Moral Responsibility and Determinism:

The Cognitive Science of Folk Intuitions

Shaun Nichols
Department of Philosophy, University of Utah
Salt Lake City, UT 84112
snichols@philosophy.utah.edu

Joshua Knobe
Department of Philosophy, Princeton University
Princeton, NJ 08544
jknobe@princeton.edu

1. INTRODUCTION

The dispute between compatibilists and incompatibilists must be one of the most persistent and heated deadlocks in Western philosophy. Incompatibilists maintain that people are not fully morally responsible if determinism is true, i.e., if every event is an inevitable consequence of the prior conditions and the natural laws. By contrast, compatibilists maintain that even if determinism is true our moral responsibility is not undermined in the slightest, for determinism and moral responsibility are perfectly consistent.[1]

The debate between these two positions has invoked many different resources, including quantum mechanics, social psychology, and basic metaphysics. But recent discussions have relied heavily on arguments that draw on people's intuitions about particular cases. Some philosophers have claimed that people have incompatibilist intuitions (e.g., Kane 1999, 218; Strawson 1986, 30; Vargas forthcoming); others have challenged this claim and suggested that people's intuitions actually fit with

compatibilism (Nahmias et al. forthcoming). But although philosophers have constructed increasingly sophisticated arguments about the implications of people's intuitions, there has been remarkably little discussion about *why* people have the intuitions they do. That is to say, relatively little has been said about the specific psychological processes that generate or sustain people's intuitions. And yet, it seems clear that questions about the sources of people's intuitions could have a major impact on debates about the compatibility of responsibility and determinism. There is an obvious sense in which it is important to figure out whether people's intuitions are being produced by a process that is generally reliable or whether they are being distorted by a process that generally leads people astray.

Our aim here is to present and defend a hypothesis about the processes that generate people's intuitions concerning moral responsibility. Our hypothesis is that people have an incompatibilist theory of moral responsibility that is elicited in some contexts but that they also have psychological mechanisms that can lead them to arrive at compatibilist judgments in other contexts.[2] To support this hypothesis, we report new experimental data. These data show that people's responses to questions about moral responsibility can vary dramatically depending on the way in which the question is formulated. When asked questions that call for a more abstract, theoretical sort of cognition, people give overwhelmingly incompatibilist answers. But when asked questions that trigger emotions, their answers become far more compatibilist.

2. AFFECT, BLAME, AND THE ATTRIBUTION OF RESPONSIBILITY

In their attempts to get a handle on folk concepts and folk theories, naturalistic philosophers have proceeded by looking at people's intuitions about particular cases (e.g., Knobe 2003a, 2003b; Nahmias et al forthcoming; Nichols 2004a; Weinberg et al. 2001; Woolfolk et al. forthcoming). The basic technique is simple. The philosopher constructs a hypothetical scenario and then asks people whether, for instance, the agent in the scenario is morally responsible. By varying the details of the case and checking to see how people's intuitions are affected, one can gradually get a sense for the contours of the folk theory. This method is a good one, but it must be practiced with care. One cannot simply assume that all of the relevant intuitions are generated by the same underlying folk theory. It is always possible that different intuitions will turn out to have been generated by different psychological processes.

Here we will focus especially on the role of *affect* in generating intuitions about moral responsibility. Our hypothesis is that, when people are confronted with a story about an agent who performs a morally bad behavior, this can trigger an immediate emotional response, and this emotional response can play a crucial role in their intuitions about whether the 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.

Consider, for example, Watson's (1987) interesting discussion of the crimes of Robert Harris. Watson provides long quotations from a newspaper article about how Harris savagely murdered innocent people, showing no remorse for what he had done. Then he describes, in equally chilling detail, the horrible abuse Harris had to endure as he was growing up. After reading all of these vivid details, it would be almost impossible for

a reader to respond by calmly working out the implications of his or her theory of moral responsibility. Any normal reader will have a rich array of reactions, including not only abstract philosophical theorizing but also feelings of horror and disgust. A reader's intuitions about such a case might be swayed by her emotions, leaving her with a conclusion that contravened her more abstract, theoretical beliefs about the nature of moral responsibility.

Still, it might be thought that this sort of effect would be unlikely to influence people's reactions to ordinary philosophical examples. Most philosophical examples are purely hypothetical and thinly described (often only a few sentences in length). To a first glance at least, it might seem that emotional reactions are unlikely to have any impact on people's intuitions about examples like these. But a growing body of experimental evidence indicates that this commonsense view is mistaken. This evidence suggests that affect plays an important role even in people's intuitions about thinly described, purely hypothetical cases (Blair 1995; Greene et al. 2001; Nichols 2002; Haidt et al. 1993).

It may seem puzzling that affect should play such a powerful role, and a number of different models of the role of emotion in evaluative thought have been proposed. We will discuss some of these models in further detail in sections 5, 6, and 7. In the meantime, we want to point to one factor that appears to influence people's affective reactions. A recent study by Smart and Loewenstein (forthcoming) shows that when a transgressor is made more 'determinate' for subjects, subjects experience greater negative affect and are more punitive towards that agent as a result. In the study, subjects play a game in which they can privately cooperate or defect. Each subject is assigned an identifying number, but none of the subjects knows anyone else's number. The

experimenter puts the numbers of the defectors into an envelope. The cooperators are subsequently allowed to decide whether to penalize a defector. The cooperator is informed that he will pick a number out of the envelope to determine which defector will be penalized (or not). The manipulation was unbelievably subtle. In the *indeterminate* condition, subjects decide how much to penalize *before* they draw the number; in the *determinate* condition, subjects decide how much to penalize *after* they draw the number. Despite this tiny difference, Smart and Loewenstein found a significant effect – subjects in the determinate condition gave worse penalties than subjects in the indeterminate condition. Furthermore, subjects filled out a self-report questionnaire on how much anger, blame, and sympathy they felt, and subjects in the determinate condition felt more anger and blame than subjects in the indeterminate condition. Finally, using mediational statistical analysis, Smart and Loewenstein found that determinateness impacts punitiveness by virtue of provoking stronger emotions.

As we shall see, previous studies of people's moral responsibility intuitions all featured determinate agents and therefore were designed in a way that would tend to trigger affective reactions. Our own study provides an opportunity to see how people's intuitions are altered when the stimuli are designed in a way that keeps affective reactions to a minimum.

3. INTUITIONS ABOUT FREE WILL AND RESPONSIBILITY

Incompatibilist philosophers have traditionally claimed both that ordinary people believe that human decisions are not governed by deterministic laws and that ordinary people believe that determinism is incompatible with moral responsibility (e.g., Kane

1999; Strawson 1986). These claims have been based, not on systematic empirical research, but rather on anecdote and informal observation. For example, Kane writes, "In my experience, most ordinary persons start out as natural incompatibilists" (1999, 217). (As will be clear below, we think Kane is actually getting at something deep about our intuitions here.) In recent years, philosophers have sought to put claims like this one to the test using experimental methods. The results have sometimes been surprising.

First, consider the claim that ordinary people believe that human decisions are not governed by deterministic laws. In a set of experiments exploring the lay understanding of choice, both children and adults tended to treat moral choices as indeterminist (Nichols 2004a). Participants were presented with cases of moral choice events (e.g., a girl steals a candy bar) and physical events (e.g., a pot of water comes to a boil), and they were asked whether, if everything in the world was the same right up until the event occurred, the event *had to* occur. Both children and adults were more likely to say that the physical event had to occur than that the moral choice event had to occur. This result seems to vindicate the traditional claim that ordinary people in our culture believe that at least some human decisions are not determined.

Experimental study has not been so kind to the traditional claim that ordinary people are incompatibilists about responsibility. Woolfolk, Doris and Darley (forthcoming) gave participants a story about an agent who is captured by kidnappers and given a powerful 'compliance drug.' The drug makes it impossible for him to disobey orders. The kidnappers order him to perform an immoral action, and he cannot help but obey. Subjects in the 'low identification condition' were told that the agent did not want to perform the immoral action and was only performing it because he had been given the

compliance drug. Subjects in the 'high identification condition' were told that the agent wanted to perform the immoral action all along and felt no reluctance about performing it. The results showed a clear effect of identification: subjects in the high identification condition gave higher ratings of responsibility for the agent than subjects in the low identification condition. This result fits beautifully with the compatibilist view that responsibility depends on identification (e.g. Frankfurt 1988). However, subjects in both conditions showed an overall tendency to give low ratings of responsibility for the agent. So these results don't pose a direct threat to the view that people are incompatibilists about responsibility.

The final set of studies we'll review poses a greater problem for the view that people are intuitive incompatibilists. Nahmias, Morris, Nadelhoffer and Turner (forthcoming) find that participants will hold an agent morally responsible even when they are told to assume that the agent is in a deterministic universe. For instance, they presented participants with the following scenario:

> 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.

Participants were subsequently asked whether Jeremy is morally blameworthy for robbing the bank. The results were striking: 83% of subjects said that Jeremy was morally blameworthy for robbing the bank. In two additional experiments with different scenarios, similar effects emerged, suggesting that lay people regard moral responsibility as compatible with determinism. These findings are fascinating, and we will try to build on them in our own experiments.

Of course, it is possible to challenge the experiments on methodological grounds. For instance, the scenarios use technical vocabulary (e.g., "laws of nature", "current state"), and one might wonder whether the subjects really understood the scenarios. Further, one might complain that determinism is not made sufficiently salient in the scenarios. The story of the supercomputer focuses on the predictability of events in the universe, and many philosophers have taken the predictability of the universe to be less threatening to free will than causal inevitability. Although one might use these methodological worries to dismiss the results, we are not inclined to do so. For we think that Nahmias and colleagues have tapped into something of genuine interest.[3] They report three quite different scenarios that produce much the same effect. In each of their experiments, most people (60-85%) say that the agent is morally responsible even under the assumption that determinism is true. Moreover, the results coincide with independent psychological work on the assignment of punishment. Viney and colleagues found that college students who were identified as determinists were no less punitive than indeterminists (Viney et al. 1982) and no less likely to offer retributivist justifications for

punishments (Viney et al. 1988).[4] So, we will assume that Nahmias et al. are right that when faced with an agent intentionally doing a bad action in a deterministic setting, people tend to hold the agent morally responsible.

But if people so consistently give compatibilist responses on experimental questionnaires, why have some philosophers concluded that ordinary people are incompatibilists?[5] Have these philosophers simply been failing to listen to their own undergraduate students? We suspect that something more complex is going on. On our view, most people (at least in our culture) really do hold incompatibilist theories of moral responsibility, and these theories can easily be brought out in the kinds of philosophical discussions that arise, e.g., in university seminars. It's just that, in addition to these theories of moral responsibility, people also have immediate affective reactions to stories about immoral behaviors. What we see in the results of the experiments by Nahmias and colleagues is, in part, the effect of these affective reactions. To uncover people's underlying theories, we need to offer them questions that call for more abstract, theoretical cognition.

4. EXPERIMENTAL EVIDENCE: FIRST PHASE

We conducted a series of experiments to explore whether participants will be more likely to report incompatibilist intuitions if the emotional and motivational factors are minimized. In each experiment, one condition, the *concrete* condition, was designed to elicit greater affective response; the other condition, the *abstract* condition, was designed to trigger abstract, theoretical cognition. We predicted that people would be more likely to respond as compatibilists in the concrete condition.

Before we present the details of the experiments, we should note that there are many ways to characterize determinism. The most precise characterizations involve technical language about, for example, the laws of nature. However, we think it's a mistake to use technical terminology for these sorts of experiments, and we therefore tried to present the issue in more accessible language.[6] Of course, any attempt to translate complex philosophical issues into simpler terms will raise difficult questions. It is certainly possible that the specific description of determinism used in our study biased people's intuitions in one direction or another. Perhaps the overall rate of incompatibilist responses would have been somewhat higher or lower if we had used a subtly different formulation.

One should keep in mind, however, that our main focus here is on the *difference* between people's responses in the concrete condition and their responses in the abstract condition. Even though we use exactly the same description of determinism in these two conditions, we predict that people will give compatibilist responses in the concrete condition and incompatibilist responses in the abstract condition. Such an effect could not be dismissed as an artifact of our description of determinism. If a difference actually does emerge, we will therefore have good evidence for the view that affect is playing some role in people's compatibilist intuitions.

All of our studies were conducted on undergraduates at the University of Utah,[7] and all of the studies began with the same setup. Participants were given the following description of a determinist universe and an indeterminist universe:

> Imagine a universe (Universe A) in which everything that happens is completely caused by whatever happened before it. This is true from the very beginning of

the universe, so what happened in the beginning of the universe caused what happened next, and so on right up until the present. For example one day John decided to have French Fries at lunch. Like everything else, this decision was completely caused by what happened before it. So, if everything in this universe was exactly the same up until John made his decision, then it *had to happen* that John would decide to have French Fries.

Now imagine a universe (Universe B) in which *almost* everything that happens is completely caused by whatever happened before it. The one exception is human decision making. For example, one day Mary decided to have French Fries at lunch. Since a person's decision in this universe is not completely caused by what happened before it, even if everything in the universe was exactly the same up until Mary made her decision, it *did **not** have to happen* that Mary would decide to have French Fries. She could have decided to have something different.

The key difference, then, is that in Universe A every decision is completely caused by what happened before the decision – given the past, each decision *has to happen* the way that it does. By contrast, in Universe B, decisions are not completely caused by the past, and each human decision *does **not** have to happen* the way that it does.

1. Which of these universes do you think is most like ours? (circle one)

Universe A **Universe B**

Please briefly explain your answer:

The purpose of this initial question was simply to see whether subjects believe that our own universe is deterministic or indeterministic. Across conditions, nearly all participants (over 90%) judged that the indeterministic universe is more similar to our own.

After answering the initial question, subjects received a question designed to test intuitions about compatibilism and incompatibilism. Subjects were randomly assigned either to the *concrete* condition or to the *abstract* condition. We ran several different versions, but we will focus on the most important ones. In one of our concrete conditions, subjects were given the following question:

> In Universe A, a man named Bill has become attracted to his secretary, and he decides that the only way to be with her is to kill his wife and 3 children. He knows that it is impossible to escape from his house in the event of a fire. Before he leaves on a business trip, he sets up a device in his basement that burns down the house and kills his family.
>
> Is Bill fully morally responsible for killing his wife and children?
>
> YES NO

In this condition, most subjects (72%) gave the compatibilist response that the agent was fully morally responsible. This is comparable to results obtained in experiments by Nahmias and colleagues. But now consider one of our abstract conditions:

> In Universe A, is it possible for a person to be fully morally responsible for their actions?
>
> YES NO

In this condition, most subjects (86%) gave the *incompatibilist* response!

In short, most people give the compatibilist response to the concrete case, but the vast majority give the *incompatibilist* response to the abstract case. What on earth could explain this dramatic difference? Let's first consider a deflationary possibility. Perhaps the concrete condition is so long and complex that subjects lose track of the fact that the agent is in a determinist universe. This is a perfectly sensible explanation. To see whether this accounts for the difference, we ran another concrete condition in which the scenario was short and simple. Subjects were given all the same initial descriptions and then given the following question:

> In Universe A, Bill stabs his wife and children to death so that he can be with his secretary. Is it possible that Bill is fully morally responsible for killing his family?
>
> YES NO

Even in this simple scenario, 50% of subjects gave the compatibilist response, which is still significantly different from the very low number of compatibilist responses in the abstract condition.[8]

As we noted above, there are many ways of describing determinism, and the overall rate of incompatibilist responses might have been higher or lower if we had used a somewhat different description. Still, one cannot plausibly dismiss the high rate of incompatibilist responses in the abstract condition as a product of some subtle bias in our description of determinism. After all, the concrete condition used precisely the same description, and yet subjects in that condition were significantly more likely to give compatibilist responses.[9]

These initial experiments replicated the finding (originally due to Nahmias et al.) that people have compatibilist intuitions when presented with vignettes that trigger affective responses. But they also yielded a new and surprising result. When subjects were presented with an abstract vignette, they had predominantly *incompatibilist* intuitions. This pattern of results suggests that affect is playing a key role in generating people's compatibilist intuitions.

5. PSYCHOLOGICAL MODELS

Thus far, we have been providing evidence for the claim that different folk intuitions about responsibility are produced by different kinds of psychological processes. But if it is indeed the case that one sort of process leads to compatibilist intuitions and another leads to incompatibilist intuitions, which sort of process should we regard as the best guide to the true relationship between moral responsibility and determinism?

Before we can address this question, we need to know a little bit more about the specific psychological processes that might underlie different types of folk intuitions. We therefore consider a series of possible models. We begin by looking at three extremely simple models and then go on to consider ways that elements of these simple models might be joined together to form more complex models.

The performance error model

Perhaps the most obvious way of explaining the data reported here would be to suggest that strong affective reactions can bias and distort people's judgments. On this view, people ordinarily make responsibility judgments by relying on a tacit theory, but

when they are faced with a truly egregious violation of moral norms (as in our concrete cases), they experience a strong affective reaction which makes them unable to apply the theory correctly. In short, this hypothesis posits an *affective performance error*. That is, it draws a distinction between people's underlying representations of the criteria for moral responsibility and the performance systems that enable them to apply those criteria to particular cases. It then suggests that people's affective reactions are interfering with the normal operation of the performance systems.

The performance error model draws support from the vast literature in social psychology on the interaction between affect and theoretical cognition. This literature has unearthed numerous ways in which people's affective reactions can interfere with their ability to reason correctly. Under the influence of affective or motivational biases, people are less likely to recall certain kinds of relevant information, less likely to believe unwanted evidence, and less likely to use critical resources to attack conclusions that are motivationally neutral (see Kunda 1990 for a review). Given that we find these biases in so many other aspects of cognition, it is only natural to conclude that they can be found in moral responsibility judgments as well.

More pointedly, there is evidence that affect sometimes biases attributions of responsibility. Lerner and colleagues found that when subjects' negative emotions are aroused, they hold agents more responsible and more deserving of punishment, *even when the negative emotions are aroused by an unrelated event* (Lerner et al. 1998). In their study, subjects in the *anger* condition watched a video clip of a bully beating up a teenager; while subjects in the *emotion-neutral* condition watched a video clip of abstract figures (Lerner et al. 1998, 566). All subjects were then presented with what they were

told was a different experiment designed to examine how people assess responsibility for negligent behaviors. Subjects in the anger condition (i.e., those who had been seen the bully video) gave higher responsibility ratings than subjects in the emotion-neutral condition. So, although the subjects' emotions were induced by the film, these emotions impacted their responsibility judgments in unrelated scenarios. The most natural way to interpret this result is that the emotion served to bias the reasoning people used in making their assessments of responsibility.

Proponents of the performance error model might suggest that a similar phenomenon is at work in the experiments we have reported here. They would concede that people give compatibilist responses under certain circumstances, but they would deny that there is any real sense in which people can be said to hold a compatibilist view of moral responsibility. Instead, they would claim that the compatibilist responses we find in our concrete conditions are to be understood in terms of performance errors brought about by affective reactions. In the abstract condition, people's underlying theory is revealed for what it is ─ incompatibilist.

Affective competence model

There is, however, another possible way of understanding the role of affect in the assessment of moral responsibility. Instead of supposing that affect serves only to bias or distort our theoretical judgments, one might suggest that people's affective reactions actually lie at the core of the process by which they ordinarily assign responsibility. Perhaps people normally make responsibility judgments by experiencing an affective reaction which, in combination with certain other processes, enables an assessment of

moral responsibility. Of course, it can hardly be denied that some people also have elaborate theories of moral responsibility and that they use these theories in certain activities (e.g., in writing philosophy papers), but the proponents of this second view would deny that people's cold cognitive theories of responsibility play any real role in the process by means of which they normally make responsibility judgments. This process, they would claim, is governed primarily by affect.

This 'affective competence' view gains some support from recent studies of people with deficits in emotional processing due to psychological illnesses. When these people are given questions that require moral judgments, they sometimes offer bizarre patterns of responses (Blair 1995; Blair et al. 1997; Hauser et al forthcoming). In other words, when we strip away the capacity for affective reactions, it seems that we are not left with a person who can apply the fundamental criteria of morality in an especially impartial or unbiased fashion. Instead, we seem to be left with someone who has trouble understanding what morality is all about. Results from studies like these have led some researchers to conclude that affect must be playing an important role in the fundamental competence underlying people's moral judgments (Blair 1995; Haidt 2001; Nichols 2004b; Prinz forthcoming).

Proponents of this view might suggest that the only way to really get a handle on people's capacity for moral judgment is to look at their responses in cases that provoke affective reactions. When we examine these cases, people seem to show a marked tendency to offer compatibilist responses, and it might therefore be suggested that the subjects in our studies should be regarded as compatibilists. Of course, we have also provided data indicating that these subjects provide incompatibilist answers when given

theoretical questions, but it might be felt that studying people's theoretical beliefs tells us little or nothing about how they really go about making moral judgments. (Think of what would happen if we tried to study the human capacity for language by asking people theoretical questions about the principles of syntax!) Thus, affective competency theorists might maintain that the best way to describe our findings would be to say that people's fundamental moral competence is a compatibilist one but that some people happen to subscribe to a theory that contradicts this fundamental competence.

Concrete competence model

Finally, we need to consider the possibility that people's responses are not being influenced by affect in any way. Perhaps people's responses in the concrete conditions are actually generated by a purely cognitive process. Even if we assume that the process at work here can only be applied to concrete cases, we should not necessarily conclude that it makes essential use of affect. It might turn out that we have an entirely cognitive, affect-free process that, for whatever reason, can be applied to concrete questions but not to abstract ones.

One particularly appealing version of this hypothesis would be that people's intuitions in the concrete conditions are generated by an innate 'moral responsibility module.'[10] This module could take as input information about an agent and his or her behavior and then produce as output an intuition as to whether or not that agent is morally responsible. Presumably, the module would not use the same kinds of processes that are used in conscious reasoning. Instead, it would use a process that is swift, automatic, and entirely unconscious.

Here, the key idea is that only limited communication is possible between the module and the rest of the mind. The module takes as input certain very specific kinds of information about the agent (the fact that the agent is a human being, the fact that he knows what he is doing, etc.), but the vast majority of the person's beliefs would be entirely inaccessible to processes taking place inside of the module. Thus, the module would not be able to make use of the person's theory about the relationship between determinism and moral responsibility. It might not even be able to make use of the person's belief that the agent is in a deterministic universe. Because these beliefs would be inaccessible inside of the module, the conclusions of the module could differ dramatically from the conclusions that the person would reach after a process of conscious consideration.

Hybrid Models

Thus far, we have been considering three simple models of responsibility attribution. It would be possible, however, to construct more complex models by joining together elements of the three simple ones we have already presented. So, for example, it might turn out that moral responsibility judgments are subserved by a module but that the workings of this module are sometimes plagued with affective performance errors, or that the fundamental competence underlying responsibility judgments makes essential use of affect but that this affect somehow serves as input to a module, and many other possible hybrids might be suggested here.

Since we are unable to consider all of the possible hybrid models, we will focus on one that we find especially plausible. On the hybrid model we will be discussing,

affect plays two distinct roles in the assignment of moral responsibility. Specifically, affect serves *both* as part of the fundamental competence underlying responsibility judgments *and* as a factor that can sometimes lead to performance errors. To get a sense for what we mean here, imagine that you are trying to determine whether certain poems should be regarded as 'moving,' and now suppose you discover that one of the poems was actually written by your best friend. Here, it seems that the basic competence underlying your judgment would involve one sort of affect (your feelings about the poems) but the performance systems enabling your judgment could be derailed by another sort of affect (your feelings about the friend). The hybrid model in question would suggest that a similar sort of process takes place in judgments of moral responsibility. The competence underlying these judgments does make use of affect, but affect can also be implicated in processes that ultimately lead to performance errors.

Proponents of this model might suggest that affect does play an important role in the competence underlying moral responsibility judgments but that the effect obtained in the experiments reported here should still be treated as a performance error.[11] In other words, even if we suppose that affect has an important role to play in moral responsibility judgments, we can still conclude that the basic competence underlying these judgments is an incompatibilist one and that the responses we find in our concrete conditions are the result of a failure to apply that competence correctly.

6. EXPERIMENTAL EVIDENCE: SECOND PHASE

Now that we have described some of the psychological models that might explain our results, we can explore a bit more deeply whether experimental evidence counts

against any of the models. One key question is whether or not the compatibilist responses in our experiments are really the product of affect. We compared concrete conditions with abstract conditions, and we suggested that the concrete descriptions triggered greater affective response, which in turn pushed subjects toward compatibilist responses. However, it's possible that what really mattered was concreteness itself, not any affect associated with concreteness. That is, it's possible that the compatibilist responses were not influenced by affect but were elicited simply because the scenario involved a particular act by a particular individual. Indeed, this is exactly the sort of explanation one would expect from the responsibility module account. Fortunately, there is a direct way to test this proposal.

To explore whether concreteness alone can explain the compatibilist responses, we ran another experiment in which the affective salience varied across the two questions, but concreteness was held constant. Again, all subjects were given the initial descriptions of the two universes, A and B, and all subjects were asked which universe they thought was most similar to ours. Subjects were randomly assigned either to the *high affect* or *low affect* condition. In the *high affect* condition, subjects were asked the following:

> 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?

In the *low affect* condition, subjects were asked:

> 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?

In addition, in each condition, for half of the subjects, the question stipulated that the agent was in Universe A; for the other half the agent was in Universe B. Thus, each subject was randomly assigned to one of the cells in Table 1.

	Agent in indeterminist universe	Agent in determinist universe
High affect case		
Low affect case		

Table 1

What did we find? Even when we used these exclusively concrete scenarios, there was a clear difference between the high affect and low affect cases. Among subjects who were asked about agents in a *determinist* universe, people were much more likely to give the incompatibilist answer in the low affect case than in the high affect case. Indeed, most people said that it is *not* possible that the tax cheat is fully morally responsible, and a clear majority said that it *is* possible that the rapist is fully morally responsible. By contrast, for subjects who were asked about an agent in an indeterminist universe, most people said that it is possible for the agent to be fully morally responsible, regardless of whether he was a tax cheat or a rapist.[12] See Table 2.

	Agent in indeterminist universe	Agent in determinist universe
High affect case	95%	64%
Low affect case	89%	23%

Table 2

These results help to clarify the role that affect plays in people's responsibility attributions Even when we control for concreteness, we still find that affect impacts people's intuitions about responsibility under determinism. The overall pattern of results therefore suggests that affect is playing an important role in the process that generates people's compatibilist intuitions.

We now have good evidence that affect plays a role in compatibilist judgments. But there remains the difficult question of whether what we see in these responses is the result of an affective competence or an affective performance error. Let's consider whether one of these models provides a better explanation of the experiment we just reported.

We think that the affective performance error model provides quite a plausible explanation of our results. What we see in the tax cheat case is that, when affect is minimized, people give dramatically different answers depending on whether the agent is in a determinist or indeterminist universe. On the performance error hypothesis, these responses reveal the genuine competence with responsibility attribution, for in the low affect cases, the affective bias is minimized. When high affect is introduced, as in the serial rapist case, the normal competence with responsibility attribution is skewed by the

emotions; that explains why there is such a large difference between the high and low affect cases in the determinist conditions.

Now let's turn to the affective competence account. It's much less clear that the affective competence theorist has a good explanation of the results. In particular, it seems difficult to see how the affective competence account can explain why responses to the low-affect case drop precipitously in the determinist condition, since this doesn't hold for the high affect case. Perhaps the affective competence theorist could say that low affect cases like the tax cheat case fail to trigger our competence with responsibility attribution, and so we should not treat those responses as reflecting our normal competence. But obviously it would take significant work to show that such everyday cases of apparent responsibility attribution don't really count as cases in which we exercise our competence at responsibility attribution. Thus, at first glance, the performance error account provides a better explanation of these results than the affective competence account.

Of course even if it is true that our results are best explained by the performance error account, this doesn't mean that affect is irrelevant to the normal competence. As noted in the previous section, one option that strikes us as quite plausible is a hybrid account on which (i) our normal competence with responsibility attribution does depend on affective systems, but (ii) affect also generates a bias leading to compatibilist responses in our experiments.

Although our experiment provides some reason to favor the performance error account of the compatibilist responses we found, it seems clear that deciding between the affective performance error and the affective competence models of compatibilist

responses is not the sort of issue that will be resolved by a single crucial experiment. What we really need here is a deeper understanding of the role that affect plays in moral cognition more generally. (Presumably, if we had a deeper understanding of this more general issue, we would be able to do a better job of figuring out how empirical studies could address the specific question about the role of affect in judgments of moral responsibility.) But our inability to resolve all of the relevant questions immediately is no cause for pessimism. On the contrary, we see every reason to be optimistic about the prospects for research in this area. Recent years have seen a surge of interest in the ways in which affect can influence moral cognition – with new empirical studies and theoretical developments coming in all the time – and it seems likely that the next few years will yield important new insights into the question at hand.

7. PHILOSOPHICAL IMPLICATIONS

Our findings help to explain why the debate between compatibilists and incompatibilists is so stubbornly persistent. It seems that certain psychological processes tend to generate compatibilist intuitions, while others tend to generate incompatibilist intuitions. Thus, each of the two major views appeals to an element of our psychological makeup.

But the experimental results do not serve merely to give us insight into the causal origins of certain philosophical positions; they also help us to evaluate some of the arguments that have been put forward in support of those positions. After all, many of these arguments rely on explicit appeals to intuition. If we find that different intuitions are produced by different psychological mechanisms, we might conclude that some of

these intuitions should be given more weight than others. What we need to know now is which intuitions to take seriously and which to dismiss as products of mechanisms that are only leading us astray.

Clearly, the answer will depend partly on which, if any, of the three models described above turns out to be the right one, and since we don't yet have the data we need to decide between these competing models, we will not be able to offer a definite conclusion here. Our approach will therefore be to consider each of the models in turn and ask what implications it would have (if it turned out to be correct) for broader philosophical questions about the role of intuitions in the debate over moral responsibility.

Performance error model

If compatibilist intuitions are explained by the performance error model, then we shouldn't assign much weight to these intuitions. For on that model, as we have described it, compatibilist intuitions are a product of the distorting effects of emotion and motivation. If we could eliminate the performance errors, the compatibilist intuitions should disappear.

Note that the performance error model does not claim that people's compatibilist intuitions are actually *incorrect*. What it says is simply that the process that generates these intuitions involves a certain kind of error. It is certainly possible that, even though the process involves this error, it ends up yielding a correct conclusion. Still, we feel that the performance error model has important philosophical implications. At the very least,

it suggests that the fact that people sometimes have compatibilist intuitions does not itself give us reason to suppose that compatibilism is correct.

The philosophical implications of the performance error model have a special significance because the experimental evidence gathered thus far seems to suggest that the basic idea behind this model is actually true. But the jury is still out. Further research might show that one of the other models is in fact more accurate, and we therefore consider their philosophical implications as well.

Affective competence model

On the affective competence model, people's responses in the concrete conditions of our original experiment are genuine expressions of their underlying competence. The suggestion is that the compatibilist responses people give in these conditions are not clouded by any kind of performance error. Rather, these responses reflect a successful implementation of the system we normally use for making responsibility judgments, and that system should therefore be regarded as a compatibilist one.

In many ways, this affective competence model is reminiscent of the view that P.F. Strawson (1962) puts forward in his classic paper 'Freedom and Resentment.' On that view, it would be a mistake to go about trying to understand the concept of moral responsibility by seeking to associate it with some sort of metaphysical theory. Rather, the best place to start is with an examination of the 'reactive attitudes' (blame, remorse, gratitude, etc.) and the role they play in our ordinary practice of responsibility attribution.

Yet, despite the obvious affinities between the affective competence model and Strawson's theory, it is important to keep in mind certain respects in which the affective

competence model is making substantially weaker claims. Most importantly, the model isn't specifically claiming that people proceed *correctly* in the concrete conditions. All it says is that people's responses in these conditions reflect a successful implementation of their own underlying system for making responsibility judgments. This claim then leaves it entirely open whether the criteria used in that underlying system are themselves correct or incorrect.

For an analogous case, consider the ways in which people ordinarily make probability judgments. It can be shown that people's probability judgments often involve incorrect inferences, and one might therefore be tempted to assume that people are not correctly applying their own underlying criteria for probabilistic inference. But many psychologists reject this view. They suggest that people actually are correctly applying their underlying criteria and that the mistaken probabilistic inferences only arise because people's underlying criteria are themselves faulty (see, e.g., Tversky and Kahneman 1981; 1983).

Clearly, a similar approach could be applied in the case of responsibility judgments. Even if people's compatibilist intuitions reflect a successful implementation of their underlying system for making responsibility judgments, one could still argue that this underlying system is itself flawed. Hence, the affective competence model would vindicate the idea that people's core views about responsibility are compatibilist, but it would be a mistake to regard the model as an outright vindication of those intuitions.

Concrete competence model

The implications of the concrete competence model depend in a crucial way on the precise details of the competence involved. Since it is not possible to say anything very general about all of the models in this basic category, we will focus specifically on the implications of the claim that people's responsibility attributions are subserved by an encapsulated module.

As a number of authors have noted, modularity involves a kind of trade-off. The key advantages of modules are that they usually operate automatically, unconsciously, and extremely quickly. But these advantages come at a price. The reason why modules are able to operate so quickly is that they simply ignore certain sources of potentially relevant information. Even when we know that the lines in the Müller-Lyer illusion are the same length, we still have the visual illusion. Perhaps in the assignment of moral responsibility, we are dealing with a similar sort of phenomenon — a 'moral illusion.' It might be that people have a complex and sophisticated theory about the relationship between determinism and moral responsibility but that the relevant module just isn't able to access this theory. It continues to spit out judgments that the agent is blameworthy even when these judgments go against a consciously held theory elsewhere in the mind.

Of course, defenders of compatibilism might point out that this argument can also be applied in the opposite direction. They might suggest that the module itself contains a complex and sophisticated theory to which the rest of the mind has no access. The conclusion would be that, unless we use the module to assess the relationship between determinism and moral responsibility, we will arrive at an impoverished and inadequate understanding. This type of argument definitely seems plausible in certain domains (e.g.,

in the domain of grammatical theory). It is unclear at this point whether something analogous holds true for the domain of responsibility attribution.[13]

Reflective equilibrium

Our concern in this section has been with philosophical questions about whether knowledge of particular mental processes are likely to give us valuable insight into complex moral issues. Clearly, these philosophical questions should be carefully distinguished from the purely psychological question as to whether people *think* that particular mental processes give them insight into these issues. Even if people think that a given process is affording them valuable moral insight, it might turn out that this process is actually entirely unreliable and they would be better off approaching these issues in a radically different way.

Still, we thought it would be interesting to know how people themselves resolve the tension between their rival intuitions, and we therefore ran one final experiment. All subjects were given a brief description of the results from our earlier studies and then asked to adjudicate the conflict between the compatibilist and incompatibilist intuitions. Given that people's intuitions in the concrete conditions contradict their intuitions in the abstract conditions, would they choose to hold on to the concrete judgment that Bill is morally responsible or the abstract judgment that no one can be responsible in a deterministic universe?[14] The results showed no clear majority on either side. Approximately half of the subjects chose to hold onto the judgment that the particular agent was morally responsible, while the other half chose to hold onto the judgment that

no one can be responsible in a deterministic universe.[15] Apparently, there is no more consensus about these issues among the folk than there is among philosophers.

9. CONCLUSION

As we noted at the outset, participants in the debate over moral responsibility have appealed to an enormous variety of arguments. Theories from metaphysics, moral philosophy, philosophy of mind and even quantum mechanics have all been shown to be relevant in one way or another, and researchers are continually finding new ways in which seemingly unrelated considerations can be brought to bear on the issue. The present paper has not been concerned with the full scope of this debate. Instead, we have confined ourselves to just one type of evidence – evidence derived from people's intuitions.

Philosophers who have discussed lay intuitions in this area tend to say either that folk intuitions conform to compatibilism or that they conform to incompatibilism. Our actual findings were considerably more complex and perhaps more interesting. It appears that people have *both* compatibilist *and* incompatibilist intuitions. Moreover, it appears that these different kinds of intuitions are generated by different kinds of psychological processes. To assess the importance of this finding for the debate over moral responsibility, one would have to know precisely what sort of psychological process produced each type of intuition and how much weight to accord to the output of each sort of process. We have begun the task of addressing these issues here, but clearly far more remains to be done.

Acknowledgments

Several people gave us great feedback on an early draft of this paper. We'd like to thank Chris Hitchcock, Bob Kane, Neil Levy, Al Mele, Stephen Morris, Thomas Nadelhoffer, Eddy Nahmias, Derk Pereboom, Lynne Rudder-Baker, Tamler Sommers, Jason Turner, and Manuel Vargas. Thanks also to John Fischer for posting a draft of this paper on the Garden of Forking Paths weblog (http://gfp.typepad.com/). Versions of this paper were delivered at the UNC/Duke workshop on Naturalized Ethics, the Society for Empirical Ethics, the Society for Philosophy and Psychology, Yale University, the University of Arizona, and the Inland Northwest Philosophy Conference. We thank the participants for their helpful comments.

Notes:

[1] Actually, compatibilists and incompatibilists argue both (1) about whether determinism is compatible with moral responsibility and (2) about whether determinism is compatible with *free will*. As Fischer (1999) has emphasized, these two questions are logically independent. One might maintain that determinism is compatible with moral responsibility but not with free will. Here, however, our concern lies entirely with the first of the two questions — whether determinism is compatible with moral responsibility.

[2] We use the term 'theory' here loosely to refer to an internally represented body of information. Also, when we claim that the folk have an incompatibilist theory, we are not suggesting that this theory has a privileged status over the psychological systems that

generate compatibilist intuitions. As will be apparent, we think that it remains an open question whether the system that generates incompatibilist intuitions has a privileged status.

[3] One virtue of Nahmias and colleagues' question about moral responsibility is that the notion of 'moral responsibility' is supposed to be common between philosophers and the folk. That is, philosophers tend to assume that the notion of moral responsibility deployed in philosophy closely tracks the notion that people express when they attribute moral responsibility. Furthermore, incompatibilists often specify that the relevant incompatibilist notion of free will is precisely the notion of free will that is required for moral responsibility (e.g., Campbell 1951). Nahmias and colleagues also ask questions about whether the agent in the deterministic scenario "acts of his own free will," and they find that people give answers consonant with compatibilism. We find these results less compelling. For the expression 'free will' has become a term of philosophical art, and it's unclear how to interpret lay responses concerning such technical terms. Moreover, incompatibilists typically grant that there are compatibilist notions of freedom that get exploited by the folk. Incompatibilists just maintain that there is also a commonsense notion of freedom that is not compatibilist.

[4] Although these results from Viney and colleagues are suggestive, the measure used for identifying determinists is too liberal, and as a result, the group of subjects coded as 'determinists' might well include indeterminists. (See McIntyre et al. 1984 for a detailed description of the measure.) It remains to be seen whether this result will hold up using better measures for identifying determinists.

[5] A related problem for the incompatibilist concerns the history of philosophy – if incompatibilism is intuitive, why has compatibilism been so popular among the great philosophers in history? An incompatibilist-friendly explanation is given in Nichols (forthcoming).

[6] In our deterministic scenario, we say that given the past, each decision *has to happen* the way that it does. This scenario allows us to test folk intuitions about the type of compatibilism most popular in contemporary philosophy. Most contemporary compatibilists argue, following Frankfurt (1969), that an agent can be morally responsible for her behavior even if she *had to* act the way she did. (As we shall see, most subjects in our concrete condition give responses that conform to this view.) However, it would also be possible for a compatibilist to maintain that (1) we can never be responsible for an event that had to occur the way it did but also that (2) even if a particular behavior is determined to occur by the laws of nature, the agent does not necessarily *have to* perform that behavior. Our experiment does not address the possibility that the folk subscribe to this type of compatibilism. With any luck, that possibility will be investigated in future research.

[7] It will, of course, be important to investigate whether our results extend to other populations. However, as we will stress throughout, we are primarily looking at how subjects from the same population give different answers in the different conditions.

[8] χ^2 (1, N=41) = 6.034, p< .05, two-tailed.

[9] We also ran an experiment that used a more real-world kind of case than the deterministic set up described in our main experiments. This was sparked by some perceptive comments from Daniel Batson, who also gave us extremely helpful

suggestions in designing the study. Again, the idea was to test whether abstract conditions were more likely to generate incompatibilist responses than affect-laden concrete conditions. All subjects were told about a genetic condition that leads a person to perform horrible actions, but they were also told that there is now an inexpensive pill that counteracts the condition and that now everyone with the condition gets this pill. In the abstract condition, subjects were then asked to indicate whether the people who had this condition before the pill was created could be held morally responsible for their actions. In the concrete condition, subjects were told that Bill had this condition before the pill was invented, and Bill killed his wife and children to be with his secretary. Subjects were then asked to indicate whether Bill was morally responsible for his action. The results were quite clear, and they were in concert with all of our earlier findings. Subjects given the abstract question gave significantly lower ratings of responsibility than subjects given the concrete question. Thus, the basic effect can be obtained using quite different materials.

[10] As far as we know, no prior research has posited a moral responsibility module, but there has been considerable enthusiasm for the more general idea that many basic cognitive capacities are driven by modules (Fodor 1983; Leslie 1994), and a number of authors have suggested that certain aspects of moral judgment might be subserved by module-like mechanisms (Dwyer 1999; Harman 1999; Hauser forthcoming).

[11] We are grateful to Jesse Prinz for suggesting this possibility.

[12] As in our previous experiments, the vast majority of subjects said that our universe was most similar to the indeterminist universe. We suspect that being a determinist might actually lead people to have more compatibilist views (see Nichols 2006), and as a result,

we antecedently decided to exclude the minority who gave the determinist response from our statistical analyses. The statistical details are as follows. The contrast between high and low affect for the determinist condition was significant ($\chi 2$ (1, N=44) = 8.066, p<.01). That is, people were more likely to say that it's possible for the rapist to be fully morally responsible. The contrast between the two high affect conditions was also significant ($\chi 2$ (1, N=45) = 7.204, p<.01); that is, people were more likely to say that it's possible that the rapist is fully morally responsible in the indeterminist universe. The contrast between the two low affect conditions was very highly significant ($\chi 2$ (1, N=45) = 26.492, p<0.0001). Subjects were dramatically more likely to say that it's possible for the tax cheat to be fully morally responsible in the indeterminist universe.

[13] The distinction between modularity hypotheses and affective hypotheses first entered the philosophical literature in the context of the debate about the role of moral considerations in intentional action (Knobe forthcoming, Malle and Nelson 2003, Nadelhoffer forthcoming; Young et al. forthcoming). In that context, modularity hypotheses are usually regarded as vindicating folk intuitions. However, there is a key difference between that context and the present one. The difference is that information about the moral status of the action might be accessible in an intentional action module, but information about determinism is unlikely to be accessible in a moral responsibility module.

[14] The design of the pilot study was modeled on the initial experiments described in section 3. Participants were asked both the high affect (Bill stabbing his wife) and the abstract questions (counterbalanced for order). They then answered the reflective equilibrium question:

Previous research indicates that when people are given question 3 above, they often say that Bill is fully morally responsible for killing his family. But when people are given question 2 above, most people say that it is not possible that people in Universe A are fully morally responsible for their actions. Clearly these claims are not consistent. Because if it is not possible to be fully morally responsible in Universe A, then Bill can't be fully morally responsible.

We are interested in how people will resolve this inconsistency. So, regardless of how you answered questions 2 and 3, please indicate which of the following you agree with most:

i. In Universe A, it is *not* possible for people to be morally responsible for their actions.

ii. Bill, who is in universe A, *is* fully morally responsible for killing his family.

[15] There were 19 subjects. Of these, 10 gave incompatibilist response to the reflective equilibrium question; 9 gave compatibilist responses.

References

Blair, R. 1995. "A Cognitive Developmental Approach to Morality: Investigating the Psychopath." *Cognition* 57.

Blair, R., Jones, L., Clark, F., Smith, M., and Jones, L. 1997. "The Psychopathic Individual: A Lack of Responsiveness to Distress Cues?" *Psychophysiology* 34.

Campbell, C. A. 1951. "Is 'Free Will' a Pseudo-problem?" *Journal of Philosophy* 60.

Dwyer, S. 1999. "Moral Competence." In K. Murasugi and R. Stainton (eds.), *Philosophy and Linguistics*. Westview Press.

Fischer, J. 1999. "Recent Work on Moral Responsibility." *Ethics* 110.

Fodor, J. 1983. *Modularity of Mind*. MIT Press.

Frankfurt, H. 1969. "Alternate Possibilities and Moral Responsibility." *Journal of* Philosophy, 66.

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

Greene, J., Sommerville, R., Nystrom, L., Darley, J., and Cohen, J. 2001. "An fMRI Investigation of Emotional Engagement in Moral Judgment," *Science* 293.

Haidt, J. 2001. "The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108.

Haidt, J., Koller, S.H., Dias, M.G. 1993. "Affect, Culture, and Morality, or Is it Wrong to Eat Your Dog?" *Journal of Personality and Social Psychology* 65.

Harman, G. 1999. "Moral Philosophy and Linguistics." In K. Brinkmann (ed.), *Proceedings of the 20th World Congress of Philosophy: Volume 1: Ethics*. Philosophy Documentation Center.

Hauser, M. forthcoming. *Moral Minds: The Unconscious Voice of Right and Wrong.* NY: Harper Collins.

Hauser, M., Young, L., and Cushman, F. forthcoming. "Reviving Rawls' Linguistic Analogy: Operative Principles and the Causal Structure of Moral Actions." In W. Sinnott-Armstrong (ed.) *Moral Psychology*.

Hume, D. 1740/1978. *A Treatise of Human Nature*. Oxford: Oxford University Press.

Kane, R. 1999. "Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism." *Journal of Philosophy* 96.

Knobe, J. 2003a. "Intentional Action and Side-Effects in Ordinary Language." *Analysis* 63.

Knobe, J. 2003b. "Intentional Action in Folk Psychology: An Experimental Investigation." *Philosophical Psychology* 16.

Knobe, J. forthcoming. "The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology," *Philosophical Studies*.

Kunda, Z. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin* 108.

Lerner, J., Goldberg, J., and Tetlock, P. 1998. "Sober Second Thought: The Effects of Accountability, Anger, and Authoritarianism on Attributions of Responsibility," *Personality and Social Psychology Bulletin* 24.

Leslie, A. 1994. "ToMM, ToBY and Agency: Core Architecture and Domain Specificity." In L. Hirschfeld and S. Gelman (eds.) *Mapping the mind*. Cambridge: Cambridge University Press.

Malle, B. and Nelson, S. 2003. "Judging Mens Rea: The Tension Between Folk Concepts and Legal Concepts of Intentionality." *Behavioral Sciences and the Law* 21.

McIntyre, R., Viney, D., and Viney, W. 1984. "Validity of a scale designed to measure beliefs in free will and determinism" *Psychological Reports*, 54.

Nadelhoffer, T. forthcoming. "Praise, Side Effects, and Folk Ascriptions of Intentional Action," *The Journal of Theoretical and Philosophical Psychology.*

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

Nichols, S. 2002. "Norms with Feeling," *Cognition,* 84.

Nichols, S. 2004a. "The Folk Psychology of Free Will: Fits and Starts." *Mind & Language*, 19.

Nichols, S. 2004b. *Sentimental Rules: On the Natural Foundations of Moral Judgment.* Oxford University Press.

Nichols, S. 2006. "Folk Intuitions about Free Will." *Journal of Cognition and Culture.*

Nichols, S. forthcoming. "The Rise of Compatibilism: A Case Study in the Quantitative History of Philosophy."

Prinz, J. forthcoming. *The Emotional Construction of Morals.* Oxford: Oxford University Press.

Pylyshyn, Z. 1999. “Is Vision Continuous with Cognition? The Case for Cognitive Impenetrability of Visual Perception.” *Behavioral and Brain Sciences* 22.

Smart, D., and Loewenstein, G. forthcoming. “The Devil You Know: The Effects of Identifiability on Punitiveness.”

Strawson, G. 1986. *Freedom and Belief*. Oxford: Oxford University Press.

Strawson, P. 1962. “Freedom and Resentment.” *Proceedings of the British Academy* 48. Reprinted in G. Watson (ed.) *Free Will,* Oxford: Oxford University Press, 1980. Page references are to the reprinted version.

Tversky, A., and Kahneman, D. 1981. “The Framing of Decisions and the Psychology of Choice.” *Science* 211.

Tversky, A., and Kahneman, D. 1983. “Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probablistic Reasoning.” *Psychological Review* 90.

Van Inwagen, P. 1983. *An Essay on Free Will*. Oxford: Oxford University Press.

Vargas, M. forthcoming. “On the Importance of History for Responsible Agency.” *Philosophical Studies.*

Viney, W., Waldman, D., and Barchilon, J. 1982. “Attitudes toward Punishment in Relation to Beliefs in Free Will and Determinism” *Human Relations* 35.

Viney, W., Parker-Martin, P., and Dotten, S. D.H. 1988. “Beliefs in Free Will and Determinism and Lack of Relation to Punishment Rationale and Magnitude.” *Journal of General Psychology* 115.

Watson, G. 1987. "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme." in *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*, ed. F. Schoeman, Cambridge University Press.

Weinberg, J., Nichols, S., and Stich, S. 2001. "Normativity and Epistemic Intuitions." *Philosophical Topics* 29.

Woolfolk, R., Doris, J., and Darley, J. forthcoming. "Attribution and Alternate Possibilities: Identification and Situational Constraint as Factors in Moral Cognition." *Cognition*.

Young, L., Cushman, F., Adolphs, R., Tranel, D., and Hauser, M. forthcoming. "Does Emotion Mediate the Effect of an Action's Moral Status on its Intentional Status? Neuropsychological Evidence." *Journal of Cognition and Culture.*