see them as mistaken, or explainable on the basis of empirical assumptions. For example, it is often claimed that the law, and our reactions, are based in part on the general assumption that severity in consequences often reflects a higher level of commitment in intentional cases.²⁰ (The successful murderers are generally more committed than the failed ones.) But when cases are described in such a way as to ensure an equal level of commitment, and to emphasize that the difference in consequences was entirely due to luck, a number of philosophers, as well as the writers of the Model Penal Code, conclude that the two are equally blameworthy, or deserving of punishment.²¹ It is worth emphasizing this point, because if in this sort of case people often revise their judgments on reflection, it is possible that they would do the same in a host of other cases, as well.

Of course, accident cases are different, since they involve no intentions, and hence, no levels of commitment whatsoever. But one suggestion, already noted by Knobe and Doris (forthcoming), is that it would be understandable if people confused degrees of compensation

owed with degrees of responsibility because they frequently go together. This sort of confusion might explain the differential judgments in this case. It is also worth noting that the theories under consideration are often limited to responsibility for actions, and not consequences. Arguably, such theories are incomplete for this reason; but the data do not directly impugn them.

Abstract v. Concrete

The difference in judgments in abstract conditions and judgments in concrete ones is fascinating. If indeed it is the case that people tend to be “compatibilists” when situations are described concretely in terms that evoke high affect, and “incompatibilists” when situations are described in abstract terms, then there is something it is important to explain.

Here is one idea: there is a kind of faulty generalization that people make when they hear about determinism (even when described instead of labeled). With no other information given, people tend to assimilate determinism to coercion; but this suggestion is cancelled when an intentional action is described in concrete terms. It is worth exploring this suggestion further. For example, it would be interesting to ask people open-ended questions of the form: “what would it be like for us in a deterministic world?” If the answers are typically “we’d be like puppets, or “we’d be controlled by fate”, that would be some support for this hypothesis.

It would also be interesting if answers typically included “we’d be like robots” or something similar; this, too, would support the more general hypothesis that some sort of (mistaken) assimilation is going on when people are presented with determinism in the abstract. Eddy Nahmias has tested the hypothesis that when hearing the description of determinism used in experiments, like those of Knobe and Nichols, subjects are assimilating determinism to some sort of reductionism, and it is this that could be problematic, and he has found some support for this conclusion.²² Thus, there are various interesting hypotheses, including combinations of

some of these simple ones, that might explain the results that at first appear to threaten any unified theory. In fact, we should expect the explanation of the results to have more than one component—after all, not everyone answers in the same way.

Additional anecdotal support for the idea that determinism is being assimilated to reductionism is provided by my experience presenting some of the relevant experimental results to an interdisciplinary academic audience that contained a number of psychologists.²³ Several objected to the set-up of the experiments in which subjects are supposed to respond to agents performing deliberate actions in a deterministic scenario. In particular, they worried that the scenarios might already beg a key question in describing actions in such terms at all in a deterministic world, and that the scenarios were actually incoherent because determinism precludes deliberate actions done for reasons. Setting aside for a moment the methodological questions they raised, it is interesting that a significant portion of an audience that is perhaps more self-conscious about isolating background assumptions than most people would be worried that a certain *non-reductive* assumption was illicitly being made. This in turn suggests that they were themselves at least tempted by an assimilation of determinism to reductionism of a kind that bypasses robust or deliberate action done for reasons. I think that this phenomenon provides at least some additional support for the idea that an assumption that determinism entails reductionism is at work for at least some of the subjects presented with the abstract cases, and yet the assumption is cancelled—rightly (as in my view) or wrongly (as in the view of some of the psychologists described)—in at least some concrete cases.²⁴

At the same time, the reaction of these psychologists gives us insight into the difficulty of isolating particular judgments in experimental design. (It should be noted that all of the experimenters note the importance of how determinism is described in the various vignettes, and

have clearly taken care to craft their presentations in ways that beg as few questions as possible.) It also reveals that some of the background assumptions being made are just the kind that philosophers devote their lives to assessing. So while on the one hand the worry expressed by several of the psychologists in the audience raises a question about what conclusions we can draw from the data, it also points at the very same time to a potentially interesting *theoretical* background assumption that might account for certain results. For this reason, we might see that at least some of what underlies the deep disagreement between compatibilists and incompatibilists is a disagreement over this sort of theoretical assumption. This is interestingly different from the concerns over leeway and the ability to be the source of one's actions.²⁵

I have just offered two different explanations of some of the data—one that points to a mistaken assimilation of determinism to coercion and one that points to a mistaken assimilation of determinism to a certain kind of reductionism in the abstract cases. But there are alternative explanations, as well, and there is at least some interesting data that suggests that a very simple assimilation of determinism to reductionism in the abstract case and a canceling of it in the concrete is not the complete explanation of what is going on.²⁶ I do not know what the full answer is, but I think that there is not sufficient reason to conclude that no unifying theory will have resources to explain the data in a unified way.

In sum, the combination strategy used here includes straightforwardly “deriving” typical judgments from a unified theory and plausible empirical assumptions (in the emotion asymmetry case), explaining the oddity of some judgments as resulting from a background empirical assumption (in the intention/action case), explaining still others by background moral assumptions concerning our duties (in the side-effects case), and still others as the result of a possible (and possibly confused) assimilation of independent situational features (in the severity

case and the abstract/concrete case). Unifiers should confront the data, but they should also avail themselves of all the resources available to answer the challenge it poses. Even if unifiers can successfully explain all of the data in ways that are compatible with unified theories, however, the experimental results will have been valuable in forcing us to confront a number of issues, including how we think about determinism, and also issues that sometimes get neglected when the focus *is* on determinism.

When it comes to determinism, the experiments and results so far suggest that it matters greatly in this area how one asks the questions. But with the complexity that this phenomenon creates also comes the possibility for eliciting more fine-grained judgments and at least the prospect of asking questions that isolate a number of importantly different features that are often run together even when conscious attempts are made to separate them. This prospect may prove very hard realize, however, for the reasons described above. It may be, for example, that it is only *after* reflection on a variety of situational features and ways they might be distinguished that we can be sure the features really are separated in subjects' minds, and it may even turn out that whether those features *can* be separated is itself a philosophical question that requires further reflection (e.g., as we saw above, whether determinism undermines agency). But this is not to say that the data so far are unhelpful. To the contrary; at the very least, the results have made it harder to make certain assumptions about people's "natural" theoretical orientation.

The experimental results are helpful in many other ways, as well. Sometimes it seems as though other issues get lost when the focus is determinism, and much of the data described earlier provides an opportunity for various unifying theorists to confront them. To name just a few, we have seen that to give a full explanation of how unifying theories fit the data, we would need to expand on how praise and blame should be treated in a complete theory, how to

understand responsibility for consequences, as opposed to actions, the role emotions play in action, and how theories of responsibility interact with other moral theories and concepts. Although I have argued that the unifiers live to see another day, it also seems clear that the experimental data provides the unifiers with a friendly invitation to continue to extend their theories in all of these directions.

REFERENCES

- Churchland, P. (2002). *Brain-Wise: Studies in Neurophilosophy*. Cambridge: MIT Press.
- Doris, J. and Knobe, J. (forthcoming). "Strawsonian Variations: Folk Morality and the Search for a Unified Theory" in *The Handbook of Moral Psychology* (Oxford: Oxford University Press).
- Fischer, J. and Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. (Cambridge: Cambridge University Press).
- Fischer, J., Kane, R., Pereboom, D., and Vargas, M. (forthcoming). *Four Views of Free Will*. (Blackwell).
- Frankfurt, H. (1971). "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* 68, pp. 5-20.

Frankfurt, H. (1988). *The Importance of What We Care About* (Cambridge: Cambridge University Press).

Kane, R. (1999). "Responsibility, Luck and Chance: Reflections on Free Will and Indeterminism," *Journal of Philosophy* 96, pp. 217-240.

Malle, B. and Bennett, R. (2004). "People's Praise and Blame for Intentions and Actions: Implications of the Folk Concept of Intentionality," Unpublished manuscript. University of Oregon.

Nagel, T. (1979). "Moral Luck," in *Mortal Questions* (Cambridge: Cambridge University Press).

Nahmias, E. (forthcoming). "Folk Fears about Freedom and Responsibility: Determinism vs. Reductionism," in *The Journal of Cognition and Culture* 6, pp. 215-237.

Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. (forthcoming-a) "Surveying Freedom: Folk Intuitions About Free Will and Moral Responsibility," *Philosophical Psychology*.

Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. (forthcoming-b). "Is Incompatibilism Intuitive?" *Philosophy and Phenomenological Research*.

Nelkin, D. (2004). "Moral Luck", *The Stanford Encyclopedia of Philosophy (Spring 2004 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2004/entries/moral-luck/>

Nichols, S. and Knobe, J. (forthcoming). "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions"

Pereboom, D. (2002). *Living Without Free Will* (Cambridge: Cambridge University Press).

Pizzaro, D., Uhlmann, E., and Salovey, P. (2003). "Asymmetry in Judgments of Moral Blame and Praise: The Role of Perceived Metadesires," *Psychological Science* 14, pp. 267-272.

Rawls, J. (1971). *A Theory of Justice*. (Oxford: Oxford University Press).

Richards, N. (1986). "Luck and Desert," *Mind*, 65, pp. 198-209.

Rosebury, B. (1995). "Moral Responsibility and Moral Luck", *Philosophical Review*, 104: 499-524.

Thomson, J.J. (1993). "Morality and Bad Luck", in *Moral Luck*, D. Statman (ed.), Albany: State University of New York Press.

Wallace, J. D. (1994). *Responsibility and the Moral Sentiments*. (Cambridge: Harvard University Press).

Walter, E., editor (2006). *Cambridge Idioms Dictionary, second edition*. (Cambridge: Cambridge University Press).

Watson, G. (1975). "Free Agency," *Journal of Philosophy* 72, pp. 205-220.

Widerker, D. and McKenna, M. (2003). *Moral Responsibility and Alternative Possibilities*. (Ashgate).

Wolf, S. (1990). *Freedom Within Reason*. (Oxford: Oxford University Press).

Woolfolk, R., Doris, J., and Darley, J. (2006). "Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility," *Cognition* 100, pp. 283-301.

Vargas, M. (2005). "The Revisionist's Guide to Responsibility," *Philosophical Studies* 125, pp. 399-429.

Vargas, M. (2006). "Philosophy and the Folk: On Some Implications of Experimental Work for Philosophical Debates on Free Will," *Journal of Cognition and Culture* 6, pp. 239-254.

Viney, W., Waldman, D., and Barchilon, J. (1982). "Attitudes Toward Punishment in Relation to Belief in Free Will and Determinism," *Human Relations* 35, pp. 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*.

¹ An earlier version of this paper was given at the Institute for Decision Sciences at the University of Oregon in January 2007, and I am very grateful to the members of the audience, and especially to Bertram Malle, for their very useful input. The paper evolved from a commentary on Joshua Knobe's presentation, "Variations on a Theme in Moral Cognition," delivered at a conference, *New Perspectives on Free Will and Moral Responsibility*, at the University of San Francisco in November 2006. Many thanks to Manuel Vargas, who organized that conference, and Howard Wettstein, who edited this volume, for giving me the opportunity to think and write about these issues. I also thank Manuel and the other conference participants, Randolph Clarke, John Martin Fischer, Michael McKenna, Eddy Nahmias, Angela Smith, and Daniel Speak for valuable discussion, and Joshua Knobe for his excellent reply. Finally, I am grateful to Sam Rickless for very helpful conversations on these issues.

² See Frankfurt (1969), and Widerker and McKenna (2003) for a sampling of subsequent responses.

³ For example, Derk Pereboom writes that "[b]eginning students typically recoil at the compatibilist response to the problem of moral responsibility" (2001, xvi), and Kane writes that "in my experience, most ordinary persons start out as natural incompatibilists" (1999, 217).

⁴ See Nahmias, Morris, Nadelhoffer, and Turner (forthcoming-a) and (forthcoming-b).

⁵ See especially Vargas (2006), and also Vargas (2005) and Fischer, Kane, Pereboom, and Vargas (forthcoming).

⁶ See Pereboom (2002) who carefully distinguishes these two motivations.

⁷ Frankfurt has offered various versions of this sort of view. See (1971) in which free action is a matter of acting on desires one endorses with second-order desires, and (1988) for other versions of the view. Watson (1975) offers a real self view according to which "the free agent has the capacity to translate his values into action; his actions flow from his evaluational system."

⁸ See Wolf (1990) for this view. Fischer and Ravizza (1998) hold a "moderate reasons-responsiveness" view, that is like this one in requiring a capacity to respond to reasons. However, it differs from Wolf's in interesting ways, including one that will be salient here, namely, that it does not require an ability to otherwise in order to be responsible for a bad action.

⁹ This view shares one aspect with the view of Patricia Churchland, who describes a set of parameters (or parameter space) relevant to being "in-control" (see, for example, Churchland 2002, 212).

¹⁰ There are some views which it might be hard to classify as either unifying or not. For example, it may not be entirely clear how to classify Strawson's theory or theories that take some inspiration from Strawson. For example, R. Jay Wallace's theory (1994), understands being responsible in terms of when it is appropriate to hold someone responsible, which in turn is understood in terms of when it is appropriate to feel the reactive attitudes. If, as Wallace believes, conditions of appropriateness can in turn be specified in terms of unifying conditions, then the theory would seem to be unifying.

¹¹ As mentioned earlier, one could remain a "prescriptive" unifier, in this case. Manuel Vargas (2006) suggests that such a position would be worth pursuing in case the empirical results hold up. In effect, one would adopt a kind of revisionism about our concepts, advocating a change to, say, a compatibilist picture, even though that would require a change in our concept of responsibility and in our self-conception.

¹² See also Viney et al (1982, 1988) and Woolfolk, Darley, and Doris (2006) for similarly suggestive results.

¹³ Pizzaro, Uhlmann, and Salovey (2003).

¹⁴ Malle and Bennett (2004, manuscript).

¹⁵ It is worth noting Knobe and Doris explore other data as well, including some that suggests that subjects' judgments of responsibility vary depending on their relationship to the agent evaluated.

¹⁶ See Rawls (1971).

¹⁷ According to the Cambridge Idioms Dictionary (Walters (2006)), the expression is something that you say which means people often intend to do good things but much of the time, they do not make the effort to do those things. "*I kept meaning to visit her but I didn't get round to it.*"

¹⁸ A very interesting set of experiments being conducted by Luke Misenheimer aims to test a similar hypothesis.

¹⁹ See Moral Luck (Stanford Encyclopedia of Philosophy, 2004)

²⁰ See, for example, Richards (1986), Rosebury (1995), and Thomson (1993).

²¹ Thomson asks, "Well *do* we regard Bert [a negligent driver who causes a death] with an indignation that would be out of place in respect to Carol [an equally negligent driver who does not]? Even after we have been told about how bad luck figured in his history and good luck in hers?" And Thomson answers: "I do not find it in myself to do so" (1993, 205).

²² Nahmias (2006). When scenarios are described as containing deterministic psychological laws, rather than deterministic physical laws, subjects make interestingly different judgments of responsibility, judging agents much more responsible when the laws are psychological than when they are physical.

²³ At a colloquium of the Institute for Decision Sciences at the University of Oregon, January 2007.

²⁴ It might be thought on the basis of these psychologists' reactions that they, at least, are natural incompatibilists. And, indeed, they may join those in the minority who give incompatibilist answers even in the concrete cases. It might be asked, however, whether the assimilation of determinism to reductionism is "natural" or acquired on the basis of reflection (in the same way that compatibilists' views are claimed not to be natural). In any case, it is interesting to ask why they have the view they do, and to identify the theoretical assumptions that distinguish them from other subjects. At the very least, identifying the theoretical assumptions they make *about* determinism allows for the possibility of more dialogue with compatibilists than would otherwise be possible.

²⁵ The idea that the very idea of action itself is what is threatened by determinism (and by mechanism more generally) is not new. See, for example, Nagel (1979): “I believe that in a sense the problem has no solution, because something in the idea of agency is incompatible with actions being events, or people being things. But as the external determinants of what someone has done are gradually exposed, in their effect on consequences, character, and choice itself, it becomes gradually clear that actions are events and people things” (68).

²⁶ Nichols and Knobe (forthcoming) set out the results of a second experiment concerning intuitive judgments in deterministic and indeterministic scenarios. This one aims to isolate the feature of concreteness from the feature of high-affect-inducement. If it turns out that it is high affect rather than concreteness that is at work in the concrete conditions described earlier, then that would provide some reason against the suggestion in the text that concrete descriptions of normal uncoerced and deliberate actions cancels one or two (possibly mistaken) theoretical assumptions. They offer all subjects a description of a completely deterministic world, Universe A, in which all events are caused by preceding events, and in which everything that happens has to happen, given the past. Universe A is contrasted with a second world, Universe B, which is exactly the same except for human decisions. In B, human decisions such as Mary’s deciding to have French Fries for lunch “*did not have to happen*” (11). Now half of the subjects were told to consider a description that takes place in Universe A, and half to consider a description that takes place in Universe B. Each of those groups was further divided, and subjects received one of two different concrete descriptions of an action, followed by a question. The first is supposed to constitute a high-affect condition: “As he has done many times in the past, Bill stalks and rapes a stranger. Is it possible that Bill is fully morally responsible for raping the stranger?” The second is meant to constitute a low-affect condition: “As he has done many times in the past, Mark arranges to cheat on his taxes. Is it possible that Mark is fully morally responsible for cheating on his taxes?” Although subjects in both high and low-affect conditions overwhelmingly answered that the agent in question was responsible in the indeterministic world (95% and 89% respectively), there was a marked difference in judgments of responsibility in the deterministic world (64% and 23% respectively). This suggests that it is not concreteness alone to which people are responding when they judge responsibility in deterministic cases. The results here are fascinating, and pose a serious challenge to the compatibilist interpretations of the data described in the text. But it is a challenge that might be answered. First, one might wonder if what is being isolated is affect, as opposed to, say, seriousness. After all, an even higher percentage of subjects agreed that Jeremy, the bank robber, was responsible in a deterministic situation, and at least at first it would seem that robbing a bank would induce lower affect than stalking and raping. This difference might be due to a difference in the way that determinism is represented in the two experiments. (It would still be puzzling why seriousness should matter to our judgments in this way.) Even if affect is involved, it is possible that the particular description of determinism used in the experiment, together with a certain priming effect of contrasting two universes that might lead subjects to look for some significance in the difference, makes it especially difficult to cancel the suggestion that determinism entails either some sort of coercion or reductionism. In fact, the highlighting of the contrast in the vignette might even *foster* some such suggestion. Thus, only a particularly high affect situation will allow for the cancellation of the suggestion in this case. This is just one possible approach one could take to the data. My guess is that the explanation will turn out not to be a single simple one, even for the subjects who answer in the same way.