EXPERIMENTAL PHILOSOPHY ON FREE WILL: AN ERROR THEORY FOR INCOMPATIBILIST INTUITIONS

for *New Waves in Philosophy of Action*

Eddy Nahmias & Dylan Murray
Department of Philosophy, Neuroscience Institute, Georgia State University

Abstract: We discuss recent work in experimental philosophy on free will and moral responsibility and then present a new study. Our results suggest an error theory for incompatibilist intuitions. Most laypersons who take determinism to preclude free will and moral responsibility apparently do so because they mistakenly interpret determinism to involve fatalism or "bypassing" of agents' relevant mental states. People who do not misunderstand determinism in this way tend to see it as compatible with free will and responsibility. We discuss why these results pose a challenge to incompatibilists.

I. Introduction

It's called "the problem of free will and determinism," but much depends on what determinism is taken to mean and entail. *Incompatibilists* claim that it is impossible for people to have free will and moral responsibility if determinism is true, and they often suggest that this is the natural position to take, supported by our pre-theoretical intuitions. Robert Kane, for instance, states that "ordinary persons start out as natural incompatibilists" (1999, 217), and Galen Strawson claims that "it is in our nature to take determinism to pose a serious problem for our notions of responsibility and freedom" (1986, 89). Sometimes people take "determinism" to *mean* "the opposite of free will," in which case incompatibilism is indeed intuitive, but at the cost of being an empty tautology. In philosophical debates, *determinism* has a technical meaning: a complete description of the state of the universe at one time and of the laws of nature logically entails a complete description of the state of the universe at any later time.[1] However, it is not obvious why determinism, defined in this way, is supposed to be incompatible with free will; rather, a further explanation of just why determinism precludes some ability associated with free will seems required. The explanations generally offered by incompatibilists are that determinism precludes either (i) the ability to choose among *alternative possibilities* for action, while holding fixed the actual past and the laws of nature (AP), or (ii) the ability to be the *ultimate source* of one's actions, such that one is ultimately responsible for some aspect of the conditions that led up to one's actions (US). To say that incompatibilism is intuitive, then, is presumably to claim that it is natural to find one or both of these conditions necessary for free will and to understand the condition in such a way that determinism precludes it.

Compatibilists, who believe that determinism does *not* preclude free will and moral responsibility, often develop arguments to show why AP and US, as defined by incompatibilists, are *not* in fact required for free will and responsibility, sometimes offering analyses of abilities meant to capture the attractive features of AP and US but in ways consistent with determinism. They often argue that certain premises and principles used in

[1] This definition is drawn from van Inwagen (1983). Two less technical, though not quite equivalent, ways of stating determinism are: 1. In a deterministic universe, necessarily, re-creating identical initial conditions and laws of nature produces identical later events; 2. Determinism is the view that every event is *completely caused* by earlier events, such that, necessarily, *given* the earlier events and the laws of nature, the later events occur. These two descriptions are more similar to the ones used in the studies described below.

incompatibilist arguments are mistaken. Compatibilists have also attempted to *explain away* the intuitions incompatibilists appeal to, sometimes suggesting that these intuitions are based on mistaken interpretations of the implications of determinism. For example, determinism may *appear* to threaten free will because it is conflated with types of coercion or manipulation, which, the compatibilist argues, are importantly different from determinism.

These conflicting views about what determinism entails and which abilities are required for free will typically lead to stalemates, often bottoming out in disagreements about which view best captures our ordinary intuitions and conceptual usage. For instance, incompatibilists claim that it is widely accepted that free will requires the ability to do otherwise. Compatibilists respond that it is not obvious that this ability must be "unconditional" (as suggested by AP); rather, free action requires a "conditional" ability to do otherwise *if* relevant earlier conditions had been different, an ability that is consistent with determinism. Given such stalemates, it would help to gain a better understanding of people's pre-philosophical intuitions about free will, moral responsibility, and determinism, as well as the sources of these intuitions. This could help to elucidate which position in fact accords best with ordinary thinking about these issues, or whether some of the intuitions supporting one position are produced in systematically unreliable ways. Though such information certainly won't *resolve* the debate, it can suggest that one side needs to answer certain questions, to motivate its views in new ways, or to take on the argumentative burden of proof.

One might attempt to uncover such information about folk intuitions and their underlying psychological processes through armchair analysis, but empirical methods will often be required to supplement such analysis, especially when philosophers on opposing sides offer conflicting claims about what is intuitive. The recent movement of "experimental philosophy" does just this, drawing on the empirical methods of psychology to systematically examine people's intuitions about philosophical issues, and then carefully considering whether and how these results impact the philosophical debates. Below, we offer a brief history of experimental philosophy on free will before presenting some of our recent studies. But first we jump ahead to the conclusion we take our results to support.

Our goal is to develop an "error theory" for incompatibilist intuitions—to show that, when people take determinism to preclude free will and moral responsibility, they usually do so because they *misinterpret* what determinism involves. In other words, we aim to explain why people *appear* to have incompatibilist intuitions, when in fact they do not. Whereas incompatibilists have suggested that "ordinary persons have to be talked out of [their] natural incompatibilism by the clever arguments of philosophers" (Kane, 1999, 217) and that "beginning students typically recoil at the compatibilist response to the problem of moral responsibility" (Pereboom, 2001, xvi), we believe that ordinary persons typically need help seeing the allure of incompatibilism. As suggested above, the proper conception of determinism needs to be given to them—without being presented in a misleading way—and then some explanation needs to be given for why determinism, so defined, is incompatible with free will and moral responsibility, perhaps by motivating the idea that they require AP and US. We suggest that in this process many ordinary persons (e.g., beginning students) come to interpret determinism to entail threats to free will that it does *not* in fact entail. We predict that laypersons often mistakenly take determinism to mean that everything that happens is inevitable—it will happen *no matter what*—or that agents' decisions, desires, or beliefs make no difference to what they end up doing, and that such mistakes then generate people's intuitions about agents' lacking free will and moral responsibility. Indeed, people may take determinism to preclude the sorts of abilities *compatibilists* associate with free will, such as the abilities to consciously deliberate about what to do and to control one's behavior in light of one's reasons. But if people's purportedly *incompatibilist* intuitions result

primarily from mistakenly interpreting determinism to preclude what *compatibilists* require for free will, then these intuitions do not support incompatibilism.

Suppose laypersons are presented with scenarios that describe a deterministic universe, and suppose that some respond that agents in that universe do *not* have free will (FW) and are *not* morally responsible (MR) for their actions—they express "incompatibilist intuitions"—while others respond that agents in these deterministic universes can have FW and MR—they express "compatibilist intuitions." One explanation for such mixed results (see below for examples) is that different people simply have different intuitions about the relationship between determinism and FW or MR, perhaps because they have different conceptions of "free will" or attribute moral responsibility in varying ways (see Knobe & Doris, forthcoming). We think that this interpretation may explain *some* of the variations in people's intuitions and may even help to explain the intractability of the philosophical debates. It may also be that some people who express compatibilist intuitions do not understand the deterministic nature of the scenario or are not drawing the intuitive connections between it and factors like AP and US. Perhaps people fail to draw these purported implications of determinism due to an emotional bias (Nichols & Knobe, 2007; see below). This would suggest an error theory for compatibilist intuitions—that is, it would suggest that these people have only *apparent*, but not *genuine* compatibilist intuitions.

However, the conflicting results might also be explained with an error theory for the *incompatibilist* intuitions people seem to have. Our hypothesis is that many people who appear to have incompatibilist intuitions are interpreting determinism to entail what we will call "bypassing," and they take *bypassing* to preclude FW and MR. While bypassing does preclude FW and MR, it is a *mistake* to interpret determinism to entail bypassing. So, if the reason people express incompatibilist intuitions is that they mistakenly take determinism to entail bypassing, then those intuitions do not in fact support the conclusion that determinism, properly understood, is incompatible with free will.

What is "bypassing"? The basic idea is that one's actions are caused by forces that bypass one's conscious self, or at least what one identifies as one's "self". More specifically, it is the thesis that one's actions are produced in a way that bypasses the abilities compatibilists typically identify with free will, such as rational deliberation, conscious consideration of beliefs and desires, formation of higher-order volitions, planning, and the like.[2] As such, bypassing might take the form of *epiphenomenalism* about the relevant mental states (i.e., that deliberations, beliefs, and desires are causally irrelevant to action), or it might take the form of *fatalism*—the belief that certain things will happen no matter what one decides or tries to do, or that one's actions *have to* happen *even if* the past had been different. Bypassing suggests that conscious agents have no control over their actions because they play no role in the causal chain that leads to their actions. For our study discussed below, we "operationalized" bypassing in a more precise way.

The crucial point is that determinism, as defined by philosophers debating free will, simply does *not* entail bypassing (certainly not in the way we operationalize it below). The history of compatibilism might be caricatured as an attempt to drive home this point. Compatibilists have emphasized that determinism does not mean or entail that all events are inevitable, in the sense that they will happen no matter what we decide or try to do. They point out that determinism does not render our beliefs, desires, deliberations, or decisions

[2] For compatibilist accounts of these abilities, see, e.g., Fischer and Ravizza (1998), Frankfurt (1971), Watson (1976), and Wolf (1990). Even incompatibilists generally take these sorts of capacities to be *necessary* for free and responsible agency—see, e.g., O'Connor (2005). The language of "bypassing" is used in Mele (1995) and Blumenfeld (1988).

causally impotent. Quite the contrary. So long as our mental states are part of the deterministic sequence of events, they play a crucial role in determining what will happen. Of course, incompatibilists generally agree with all this, but claim their arguments are not based on such mistakes. Nonetheless, the pre-philosophical *intuitive* appeal of incompatibilism may rest largely on such mistakes, and to the extent that it does, incompatibilists either need to abandon the appeal to wide-scale intuitive support as a motivation or basis for their position, or they need to demonstrate that incompatibilism remains intuitive *even when* people properly recognize that determinism does not entail bypassing. Put another way, since incompatibilists generally allow that determinism is indeed compatible with the abilities compatibilists associate with free will, "*genuine* incompatibilist intuitions" are those that do not involve misinterpreting determinism to involve bypassing of these compatibilist abilities (see Figure 1). If it can be shown, then, that most people who take determinism to preclude FW and MR do so on the basis of such a mistake, this should at least shift the burden onto the incompatibilist to demonstrate that people nonetheless have *genuine* incompatibilist intuitions. Our goal here is to offer evidence that most laypersons who respond that agents do *not* have FW and MR in deterministic universes are in fact only expressing "*apparent* incompatibilist intuitions" because they misunderstand determinism to involve bypassing (see Figure 2).

[INSERT FIGURES 1 & 2 NEAR HERE]

II. Experimental Philosophy on Free Will

As we have seen, philosophers often appeal to ordinary intuitions and common sense about free will and moral responsibility. We think such appeals have a legitimate place in the philosophical debate. This debate, unlike others about more technical concepts, is about concepts that are intimately connected to ordinary people's beliefs about and practices concerning morality, agency, praise, blame, punishment, reward, etc. The claim is not that ordinary intuitions or conceptual usage should *exhaust* the philosophical analysis of free will, much less that they will inform us about any extra-semantic facts about the nature of human decision-making. Rather, the claim is that folk intuitions provide important information about *which* extra-semantic facts we should be looking for when we want to know, for instance, whether humans have free will and are morally responsible for their actions. Philosophical theories should systematize such intuitions as much as possible, revise them when they are inconsistent or when competing theoretical advantages (such as consistency with scientific facts) call for it, or explain the intuitions away—that is, offer an error theory for why they *appear* to support a particular position when in fact they do not. If, instead, philosophers end up mired in disputes about the proper analysis of a technical concept of "free will" that no longer connects with ordinary concepts and practices, then these debates risk being irrelevant.

We take it to be particularly important for incompatibilists to establish the intuitive plausibility of their position, primarily because incompatibilist theories of free will are generally more metaphysically demanding than compatibilist alternatives. Incompatibilist theories require indeterminism in the agent at the right time and place, and often, additionally, agent causal powers. These conditions are typically required *in addition to*, rather than *instead of*, compatibilist conditions. Other things being equal, incompatibilists should motivate the need for these extra metaphysical conditions. Many incompatibilists have motivated their more metaphysically demanding theories, at least in part, by claiming that other things are *not* equal, because our ordinary intuitions, as well as our phenomenology of

decision-making, support incompatibilist views. It is certainly unclear why, *without* wide-scale intuitive support for incompatibilism, the burden of proof would be on compatibilists.[3]

Motivated by these considerations and the lack of any empirical data on what intuitions laypersons actually have, Eddy Nahmias, Stephen Morris, Thomas Nadelhoffer, and Jason Turner (2005, 2006) developed the initial experimental philosophy studies on folk intuitions about FW, MR, and determinism. Using three different descriptions of determinism, they found that a significant majority of participants (typically 65-85%) judged that agents in a deterministic scenario act of their own free will and are morally responsible. One of the descriptions of determinism was the following "re-creating universe" scenario:

> Imagine there is a universe (Universe C) that is re-created over and over again, starting from the exact same initial conditions and with all the same laws of nature. In this universe the same initial conditions and the same laws of nature cause the exact same events for the entire history of the universe, so that every single time the universe is re-created, everything must happen the exact same way. For instance, in this universe a person named Jill decides to steal a necklace at a particular time and then steals it, and *every* time the universe is re-created, Jill decides to steal the necklace at that time and then steals it.[4]

After reading the scenario, participants were asked to judge whether Jill decided to steal the necklace of her own free will and whether "it would be fair to hold her morally responsible (that is, blame her) for her decision to steal the necklace": 66% of subjects judged that Jill acted of her own free will, and 77% judged her to be morally responsible. Similar results were found using two other scenarios, some of which included agents' performing positive actions (saving a child) or neutral actions (going jogging).[5] These results offer evidence that a significant majority of laypersons are in fact "natural *compatibilists*" and thus call for an explanation for why so many philosophers have assumed that determinism is intuitively threatening to free will and moral responsibility.

In response to these results from Nahmias, Morris, Nadelhoffer, and Turner (NMNT), one might argue that people only *appear* to have compatibilist intuitions, when in fact they do not. Rather, such judgments might be unreliable, or not reflect people's *considered* beliefs or folk theories about FW and MR. In part to provide such an error theory for people's compatibilist judgments, Shaun Nichols and Joshua Knobe (2007) developed experiments aimed to explore the psychological mechanisms that generate intuitions about moral responsibility. In their studies, participants were randomly assigned to one of two groups, one of which was presented with a scenario in the "abstract" condition, and the other in the "concrete" condition. The scenario in the *abstract* condition read:

> 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.

---

[3] For further development of the points raised in this paragraph see Nahmias *et al.* 2006, 30-33. Moreover, if *revisionism* is called for (Vargas, 2005), it's unclear why philosophers should revise the concept of free will to be more metaphysically demanding than required by our ordinary intuitions.

[4] The wording of the scenario as presented here is very slightly altered from that used by Nahmias *et al*. (2006) in order to reflect the exact wording we used in one of our new studies presented below.

[5] In one scenario (Jeremy), affirmative responses for FW were 76% (negative action), 68% (positive), and 79% (neutral), and for MR they were 83% (negative) and 88% (positive). For the other scenario (Fred & Barney), affirmative responses for FW were 76% (negative action) and 76% (positive), and for MR they were 60% (negative) and 64% (positive) (see Nahmias *et al.,* 2006, 39). All results were significantly different from chance, as determined by χ^2 goodness-of-fit tests.

> 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.

In the *concrete* condition, this scenario was followed by another paragraph:

> In Universe A, a man named Bill has become attracted to his secretary, and he decided 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.

In the abstract condition, participants were asked:

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

and in the concrete condition:

> Is Bill fully morally responsible for killing his wife and children?
>
> Yes No

In the concrete condition, 72% of subjects gave the compatibilist response that Bill *is* fully morally responsible. In the abstract condition, however, 84% gave the purportedly *incompatibilist* response that it is *not* possible for a person to be fully morally responsible in Universe A.

Nichols and Knobe (N&K) claim that this disparity between participants' responses to the abstract and concrete cases is due to the psychological mechanisms driving people's intuitions. They suggest that the immoral action in the concrete case engages people's emotions in a way that leads them to offer compatibilist judgments, but that these judgments are the result of a performance error of one's normal capacity to make correct attributions of moral responsibility. According to this *affective performance error model*, the emotions induced by high-affect scenarios, such as an agent's murdering his spouse and children, skew people's attributions of MR. Compatibilist responses are "performance errors brought about by affective reactions. In the abstract condition, people's underlying theory is revealed for what it is—incompatibilist" (2007, 672). The performance error model thus presents an error theory for compatibilist intuitions. Because they are the result of affect-produced error, these intuitions should not be assigned any real significance in a theory of moral responsibility. As N&K say, "if we could eliminate the performance errors, the compatibilist intuitions should disappear" (2007, 678). N&K are rightly tentative about the affective performance error model, and also discuss the possibility that an *affective competence model*, a *concrete competence model*, or some hybrid model might instead be correct. N&K note that "we don't

yet have the data we need to decide between these competing models," but they claim that "the philosophical implications of the performance error model have a special significance because the experimental evidence gathered thus far [including NMNT's] seems to suggest that the basic idea behind this model is actually true" (2007, 678).[6]

We believe that there are several problems with Nichols and Knobe's (2007) study, as well as with their favored performance error model. For instance, N&K do not ask any questions about free will because, they say, "the expression 'free will' has become a term of philosophical art, and it's unclear how to interpret lay responses concerning such technical terms" (682, note 3). We do not think 'free will' should be treated as a technical term nor that ordinary intuitions about free will are irrelevant to philosophical debates, so we continue to ask questions about it in our studies here. N&K instead use just one experimental question, asking whether it is possible for a person to be "fully morally responsible" for their actions. This phrase (itself somewhat technical sounding) is ambiguous and likely to be understood differently by different people. First, the "fully" may be contrasted with being *partially* responsible or contrasted with being responsible in some *lesser* sense not involving moral desert. Indeed, attributions of MR are notoriously ambiguous between holding people responsible for forward-looking (e.g., deterrent) reasons and holding them responsible for backward-looking (e.g., retributive) reasons. Thus, we think it is difficult to move directly from people's responses to questions about being "fully morally responsible" to their intuitions about the compatibility of determinism and the sort of MR (or FW) involved in philosophical debates. Though we're unsure how best to address these issues, we think questions about whether agents *deserve* blame (or praise) are more likely to elicit the relevant notion of moral responsibility, and we use this language in our studies.[7]

We are more concerned with N&K's description of determinism. For instance, they write that in Universe A, "given the past, each decision *has to happen* the way that it does." This wording leaves the scope of the modal operator ambiguous. It may be interpreted as: "Given past events, it is necessary (or inevitable) that latter events (e.g., decisions) happen the way they do" rather than: "It is necessary that, given past events, later events (e.g., decisions) occur."[8] The latter, correct reading allows that later events (effects) *could* be

---

[6] Interestingly, Paul Edwards anticipated this model 50 years earlier: "The very same persons, whether educated or uneducated, use it [MR] in certain contexts in the one sense and in other contexts in the other. Practically all human beings … use what [C.A.] Campbell calls the unreflective conception when they are dominated by violent emotions like anger, indignation, or hate, and especially when the conduct they are judging has been personally injurious to them. On the other hand, a great many people, whether they are educated or not, will employ what Campbell calls the reflective conception when they are not consumed with hate or anger—when they are judging a situation calmly and reflectively and when the fact that the agent did not ultimately shape his own character has been vividly brought to their attention" (1958, 111). Edwards goes on to suggest that the "reflective conception" is the right one because the "unreflective conception" is driven by emotional bias.

[7] Moreover, some participants may interpret the MR question in the concrete case as: "Should Bill be punished for his action?" and even if they do *not* think he has free will or "full moral responsibility," they may think he needs to be punished. This interpretation might not be primed in the abstract case since there is no specific human action to be (potentially) punished. Finally, notice a subtle but important difference: Universe A does not mention *humans*, whereas Universe B explicitly mentions that the "one exception is human decision making," and it concludes: "each human decision *does* ***not*** *have to happen* the way that it does." This may prime some readers to think that Universe B is more like our universe, especially in the abstract condition, which is *not* then followed by the description of Bill, whose behaviour suggests that he is human. In that case, some of the differences in results between N&K's abstract and concrete cases might also be explained by a difference in intuitions people have about moral responsibility when asked about an 'alternate universe' (A) vs. a 'real-world universe' (B), differences that have been demonstrated in Nahmias, Coates, and Kvaran (2007) and Nichols and Roskies (2008).

[8] That is, the description suggests that determinism entails [(Po & L) ⊃ □P], rather than the proper □[(Po & L) ⊃ P].

otherwise as long as earlier events (causes) were otherwise. The former reading, however, mistakenly conflates determinism with fatalism (that all actual events are necessary or inevitable), and it negates a compatibilist understanding of the ability to do otherwise (see above), because one's actions *have to* happen *even if* the past (e.g., one's reasons) had been different (see Nahmias 2006 and Turner and Nahmias 2006). Furthermore, the concluding sentence of N&K's abstract scenario reads, "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." Some participants may read this to mean that (by contrast), in Universe A, each human decision *does* ***have*** *to happen* the way it does (full stop). We suspect that this reading, perhaps along with the other issues we've raised, may lead people to interpret N&K's description of determinism to suggest one or more of the following: that agents' actions could not happen otherwise *even if* the past had been different; that agents' decisions, beliefs, and desires are not playing a role in influencing their actions; or that agents have no control over what they do. That is, we predict that N&K's scenario will lead many people to interpret determinism to involve bypassing.

N&K claim that "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" (670-71). This response, however, neglects the possibility that the description of determinism has potentially misleading features that imply bypassing in the abstract case but *not* in the concrete case. We do not dispute the idea that the high negative affect likely induced in N&K's concrete case—with Bill's selfish, premeditated, and wanton murder of his wife and three children—may bias many participants to judge Bill to be morally responsible, but it may also lead participants to neglect features of the scenario that might otherwise mitigate their responsibility attributions. Indeed, it may be that the high negative affect causes participants to neglect the *bypassing* feature of the scenario. In other words, N&K's description of determinism may lead people to make a mistake, which is then "cancelled out" in the concrete case—but not in the abstract case—by high negative affect. Hence, we predict that most people will *not* read N&K's *concrete* scenario to involve bypassing, which may help to explain why they are generally willing to attribute MR to Bill.

While we agree with N&K that very high negative affect likely biases people to neglect potential responsibility-mitigating factors, we do not agree with the more general assumption that people are more competent in making judgments about FW, MR, and determinism, when they consider *abstract* cases than when they consider *concrete* cases including specific agents performing specific actions. On the contrary, assuming all else is equal (such as degree of affect), we believe that concrete conditions likely *facilitate* participants' comprehension and capacity to make accurate attributions of responsibility (in N&K's terms, we advocate a type of "concrete competence model"). Specifically, we believe that judgments about responsibility—including whether agents deserve credit or blame for their actions—will be more reliable if they engage our capacities to think about the beliefs, desires, and intentions of agents (e.g., our "theory of mind" capacities), and these are presumably more likely to be engaged when we consider specific agents in specific circumstances. More generally, it may be that people's intuitions are more reliable when they have more details about a scenario, which is likely part of the reason why philosophers construct thought experiments with specific details to probe (or prime) our intuitions.[9] Hence,

[9] Consider, by analogy, linguistic surveys intended to elicit people's intuitions concerning grammaticality. These surveys generally ask people to consider specific sentences rather than asking them to consider abstract questions about whether various constructions of sentences could be grammatical.

while we agree with N&K that concrete cases that *also* involve high affect may lead to errors, we do not believe this is a product of concreteness *per se*. Rather, we believe that, in general, concrete cases are more likely to reveal reliable intuitions about MR and FW than are abstract cases. For instance, we believe that N&K's description of determinism is more likely to lead to errors of interpretation in the abstract case.[10]

Finally, some other explanation is required for why so many more participants express incompatibilist intuitions in N&K's abstract scenario than in NMNT's cases, since the performance error model simply cannot account for this difference. The majority of participants in NMNT's (2006) studies gave compatibilist responses, even for those scenarios that *did not involve high affect*—i.e., those that involved positive actions, such as saving a child from a burning building or returning money one finds in a lost wallet, as well as those that involved neutral actions, such as going jogging. Moreover, no significant differences in responses were found between these cases and those that did involve negative actions—respectively, robbing a bank, stealing a necklace, and keeping the money one finds in a lost wallet (see note 5 above).[11] Thus, the performance error model does not provide an explanation for these previous results. Some other explanation for the difference in responses to N&K's and NMNT's cases is required. One possibility is that all of NMNT's scenarios describe *concrete* agents and actions and ask about those specific agents' FW and MR, whereas N&K's abstract scenario does not include or ask about specific agents or actions, but again, we think there is no good reason to think concreteness alone leads to performance errors. Another (non-exclusive) possibility is that N&K's description of determinism primes bypassing judgments more than NMNT's descriptions. Our study explores these possibilities.

An initial attempt to explore the issue of bypassing was developed in Nahmias, Justin Coates, and Trevor Kvaran (2007). They found that, across several different scenarios, most people responded that MR and FW were possible in a deterministic universe *if* the scenario described the decisions of agents in that universe as being "completely caused by the specific thoughts, desires, and plans occurring in our minds." In contrast, most people responded that FW and MR were *not* possible in a deterministic universe *if* the scenario described agents' decisions as "completely caused by the specific chemical reactions and neural processes occurring in our brains."[12] The latter, reductionistic description seems to prime people to think that agents' mental states are not playing the proper role in their actions—that their conscious self is bypassed. Thus, even though determinism in the technical sense is equally present in both scenarios, people tend to think determinism is compatible with FW and MR unless they take determinism to involve bypassing.

We designed our current studies in order to further explore the possible effects of bypassing on people's judgments of FW and MR, and to test our error theory for incompatibilist intuitions. We presented participants with different descriptions of determinism (N&K's scenario versus NMNT's "re-creating universe" scenario, with abstract

[10] For more discussion of differences in judgments about FW and MR based on abstract/concrete differences, as well as real world/alternate world differences, see Nahmias, Coates, and Kvaran (2007) and Nichols and Roskies (2008).

[11] Even NMNT's scenarios that do involve negative actions are not very "high-affect." Stealing a necklace, robbing a bank, and keeping $1000 found in a wallet seem more similar to N&K's (2007) *low-affect* condition, in which Bill cheats on his taxes, than N&K's *high-affect* condition, in which Bill murders his family.

[12] For instance, in one version (the real world cases), when the agents were described with the psychological predicates, 89% of participants said that agents should be held morally responsible and 83% said they had free will, whereas when the agents were described with the "neuro-reductionistic" predicates, only 40% said they had MR and 38% said they had FW.

and concrete versions of each), and then asked participants not only about FW and MR but also about bypassing. We predicted that:

1. In general, participants' judgments about bypassing would correlate significantly with their judgments about MR and FW. That is, when making judgments about agents in a deterministic universe, (a) most participants who respond that the agents do *not* have MR and FW would also agree that the agents' decisions, beliefs, and desires do *not* affect what happens—i.e., such participants would interpret the deterministic nature of the scenario to involve bypassing—whereas (b) most participants rejecting the bypassing claims would respond that the agents *do* have MR and FW. That is, bypassing judgments would explain away most *apparent* incompatibilist intuitions, whereas most people who do *not* misunderstand determinism to involve bypassing would express *prima facie* compatibilist intuitions.[13]
2. Judgments of FW and MR would be *lower*, while judgments of bypassing would be *higher*, in N&K's abstract scenario compared to NMNT's abstract scenario. That is, N&K's description of determinism would, in the abstract case, lead more people to misunderstand determinism.
3. Judgments of FW and MR would be *lower*, while judgments of bypassing would be *higher*, in the abstract scenarios compared to the concrete scenarios, with this difference especially pronounced in N&K's high-affect scenario.

III. Methods

Participants (included in the analysis) were 249 undergraduate students at Georgia State University (Atlanta, GA) who were randomly assigned to complete one of four versions of the experimental task.[14] We used software from QuestionPro to develop and administer these surveys online. Using a 2x2 between-subjects design, four scenarios were generated by systematically varying (1) whether the deterministic scenario was *N&K's* or *NMNT's*, and (2) whether the scenario was *abstract* or *concrete*.

Participants began by reading a general description of the task, providing informed consent, and then reading one of the four scenarios. After reading the scenario, participants answered a series of experimental questions designed to probe their intuitions about FW and MR, as well as whether they interpreted the scenario to involve bypassing. N&K's abstract and concrete scenarios read exactly as they are presented above, as did NMNT's concrete scenario. NMNT's abstract scenario replaces the last sentence of the concrete version with:

> For instance, in this universe whenever a person decides to do something, *every* time the universe is re-created, that person decides to do the same thing at that time and then does it.

In order to replicate N&K's study, participants given those surveys were first asked:

> Which of these universes do you think is most like ours? Universe A Universe B

[13] We call them "*prima facie* compatibilist intuitions" in part because we are not committed to the idea that ordinary people have the (positive) intuition that determinism is compatible with FW and MR—their intuitions may not be so theoretically rich. But we think that lacking intuitions that (genuinely) support incompatibilism is sufficient to say that people are "natural compatibilists." We also accept that there may be alternative explanations that suggest people are expressing only *apparent* compatibilist intuitions (see section V).

[14] Participants were 42% male and 58% female. They completed the surveys for class credit in critical thinking or psychology courses. We excluded from analysis 187 participants who (a) did not complete the entire survey, (b) responded incorrectly to either of two comprehension questions, or (c) completed the survey too quickly (i.e., in less time than two standard deviations from the mean time for completion). Studies were carried out under previous approval of the GSU Institutional Review Board.

Participants who were given the NMNT surveys were first asked:

> Is it possible that our universe could be like Universe C, in that the same initial conditions and the same laws of nature cause the exact same events for the entire history of the universe? Yes No

Participants were next asked to indicate their level of agreement with each of a series of statements using a 6-point rating scale (strongly disagree, disagree, somewhat disagree, somewhat agree, agree, strongly agree). The first statement was always the *moral responsibility* (MR) question (replicating N&K's format). The remaining statements in each survey were randomized to decrease the likelihood of order effects. The most important experimental questions we asked read as follows (variations between scenarios are in brackets: *N&K abstract* scenario asks about Universe A; *NMNT abstract* asks about Universe C; *N&K concrete* asks about Bill; *NMNT concrete* asks about Jill):

<u>The MR/FW questions</u>

> **MR**: In Universe [A/C], it is possible for a person to be fully morally responsible for their actions.
>
> [Bill/Jill] is fully morally responsible for [killing his wife and children / stealing the necklace].
>
> **FW**: In Universe [A/C], it is possible for a person to have free will.
>
> It is possible for [Bill/Jill] to have free will.
>
> **Blame**: In Universe [A/C], a person deserves to be blamed for the bad things they do.
>
> [Bill/Jill] deserves to be blamed for [killing his wife and children / stealing the necklace.]

<u>The Bypassing questions</u>

(These questions represent our way of operationalizing "bypassing"; we take it that philosophers on all sides of the free will debate should agree that if one *properly* understands determinism, one should *not* agree with these statements.)

> **Decisions**: In Universe [A/C], a person's decisions have no effect on what they end up being caused to do.
>
> [Bill's/Jill's] decision to [kill his wife and children / steal the necklace] has no effect on what [he/she] ends up being caused to do.
>
> **Wants**: In Universe [A/C], what a person wants has no effect on what they end up being caused to do.
>
> What [Bill/Jill] wants has no effect on what [he/she] ends up being caused to do.
>
> **Believes**: In Universe [A/C], what a person believes has no effect on what they end up being caused to do.
>
> What [Bill/Jill] believes has no effect on what [he/she] ends up being caused to do.
>
> **No Control**: In Universe [A/C], a person has no control over what they do.
>
> [Bill/Jill] has no control over what [he/she] does.[15]

[15] One might, however, think the "no control" question is inappropriate to indicate bypassing, perhaps because one believes that determinism *should* be interpreted to mean that agents have no control over what they do. We disagree with this view. In any case, while removing this question from our analyses does reduce the statistical strength of some of our results, it does not alter our general findings (see also note 17).

Past Different: In Universe A, everything that happens *has to* happen, even if what happened in the past had been different.

Bill *has* to kill his wife and children, even if what happened in the past had been different.[16]

After providing responses to these questions, participants then answered two comprehension questions to ensure that they understood the scenario and several demographic questions (e.g., gender, age, religious affiliation, etc.).

IV. Main Results

In order to examine the relationship between participants' judgments about bypassing and their judgments about MR and FW, we created two composite scores which we used for the analyses below: an *MR/FW composite score*, which was obtained by computing the average of each participant's responses to the MR, FW, and Blame questions, and a *composite Bypassing score,* obtained by computing the average of each participant's responses to the Decisions, Wants, Believes, and No Control questions.[17] Initial visual inspection of the data suggested that, as we predicted, (i) *MR/FW* scores were lower and *Bypassing* scores higher in response to the abstract scenarios compared to the concrete scenarios, and (ii) *MR/FW* scores were lower, while *Bypassing* scores were higher, in response to *N&K's abstract* scenario compared to *NMNT's abstract* scenario (see Figure 3). Moreover, across conditions, (i) the majority of participants who gave *apparent* incompatibilist responses (*MR/FW* scores less than the 3.5 midpoint) also gave *high* bypassing responses (Bypassing scores > 3.5), whereas (ii) most participants who gave *prima facie* compatibilist responses (*MR/FW* scores > 3.5) also gave *low* (< 3.5) bypassing responses (see Figure 4).[18] Given these findings, we employed a series of analyses in order to determine whether these results were statistically significant.[19]

[INSERT FIGURES 3 and 4 NEAR HERE]

To determine whether *MR/FW* scores were significantly lower in N&K's surveys than in NMNT's surveys and lower in the abstract conditions than in the concrete conditions, we ran a 2 (survey: N&K, NMNT) x 2 (condition: abstract, concrete) Analysis of Variance (ANOVA) on the mean *MR/FW* composite scores (see Figure 5). The ANOVA showed a significant main effect for survey: $F(1, 245) = 5.396$, $p = .021$, a significant main effect for condition: $F(1, 245) = 61.058$, $p < .001$, and a marginally significant interaction effect: $F(1, 245) = 3.297$, $p < .071$. Thus, N&K's surveys received significantly lower *MR/FW* scores

[16] This question was not asked in the NMNT surveys, though the following similar statements were used: "If Universe C were re-created with *different* initial conditions or *different* laws of nature, it is possible Jill would *not* [mow her lawn/steal the necklace] at that time."

[17] These composite scores provide a more robust measure of people's intuitions concerning MR, FW, and bypassing. Lest one worry about averaging these scores, and in so doing losing information about responses to each particular question, responses to all questions factored into each composite score were, with very few exceptions, highly positively intracorrelated with one another in all conditions. Across conditions, reliability analyses produced a Cronbach's alpha of .807 among the questions used to compute the *MR/FW* composite score, and a Cronbach's alpha of .823 among the questions used to compute the *Bypassing* composite score, indicating that both composite scores were strongly internally consistent.

[18] Data reported in Figures 3 and 4 does not include the composite scores for 20 of the 249 participants whose *Bypassing* composite scores were equal to the 3.5 midpoint.

[19] While we did replicate N&K's overall findings, our results were slightly different than theirs. In our study, 68% of participants in the abstract condition gave the apparent incompatibilist response that it is *not* possible for a person to be fully morally responsible in Universe A, compared to the 84% in N&K's previous study. Also, 87.5% of participants in our present study responded that Bill is fully morally responsible in N&K's concrete scenario, compared to 72% in N&K's study.

than NMNT's, independent of whether the condition was abstract or concrete, and the abstract scenarios received lower *MR/FW* ratings than the concrete scenarios, independent of whether the survey was N&K's or NMNT's. We ran an additional pre-planned *t*-test specifically comparing the mean *MR/FW* composite responses to *N&K abstract* vs. *NMNT abstract*. As hypothesized, we found that the mean *MR/FW* score was significantly lower in *N&K abstract* than in *NMNT abstract*: $t(1, 131) = -2.973$, $p = .004$.

[INSERT FIGURE 5 NEAR HERE]

To determine whether *Bypassing* scores were significantly higher in N&K's surveys than in NMNT's surveys and higher in the abstract conditions than in the concrete conditions, we ran a 2 (survey: N&K, NMNT) x 2 (condition: abstract, concrete) ANOVA on the mean *Bypassing* composite scores (see Figure 5). The ANOVA showed a significant main effect for condition: $F(1, 245) = 20.665$, $p < .001$, but only a near-significant effect for survey: $F(1, 245) = 3.463$, $p = .064$ (see note 20). There was no significant interaction effect. Thus, the abstract scenarios received significantly higher *Bypassing* scores than the concrete scenarios, independent of whether the survey was N&K's or NMNT's, and N&K's surveys received marginally higher *Bypassing* scores than NMNT's, independent of whether the condition was abstract or concrete. We ran an additional pre-planned *t*-test specifically comparing the mean *Bypassing* responses to *N&K abstract* vs. *NMNT abstract*. As hypothesized, we found that the mean *Bypassing score* was significantly higher in *N&K abstract* than in *NMNT abstract*: $t(1, 131) = 2.319$, $p = .022$.[20]

In order to statistically assess the relationship between these two variables of interest, we computed Pearson correlation coefficients between the *MR/FW* and *Bypassing* composite scores for each scenario. Consistent with our hypothesis, but even more dramatically than we expected, we found a strong inverse correlation between *Bypassing* and *MR/FW* scores—that is, the higher a participant's *Bypassing* score, the lower his or her *MR/FW* score, and vice versa—in each of the four scenarios (*N&K abstract*: $r(75) = -0.695$, $p < .001$; *N&K concrete*: $r(54) = -0.569$, $p < .001$; *NMNT abstract*: $r(54) = -0.803$, $p < .001$; *NMNT concrete*: $r(58) = -0.708$, $p < .001$). Collapsing across all four surveys, the correlation coefficient between *Bypassing* and *MR/FW* scores was strikingly high: $r(247) = -0.734$, $p < .001$.

Consistent with our hypothesis, then, average scores in response to the Bypassing questions were significantly *higher* in *N&K's abstract* scenario than in *NMNT's abstract* scenario, average scores in response to the MR/FW questions were significantly *lower* in *N&K abstract* than in *NMNT abstract*, and responses to the Bypassing and MR/FW questions were strongly inversely correlated across scenarios. Given these results, we further suspected that responses to the MR/FW questions were lower for *N&K abstract* compared to *NMNT abstract* precisely *because* participants interpreted N&K's scenario to involve a higher degree of bypassing. We hypothesized that the degree to which one interpreted a scenario to involve bypassing would *mediate* the relationship between survey and MR/FW responses. That is, we hypothesized that the difference in MR/FW responses between the two *abstract* conditions of the surveys was *caused* largely by people's bypassing judgments. In order to test this causal hypothesis more directly, we used a mediation analysis.[21]

---

[20] A *t*-test comparing the mean *Bypassing* responses in *N&K concrete* and *NMNT concrete* was *not* significant, which explains why the ANOVA did not show a significant main effect for survey. We suspect that this lack of effect is due to the very high affect in the *N&K concrete* scenario driving down participants' Bypassing scores.

[21] Because *N&K's concrete* scenario involves high affect in a way that *NMNT's concrete* scenario does not, we did not include either concrete scenario in the mediation analysis, as doing so would introduce another, potentially confounding, variable (high affect) in addition to condition (concrete vs. abstract) (see also note 20).

We conducted three regression analyses to test for mediation (see Figure 6), as outlined by Baron and Kenny (1986) and more recently by MacKinnon *et al*. (2002).[22] The first regression equation used survey type (*N&K abstract, NMNT abstract*) to predict *MR/FW* score (path c), and yielded a significant effect: $t(132) = 2.973$, $p = .004$, corroborating the above results showing that *N&K abstract* prompts participants to respond that agents do not have MR and FW more than does *NMNT abstract*. The second regression equation estimated changes in *Bypassing* score using survey (path a) and also yielded a significant effect: $t(132) = -2.319$, $p = .022$, corroborating the above results showing that participants interpreted *N&K abstract* to involve bypassing more than they did *NMNT abstract*. The third equation estimated *MR/FW* score using both survey type and *Bypassing* score. The link between *Bypassing* and *MR/FW* scores (path b) was highly significant: $t(132) = -12.799$, $p < .001$, and the relation (path c) between survey and *MR/FW* score was reduced to *non-significance* once *Bypassing* was included in the model: $t(132) = 1.821$, *ns*; Sobel test = 2.286, $p < .022$. Thus, all the conditions of mediation were met: survey type was a significant predictor of *MR/FW* and of *Bypassing* scores, and *Bypassing* was a significant predictor of *MR/FW* scores while controlling for survey.

[INSERT FIGURE 6 NEAR HERE]

Thus, as we hypothesized, the effect that survey type (i.e., the description of determinism in the abstract scenarios) had on a participant's MR/FW responses was mediated by whether the participant interpreted the description to involve bypassing. That is, these results suggest that which survey participants read (*N&K* versus *NMNT*) had *no significant causal effect* on their MR/FW scores *over and above* the effect it had in virtue of causing different interpretations of whether the scenario described involved bypassing.

V. Discussion

We predicted that participants across surveys and conditions would be *more* likely to judge determinism to threaten free will and moral responsibility when they interpreted determinism to involve bypassing. We also predicted that participants would be *more* likely to interpret determinism to involve bypassing in N&K's abstract scenario than in NMNT's abstract scenario, and that this would help to explain why the description of determinism in N&K's scenario leads people to be *less* likely to attribute FW and MR to agents. Finally, we predicted that participants would be *less* likely to interpret determinism to involve bypassing in concrete scenarios, especially in N&K's high-affect case, and hence *more* likely to attribute FW and MR to the agents in those scenarios. Our results strongly support each of these predictions. Indeed, not only do our results show that (i) the vast majority of participants who express *apparent* incompatibilist intuitions interpret determinism to involve bypassing, while those who express *prima facie* compatibilist intuitions tend *not* to, and that (ii) there is a dramatic correlation between the degree to which participants take a scenario to

[22] Mediation analysis involves the specification of a causal model between three variables. Suppose that an *initial variable*, *X* (in our case, survey type), is assumed to have a causal effect on an *outcome variable*, *Y* (in our case, MR/FW responses). Call *c* the direct effect of *X* on *Y*. A mediational model of the relationship between them, then, is one in which the causal effect of *X* on *Y* is *mediated* by an *intervening variable*, *M* (in our case, bypassing judgments). Call *a* the effect of the initial variable *X* on *M*, and *b* the effect of *M* on the outcome variable *Y*. *Complete mediation* obtains when variable *X* no longer has any direct effect on *Y* when *M* is controlled for, such that path *c* is zero. *Partial mediation* obtains when *c* is reduced if *M* is controlled for, but not to zero, because paths *a* and *b* account for some, but not all, of the overall causal effect of *X* on *Y*. Mediational models can be assessed statistically by mediation analysis, which uses multiple regression analyses to estimate the values of paths *a*, *b*, and *c*.

involve bypassing and the degree to which they attribute MR and FW to agents in that scenario, but the results also suggest that (iii) the difference in attributions of MR and FW between the abstract conditions is *caused by* people's bypassing interpretations.

We think it is safe to conclude from our results that there is an important connection between (a) whether people take a description of determinism to entail bypassing and (b) whether people take the scenario so described to preclude free will and moral responsibility. The most plausible interpretation of this connection is that when a person takes determinism to entail bypassing it generally *causes* him or her to judge that determinism precludes MR and FW, and hence to offer *apparent* incompatibilist intuitions. Conversely, when a person does *not* take determinism to entail bypassing, they are likely to offer compatibilist intuitions—they do not see any conflict between determinism and FW or MR (below we will further consider whether such intuitions should count as supporting compatibilism). This causal conclusion also draws support from Nahmias *et al*.'s (2007) study, which manipulated a type of bypassing directly—rather than measuring responses to it—and found that it significantly influenced participants' attributions of MR and FW. Thus, previous evidence supports the conclusion that bypassing judgments mediate attributions of MR and FW, rather than the other way around.

If these interpretations of the data are correct, then people's interpreting determinism to entail bypassing may be the best explanation for ordinary people's intuitions that *appear* to support the incompatibility of determinism and FW and MR. But these intuitions do *not* properly support incompatibilism because determinism does *not* properly entail bypassing as we have operationalized it here. Determinism, properly understood, simply does *not* entail that our decisions, beliefs, and desires have no effect on what we end up doing, nor that we have no control over what we do (see note 15), nor that our actions *have to* happen just as they do *even if* the past had been different (see below). When people do not misinterpret determinism in these ways, they usually do not take it to threaten FW and MR.

Furthermore, our results suggest that certain descriptions of determinism and certain conditions increase the degree to which people will misinterpret determinism to entail bypassing (and hence to lower their attributions of MR and FW). Specifically, Nichols and & Knobe's description of determinism in the abstract condition has this effect, and this likely occurs because of the problems we pointed out earlier with their description of determinism, including the use of language that may suggest that decisions and actions in Universe A *have to* happen *even if* the past had been different. Indeed, 48% of participants in the *N&K abstract* condition responded that, in Universe A, everything that happens has to happen the way it does even if the past had been different, and these responses were significantly correlated with the *MR/FW* score.[23] The degree to which people interpret determinism to involve bypassing is also higher in the abstract conditions than in the concrete conditions, perhaps because descriptions of concrete agents and actions prime people to think about the effectiveness of agent's beliefs, desires, and decisions (e.g., consider how difficult it is to agree with the statement: **"**What Bill wants has no effect on what he ends up being caused to do"—his desire to be with his secretary was precisely what led to his murderous actions!). And when the concrete conditions induce *high* negative affect, as with N&K's murderous

[23] Pearson correlation coefficient between Past Different (M = 3.66, SD = 1.57), and *MR/FW* composite scores in *N&K abstract*: $r(75) = -.277$, $p = .015$. By comparison, in the *NMNT abstract* scenario only 7% disagreed with (i.e., "missed") the following similar question (M = 2.68, SD = 1.40): "Suppose that in Universe C, a person named Jill decides to mow her lawn at a particular time and then does it. If Universe C were re-created with *different* initial conditions or *different* laws of nature, it is possible Jill would *not* mow her lawn at that time."

Bill, this tends to lower people's bypassing judgments while leaving their judgments of FW and MR high.

As we said, we find it plausible that very high negative affect can induce biases in MR judgments; there is, after all, evidence of such biases (see Nichols & Knobe 2007, 672). But this does not suggest that, in general, people get things wrong in concrete cases and right in abstract cases. Indeed, given that interpreting determinism to entail bypassing is a *mistake*, it appears that scenarios with abstract descriptions of unspecified agents performing unspecified actions may bias people toward making this mistake. This is unsurprising, since thinking in terms of the efficacy of agents' mental states—taking the intentional stance—is more likely to occur when one is thinking about specific agents performing specific actions.

Due to these considerations, we believe that concrete scenarios involving low affect are more useful for testing people's intuitions about FW, MR, and (in)compatibilism than either abstract or high-affect scenarios. Indeed, fewer participants made the mistake of interpreting determinism to involve bypassing in *NMNT concrete* than in any other scenario. Furthermore, every one of the few participants who did respond as an apparent incompatibilist in *NMNT concrete* also interpreted the scenario to involve bypassing, while *every* participant who did *not* interpret the scenario to involve bypassing responded as a *prima facie* compatibilist. Thus, the scenario elicited fewer mistakes regarding bypassing and less variability in responses among participants' who did not make that mistake. Future research should explore these issues. For instance, it would be useful to compare responses to bypassing and MR/FW questions in other concrete cases, such as N&K's low-affect concrete case of the tax cheat and NMNT's cases involving positive actions (e.g., agent saving a child) and neutral actions (e.g., agent mowing the lawn), and perhaps also to develop high-affect *abstract* cases.

Hence, we conclude that the extant research in experimental philosophy on free will converges on the conclusion that most laypersons do *not* have *genuine* incompatibilist intuitions. They do have the intuition that bypassing undermines FW and MR, and they can be primed to judge that determinism entails bypassing. But the latter judgment is based on a mistaken interpretation of determinism. The judgment that bypassing undermines FW and MR does not support incompatibilism. If anything, it supports compatibilist theories of freedom and responsibility, since those theories emphasize that FW and MR require that our conscious, rational deliberative processes play the right role in producing our decisions and actions. So, folk intuitions that take bypassing of these processes to preclude FW and MR support compatibilist theories. People seem to be attending to compatibilist conditions for FW and MR—whether agent's actions are properly caused by their decisions, beliefs, desires, etc. The more likely people believe that these conditions are met, the more likely they are to attribute FW and MR; the more likely people believe that these conditions are *not* met—e.g., because of bypassing—the less likely they are to attribute FW and MR. These results support the claim that people have merely *apparent* incompatibilist intuitions—most people do *not* seem to think that determinism—*without* bypassing—precludes FW and MR (see Figure 2).

There are several ways that incompatibilists might object to these interpretations of our results, but space limits us to consider only some of them briefly—we leave it to our critics to complete the task.

First, one might argue that even *after* recognizing that determinism does not involve bypassing, people would (or should) recognize that determinism threatens FW and MR, and that those who do not recognize this are likely failing to understand the deterministic nature of the scenario. This response, in effect, offers a debunking explanation for *prima facie* compatibilist intuitions by arguing that participants who express such judgments do so

because they fail to understand determinism or its implications. Perhaps they do not recognize that determinism is incompatible with AP or US, but if they *did,* they would think it was incompatible with MR and FW (even while also recognizing that determinism does not involve bypassing).

This is an interesting objection, and we would be intrigued to see experiments to test for it. Recall that the vast majority of our participants who did *not* take determinism to involve bypassing also attributed FW and MR to agents in those scenarios, so this objection requires that most of these participants are failing to understand the deterministic nature of the scenario or failing to understand that these agents do not meet conditions *the participants themselves* take to be necessary for free will (e.g., AP or US *in the incompatibilists' sense*).[24] Future research should try to elicit whether people understand determinism to conflict with AP and US and whether people take AP and US—understood in ways that are incompatible with determinism—to be necessary for free will or moral responsibility. Unfortunately, it is difficult to properly describe what it means for an agent to be the ultimate source of her decisions, or to have an *unconditional* ability to do otherwise, without a good bit of explanation, which might verge on "intuition coaching." While we are inclined to think that this difficulty suggests that AP and US are not particularly "natural" or intuitive to non-philosophers, others may argue that they *are* intuitive once people properly understand the relevant ideas. This suggests a second objection to our interpretation of the data.

The incompatibilist might argue that untutored intuitions are simply *irrelevant* to the philosophical debates about free will and moral responsibility or, what is different, that the sort of studies carried out by experimental philosophers cannot uncover information about the relevant intuitions. These objections might be motivated by the belief that the issues are so complex that responses from untrained individuals reveal little to nothing about the truth, by the conviction that philosophers can discern the relevant intuitions by considering their own intuitions or those adduced from their students and other folk, or by the belief that the methods employed by experimental philosophers simply cannot do the job they aim to do (see, e.g., Kauppinen, 2007).

Extensive responses to such objections applied to experimental philosophy in general have been offered elsewhere (e.g., Nadelhoffer and Nahmias, 2007; Nahmias *et al*., 2006; Knobe and Nichols, 2008; Weinberg, 2007). We reiterate that philosophical debates about free will and moral responsibility require an understanding of the way non-philosophers think about these issues in order to develop a theory that accords with folk intuitions, where possible, and to know what it is that we're revising (and why) where revision is advocated. We agree with David Lewis when he says that in developing philosophical theories "we are trying to improve *that* theory, that is to leave it recognizably the same theory we had before" (1986, 134). If the studies we have presented here or previous studies have design flaws, then attempts should be made to improve them, rather than abandoning the very idea of understanding folk intuitions in an empirically informed way. Part of this process might

[24] The only data we have that appears relevant does *not* support this hypothesis. In the *NMNT abstract* scenario we asked: "Suppose that in Universe C, a person named Jill decides to mow her lawn at a particular time and then does it. If Universe C were re-created with the *same* initial conditions and the *same* laws of nature, it is possible Jill would *not* mow her lawn at that time." (In *NMNT concrete* we asked whether it is possible Jill would *not* steal the necklace). The objection under consideration would predict there to be a significant correlation between responses to these questions and responses to the MR/FW questions—e.g., people who say it *is* possible for Jill to do otherwise may be neglecting to see that determinism rules out AP and so should be more likely to attribute MR and FW to her. However, there was no significant correlation between responses to this question and MR/FW composite scores in *NMNT abstract*: $r(54) = .137$, *ns*, and only a marginally significant correlation in *NMNT concrete*: $r(58) = .242$, $p = .06$.

include making sure that participants understand the concepts involved as clearly as possible, but one of the motivations for surveying people untrained in the philosophical debates is the worry that philosophical training may end up shaping intuitions toward a certain theory. For instance, it is not uncommon for philosophy teachers to initially present determinism using metaphors that suggest fatalism or epiphenomenalism (or, on the other side, to present indeterminism as involving entirely random uncaused events). Finally, since professional philosophers writing about free will tend to have theoretical commitments and hence *post*-theoretical "intuitions," and since these "intuitions" as well as reports about folk (e.g., students') intuitions tend to conflict with each other, it is appropriate to attempt, as best we can, to uncover information about *pre*-theoretical intuitions, their sources, and their reliability.

VI. Conclusion

Incompatibilists suggest that free will and moral responsibility require conditions that are incompatible with determinism—conditions that are generally more demanding than those required by compatibilists. One way to motivate the claim that these conditions are indeed necessary is to argue that incompatibilism is intuitive, and that compatibilism is thus a "quagmire of evasion," a revision of the way ordinary people think about these issues—to suggest that "ordinary persons have to be talked out of this natural incompatibilism by the clever arguments of philosophers" (Kane, 1999, 217). Our evidence here suggests that people may instead need to be talked *into* incompatibilism by the clever arguments of philosophers. We suggest that incompatibilism only *appears* to be intuitive, largely because determinism is misinterpreted. Indeed, it is misinterpreted such that it precludes the very conditions compatibilists identify with free and responsible agency. It is possible that incompatibilism is actually intuitive even *after* this mistake is corrected—that people find determinism threatening even if they understand that it does *not* involve bypassing (e.g., fatalism or epiphenomenalism). We await the evidence. And obviously, *some* people—for instance, some philosophers—do have genuine incompatibilist intuitions. But if most people think that free will and moral responsibility can exist even if determinism (properly construed) is true, the argumentative burden shifts to these philosophers to explain why people's intuitions need to be revised so that they accept a more demanding theory. We await the argument.[25]

[25] We would like to thank the following people for helpful comments on earlier drafts: Shaun Nichols, Al Mele, Jason Turner, Stephen Morris, Neil Levy, Tamler Sommers, George Graham, Dan Weiskopf, Joshua Knobe, Reuben Stern, Jason Shepard, Thomas Nadelhoffer, Trevor Kvaran, Fiery Cushman, and especially Bradley Thomas. This article was completed in part with support from a grant (for E.N.) from the University of Chicago Arete Initiative and the John Templeton Foundation.

References

Baron, Reuben M. & Kenny, David A. 1986. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* 51: 1173-1182.

Blumenfeld, D. 1988. Freedom and mind control. *American Philosophical Quarterly* 25: 215-227.

Chalmers, David. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.

Fischer, John and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility.* Cambridge: Cambridge University Press.

Frankfurt, Harry. 1971. Freedom of the Will and the Concept of a Person. In *The Importance of What We Care About* (Cambridge University Press, 1988), 11-25.

Kane, Robert. 1999. Responsibility, luck, and chance: reflections on free will and indeterminism. *Journal of Philosophy* 96: 217-240.

Kauppinen, Antti. 2007. The rise and fall of experimental philosophy. *Philosophical Explorations*.

Knobe, Joshua & Nichols, Shaun. 2008. An experimental philosophy manifesto. In *Experimental Philosophy*, *ed.* J. Knobe and S. Nichols (Oxford University Press).

Knobe, Joshua & Doris, John. Forthcoming. Strawsonian Variations: Folk Morality and the Search for a Unified Theory. In *The Handbook of Moral Psychology, ed.* J. Doris (Oxford University Press).

Lewis, David. 1986. *The Plurality of Worlds*. Oxford: Blackwell Publishers.

Lycan, William. 2003. Free will and the burden of proof. In *Proceedings of the Royal Institute of Philosophy for 2001-02*, *ed.* Anthony O'Hear. Cambridge University Press, 107-122.

MacKinnon, David P., Lockwood, Chondra M., Hoffman, Jeanne M., West, Stephen G., and Sheets, Virgil. 2002. A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods* 7: 83-104.

Mele, A. 2005. *Autonomous Agents*. New York: Oxford University Press.

Nadelhoffer, Thomas and Eddy Nahmias. 2007. The Past and Future of Experimental Philosophy. *Philosophical Explorations* 10.2: 123-149.

Nahmias, Eddy. 2006. Folk fears about freedom and responsibility: Determinism vs. reductionism. *Journal of Cognition and Culture* 6: 215-237.

Nahmias, Eddy, Stephen Morris, Thomas Nadelhoffer, and Jason Turner. 2005. Surveying Freedom: folk intuitions about free will and moral responsibility. *Philosophical Psychology* 18: 561-584.

Nahmias, Eddy, Stephen Morris, Thomas Nadelhoffer, and Jason Turner. 2006. Is incompatibilism intuitive? *Philosophy and Phenomenological Research* 73: 28-53.

Nahmias, Eddy, D. Justin Coates & Trevor Kvaran. 2007. Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions. *Midwest Studies in Philosophy* 31: 214-242.

Nichols, Shaun & Knobe, Joshua. 2007. Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions. *Nous* 41: 663-685.

Nichols, Shaun & Roskies, Adina. Forthcoming. Bringing moral responsibility down to earth.

O'Connor, Timothy. 2005. "Freedom With a Human Face," *Midwest Studies in Philosophy*, 29, 207-227.

Pereboom, Derk. 2001. *Living Without Free Will.* Cambridge University Press.

Strawson, Galen.1986. *Freedom and Belief*. Oxford: Clarendon.

Turner, Jason & Nahmias, Eddy. 2006. Are the Folk Agent Causationists? *Mind and Language* 21: 597-609.

van Inwagen, Peter. 1983. *An Essay on Free Will*. Oxford: Clarendon Press.
Vargas, Manuel. 2005. The Revisionist's Guide to Responsibility. *Philosophical Studies* 125: 399-429.
Watson, Gary. 1975. Free Agency. In *Free Will, ed.* G. Watson (Oxford University Press, 1982), 96-110.
Weinberg, Jonathan. 2007. How to Challenge Intuitions Empirically Without Risking Skepticism. *Midwest Studies in Philosophy* 31: 318 – 343.
Wolf, Susan. 1990. *Freedom within Reason*. Oxford University Press.

FIGURE 1: *Genuine* Incompatibilist Intuitions

Genuine incompatibilists take determinism to preclude unconditional alternative possibilities (AP) and/or ultimate sourcehood (US), and in turn take the lack of these abilities to preclude free will (FW) and moral responsibility (MR). Genuine incompatibilists do *not* misunderstand determinism to involve or entail bypassing.

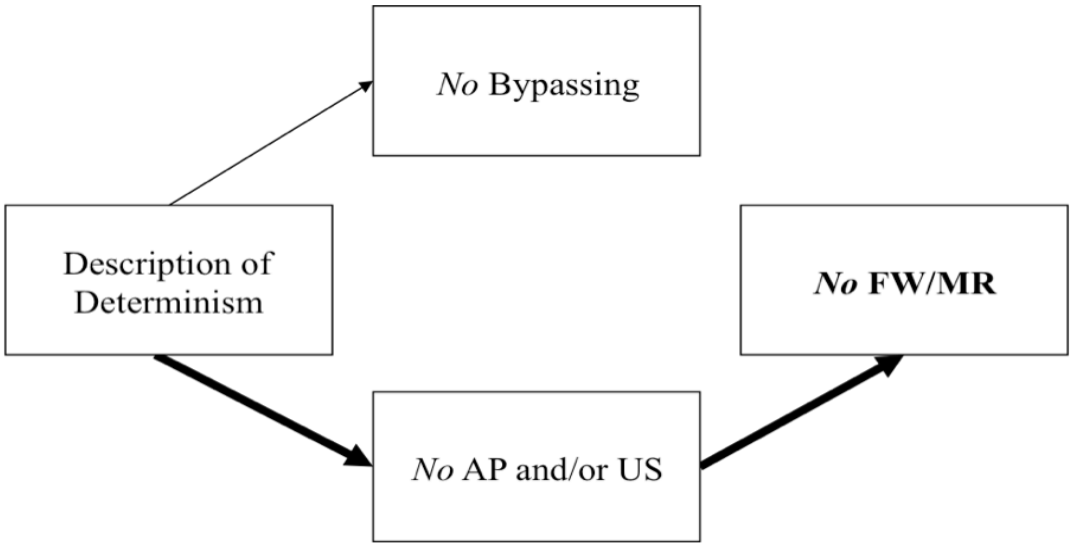

FIGURE 2: *Apparent* Incompatibilist Intuitions

Apparent incompatibilists *mistakenly* take determinism to involve or entail bypassing, and in turn take bypassing to preclude free will (FW) and moral responsibility (MR). Apparent incompatibilists may also take determinism to preclude AP and/or US, but they do not infer incompatibilism from the lack of these abilities.

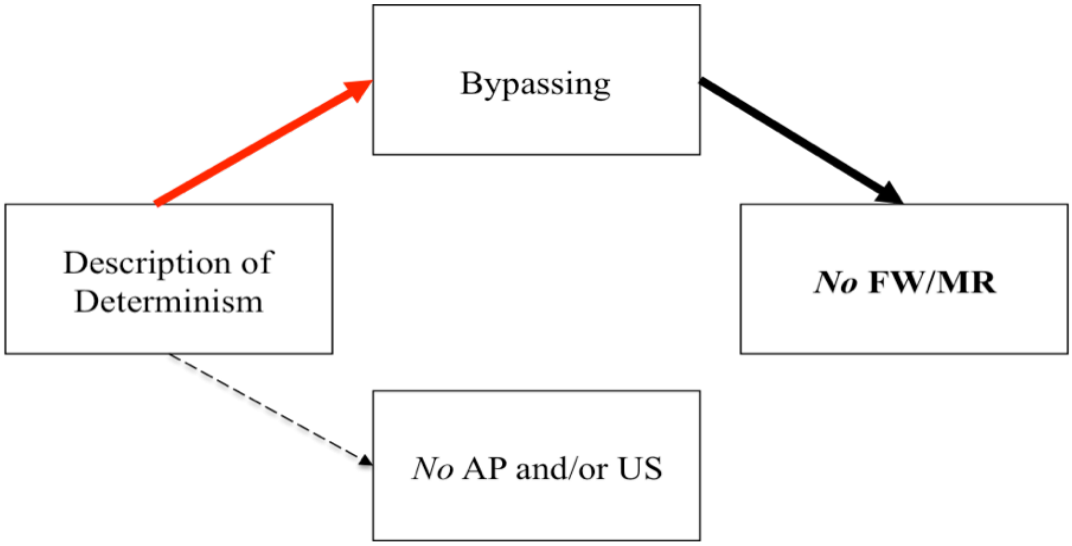

FIGURE 3: Judgments about MR, FW, and Bypassing

Percentage of "apparent incompatibilists" (participants with MR/FW composite scores < 3.5 midpoint, indicating disagreement on questions about MR, FW, and Blame) and percentage of "bypassers" (participants with Bypassing composite score > 3.5 midpoint, indicating agreement on questions about bypassing: Decisions, Wants, Believes, and No Control).

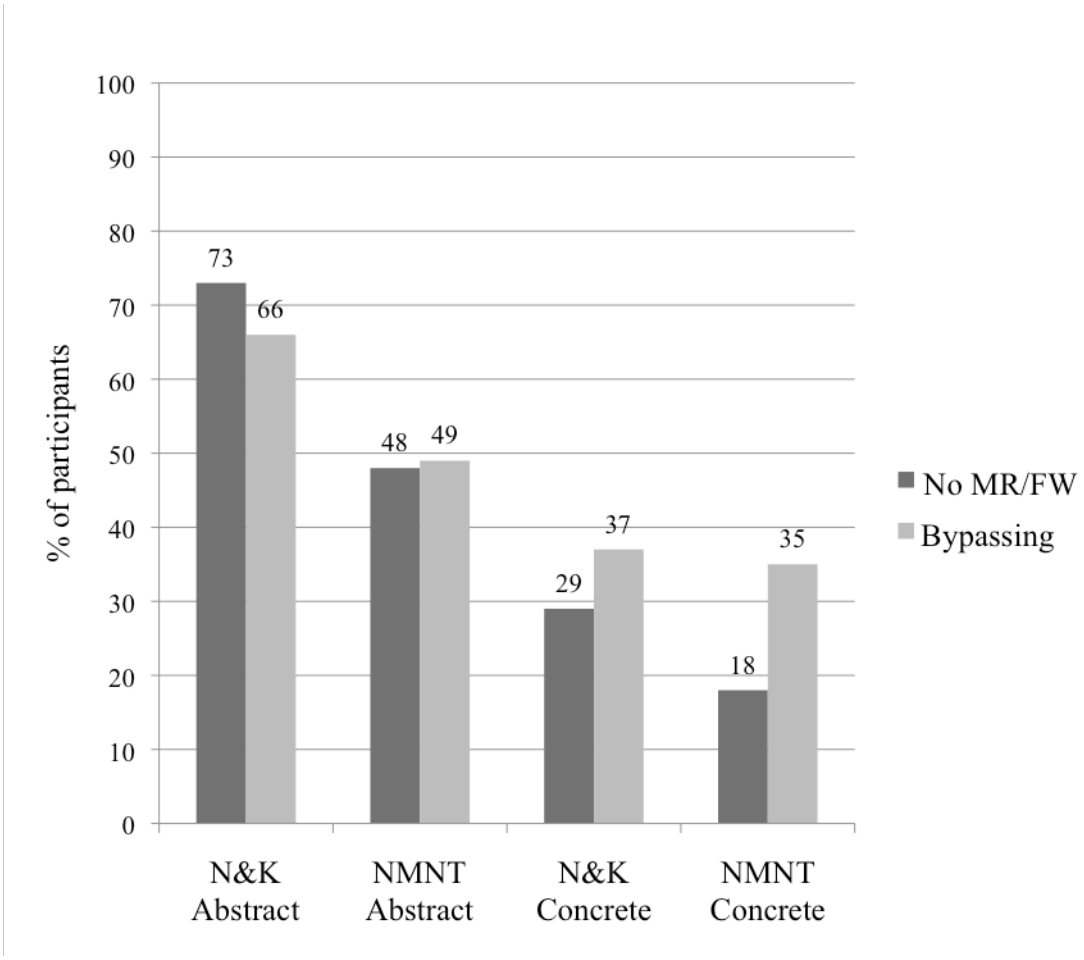

FIGURE 4: Apparent Incompatibilists who are Bypassers and Prima Facie Compatibilists who are not

Percentage of "apparent incompatibilists" who are "bypassers," and percentage of "prima facie compatibilists" (MR/FW composite score > 3.5 midpoint, indicating *agreement* on questions about MR, FW, and Blame) who are *not* "bypassers" (Bypassing composite score < 3.5 midpoint, indicating *disagreement* on questions about bypassing: Decisions, Wants, Believes, and No Control).

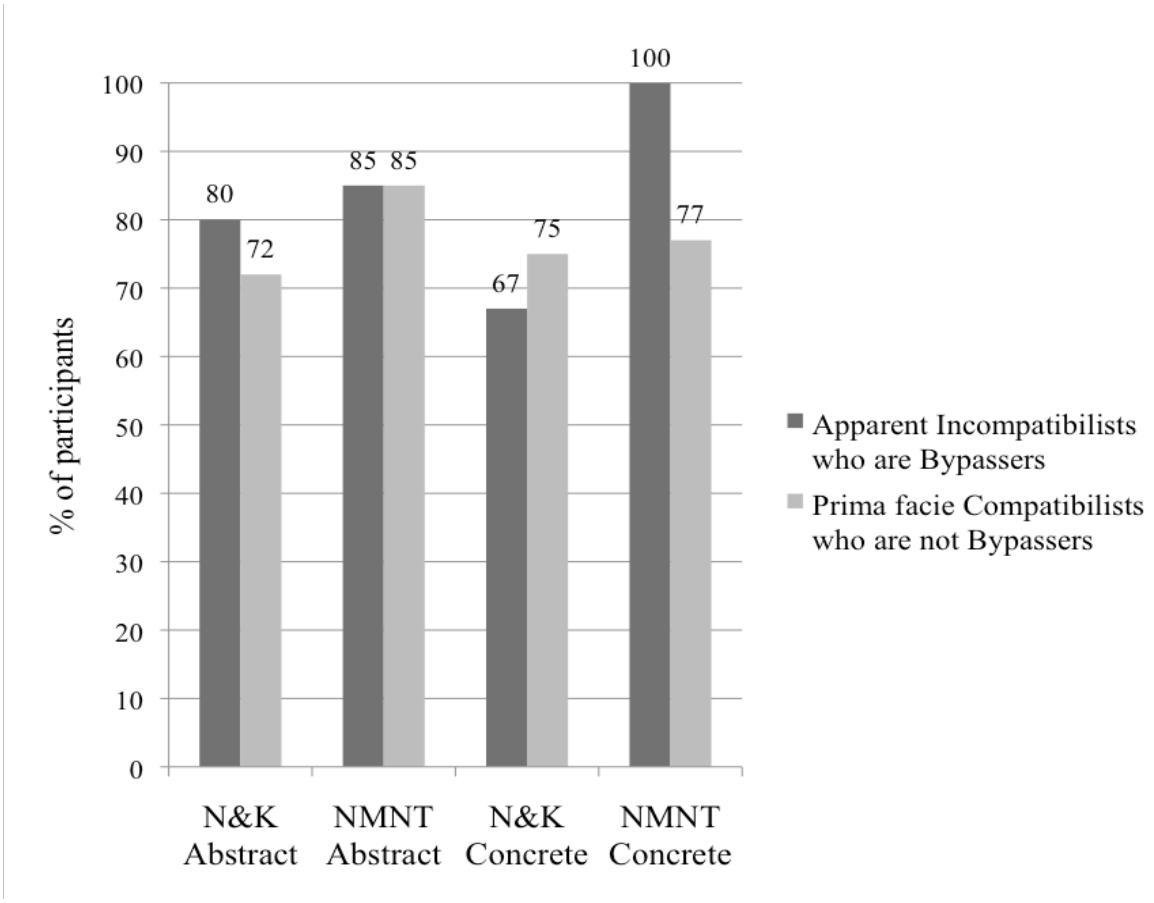

FIGURE 5: MR/FW and Bypassing Descriptive Statistics

Means and standard deviations for MR/FW and Bypassing composite scores for each scenario. Responses were given using a 6-point rating scale (midpoint is 3.5): strongly disagree = 1, disagree = 2, somewhat disagree = 3, somewhat agree = 4, agree =5, strongly agree = 6.

			MR/FW		Bypassing	
Survey	Condition	N	Mean	Std. Dev.	Mean	Std. Dev.
N&K	Abstract	77	2.818	1.204	3.958	1.205
	Concrete	56	4.363	1.298	3.018	1.216
NMNT	Abstract	56	3.482	1.360	3.442	1.346
	Concrete	60	4.444	1.174	2.946	1.183

FIGURE 6: Mediation Analysis

Standardized regression coefficients for the relationships between Survey (N&K, NMNT), MR/FW judgments, and Bypassing judgments. (Sobel test = 2.286, $p < .022$).

* $p = .022$, ** $p < .001$.