Incompatibilism and "Bypassed" Agency

Gunnar Björnsson, Umeå University, University of Gothenburg
Forthcoming in Al Mele (ed.) *Surrounding Free Will*, New York: OUP

1. Introduction

Both compatibilist and incompatibilist theories of moral responsibility are largely supported with reference to intuitions about cases. However, such intuitions vary among philosophers and laymen alike, and even people theoretically committed to compatibilism or incompatibilism can often feel the pull of intuitions in line with the opposite view. While our understanding of various arguments and of practices of holding responsible has made tremendous progress over the last few decades, it is fair to say that the basic disagreements over incompatibilism have remained.

One way to try to break this stalemate is to look not at the direct arguments for or against incompatibilism, but at the intuitions that seem to drive the debate. For example, if it could be shown, empirically, that pre-theoretical incompatibilist commitments are typically based on some clearly identifiable mistake, this might give us reason to doubt intuitions that flow from such commitments. (Similarly, of course, for compatibilist commitments.)

In earlier work, Karl Persson and I have argued that a certain independently supported general account of responsibility judgments gives us reason to disregard the basic intuitions grounding incompatibilist or skeptical convictions (Björnsson 2011, Björnsson and Persson 2009, 2012, 2013). According to this account, the *Explanation Hypothesis*, attributions of responsibility are implicit explanatory judgments, understanding the object of responsibility as straightforwardly explained by the agent's motivational structures. Incompatibilist intuitions arise from shifts in salient explanatory models, shifts that, we argue, are predictable but epistemically weightless side effects of mechanisms the function of which is to keep track of mundane relations between agents and outcomes.

A competing error theory for intuitions supporting incompatibilism has been proposed by Eddy Nahmias and Dylan Murray (N&M). Their proposal, the *Bypass Hypothesis*, is that when people take responsibility to be undermined by determinism, they do so because they take determinism to imply that the agent's beliefs, desires and decisions are *bypassed*, playing no role in bringing about or determining the agent's actions (Nahmias and Murray 2010; Murray and Nahmias 2012). This might seem like an improbable mistake, but the Bypass Hypothesis is bolstered by intriguing experimental data. Moreover, the attribution of error seems more straightforward than in the account provided by the Explanation Hypothesis, as N&M seem to have identified what is *obviously* a mistaken understanding of determinism. By contrast, the

Explanation Hypothesis only provides a credible error theory if it can be made plausible that judgments are illegitimate when based on certain explanatory interests and models.

The overall purpose of this paper is two-fold: to assess N&M's proposal and to see whether the Explanation Hypothesis is compatible with or capable of accounting for the relevant data. Sections 2 through 4 provide the background: a brief overview of some of the recent studies of folk intuitions about determinism and moral responsibility, an outline of how the Explanation Hypothesis accounts for some results from these studies, and a presentation of the experiments that seem to support the Bypass Hypothesis. In sections 5 through 9, I present a number of problems for the Bypass Hypothesis and alternative interpretations of the experimental data adduced in its support. I also argue that a variety of experimental studies by myself and others provide strong reason to reject the Bypass Hypothesis and accept the alternative interpretations, interpretations consonant with the Explanation Hypothesis.

2. The variety of compatibilist and incompatibilist intuitions

The last decade has seen numerous studies taking on the task of characterizing folk intuitions about responsibility and determinism. As is clear for anyone looking at these studies, the resulting picture is messy: intuitions vary interpersonally and depend in various ways on subtle variations in the questions asked and the ways determinism is presented. A study from Shaun Nichols and Joshua Knobe (2007) provides a useful example. (Though well known, I present it here in some detail, as most of the studies considered later build on the same paradigm.) Like several other studies, it has a straightforward format: subjects are presented with a deterministic scenario and are then asked whether an agent in that scenario is or could be morally responsible. In this case, subjects were introduced to a deterministic scenario characterized in terms of events being "completely caused" by prior events, such that the latter "have to happen" given the former. This scenario was contrasted with an indeterministic scenario (ibid. 669–70):

> 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.

After reading this vignette and indicating whether they think that the actual world is more like Universe A or Universe B (over 90% think the latter), subjects were asked whether they would attribute full moral responsibility to agents in Universe A. This question was asked in two quite different ways to different subjects: half the subjects were assigned the "concrete" question below, while the other half were assigned the "abstract" question:

CONCRETE CONDITION:

> 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

ABSTRACT CONDITION:

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

Only 14% of subjects in the abstract condition thought that it would be possible for an agent to be fully morally responsible in Universe A, while 72% in the concrete condition thought that Bill was fully morally responsible for his action. Judging from these results, different ways of asking about responsibility in deterministic scenarios can trigger contradictory intuitions.[1] For our purposes, this is interesting in several ways.

---

[1] Each condition had a little more than 40 subjects. Another group of subjects were in a concrete condition with less elaborate description of the action in 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?" Here, 50% answered "yes".

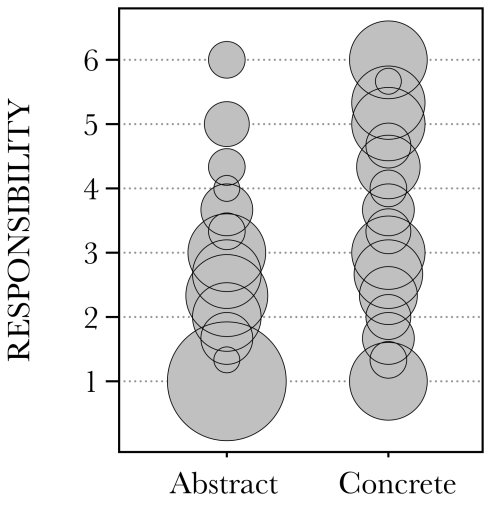
First, since a substantial majority of subjects gave incompatibilist answers to the abstract question, incompatibilist reactions seem to be grounded in a common, non-idiosyncratic, understanding of responsibility and determinism.

Second, since answers in the concrete condition seem to contradict those in the abstract condition, it is natural to assume that most judgments in one of these conditions are based on some sort of mistake: perhaps the concrete details in the former remind us of something required for responsibility, or obscure the deterministic character of the scenario or some important consequences of determinism.

Third, since incompatibilist reactions are substantially undermined when people are asked about concrete acts of wrongdoing, they are unlikely to rely on what is *front and center* in this common understanding of moral responsibility and determinism. Though pervasive, the mechanisms by which determinism undermines responsibility judgments seem to be relatively subtle.

The latter point is strengthened by variation in the extent to which subjects have been willing to attribute responsibility to agents in various studies: there is considerable variation in responsibility attributions depending both on the ways in which determinism is characterized in the relevant scenarios and on how the questions about responsibility are asked. In particular, descriptions of determinism in terms of how earlier events *cause* later events or make them *predictable* seem to undermine responsibility attributions to a much lesser extent than scenarios also stressing that prior events *necessitate* later events, as in the scenario above (e.g. Nahmias et al. 2006; Nahmias et al. 2007). The non-obviousness or non-centrality of assumptions underpinning incompatibilist reactions is also highlighted by considerable interpersonal variation in responsibility attributions. For example, in a study using the Nichols and Knobe (2007) vignettes and abstract / concrete conditions, I asked subjects to indicate their level of a agreement with a statement saying that in Universe A it is possible for Bill to be fully morally responsible for killing his wife and children (concrete conditions) or for a person to be fully morally responsible for their actions (abstract condition). Answers, which were given on a 1-to-6 scale ranging from *strongly disagree* to *strongly agree*, are represented in Figure 1, where the size of each bubble indicates the number of replies at that point, ranging from 1 to 21 (subjects (N=155) were recruited from Amazon Mechanical Turk). The difference between the abstract and concrete conditions is in line with the replies in

Figure 1: ABSTRACT, RESPONSIBILITY



Nichols' and Knobe's experiment (M=2.37 and 3.58, respectively), but a striking *spread* of answers is revealed in the diagram. This is not what one would expect if attributions of responsibility were straightforwardly governed by some easily applied compatibilist or incompatibilist rule.

What is clear is that some sort of explanation is needed that allows for significant interpersonal variation, and significant effects of seemingly irrelevant factors, such as the concreteness of the questions asked.

3. The Explanation Hypothesis

In earlier papers, partly in collaboration with Karl Persson, I have argued that a wide variety of phenomena involving judgments of responsibility, including those mentioned above, can be given a unified explanation if we understand such judgments as a species of explanatory judgment (see Björnsson 2011; Björnsson and Persson 2009; 2012). More specifically, responsibility judgments see the object of responsibility as explained (in normal ways) by some relevant "motivational structure" of the agent, i.e. a motivational structure of a kind that is generally an appropriate target for practices of holding responsible (for our purposes here, we might think of these as structures that are responsive to reasons). So when we think that an agent is morally to blame for an act or event, we think that it happened because the agent didn't care enough about morally important matters, or cared about the wrong things. Similarly, when we think that an act or event is to an agent's moral credit, we think that it happened because the agent balanced morally relevant concerns in a good way.

This might seem trite, but ordinary explanatory judgments are known to have a number of interesting features. Most importantly for our purposes, they are *selective*. If we are thinking about why some event E happened, we will focus only on one (or perhaps a few) events or conditions that were part of the causal prehistory of E, at the exclusion of others. If we are thinking about why a house burnt down, for example, we might focus on the fact that the house was hit by lightning, but not on the fact that the air contained oxygen, or on the fact that the house was built by combustible matter, or lacked a first-class sprinkler system. Though we understand that these other factors were necessary conditions for E, they are part of the explanatory background, as it were, typically because they are more generally expected and so less informative than the factors that we do focus on. Moreover, we naturally focus on factors that have a comparatively straightforward or familiar explanatory connection to E. Though we think that the lightning that hit the house had a causal prehistory—a separation of charges in the neighboring atmosphere, say—our focus will be on the lightning, as the lightning is causally related in a more straightforward and familiar way to the burnt down house than events leading up to the lightning.

Let us say that to focus on some factors as explaining E is to see these factors as the "significant" explanation of E. Then the following is our proposed account of responsibility judgments:

> THE EXPLANATION HYPOTHESIS: We take A to be responsible for X if we see some relevant motivational structure of A as (part of) a significant normal explanation of X.

Elsewhere we detail how the Explanation Hypothesis and the selectivity of explanatory judgments might account for a number of features of responsibility judgments, including the fact that responsibility judgments display so-called side-effect asymmetries and are closely statistically correlated with explicit explanatory judgments (Björnsson 2011; Björnsson and Persson 2012). Many of these features are relatively disconnected from issues of incompatibilism. But there is a further aspect of the selectivity of explanatory judgments that we suspect explains the seeming force of standard skeptical arguments about moral responsibility as well as the results recounted in the previous section: the selection of explanatory factors is relative to explanatory interests and salient explanatory models. Though you and I might ordinarily focus on the lightning when thinking about why the house burnt down, a fire engineer might instead focus on the lack of a lightning rod, treating the fact that the house was hit by lightning as part of the explanatory background. Similarly, a politician thinking about the same event might focus on inadequate funding for the fire brigade, and a physicist on specific properties of the building materials. Because of different explanatory interests, they might relegate different factors to the explanatory background, and employ explanatory models relating different variables. And because of this, they will think of different things as the *significant explanation* of the event.

Here is how the combination of this interest relativity of explanatory judgments and the Explanation Hypothesis might account for subjects' general but non-universal reluctance to attribute responsibility to agents in deterministic scenarios (Björnsson and Persson 2012; 2013):

First, people ordinarily attribute moral responsibility to agents on the basis of applying ordinary folk-psychological models, explaining actions and outcomes in terms of the beliefs and motivational structures of agents.

Second, what deterministic scenarios do is to introduce abstract deterministic explanatory models saying that every event is causally determined by earlier events (back to the beginning of the universe). In such models, human motivational structures, deliberation and decision-making play no privileged role, being mere causal intermediaries and providing no independent input into the general unfolding of events. Given the Explanation Hypothesis, someone looking at things from the perspective of this explanatory model will not see agents as responsible for their actions. This explains the tendency towards incompatibilist judgments.

Third, although deterministic scenarios introduce abstract explanatory models, folk-psychological models might nevertheless be more salient for particular subjects, especially since the latter are central parts of our everyday explanatory repertoire. This explains why the incompatibilist tendencies are limited.

Fourth, questions about responsibility asked about concrete cases are likely to activate folk-psychological models capable of explaining the specifics of such cases, at the expense of abstract deterministic models incapable of explaining any such particulars. This explains why subjects agree more with responsibility attributions in deterministic scenarios when these attributions concern concrete cases.

This explanation might itself be accepted by defenders of incompatibilism and compatibilism alike: it tells us that incompatibilist intuitions stem from a certain kind of explanatory perspective rather than another, but does not tell us which perspective is correct. While I think that it ultimately supports a comprehensive error theory for central incompatibilist intuitions (Björnsson and Persson 2012: 345–8; Björnsson ms), the argument needed for such a conclusion are complex and predictably contentious. In comparison, the Bypass Hypothesis offered by Nahmias and Murray is much more straightforward.

4. The Nahmias and Murray Bypass Hypothesis

N&M's hypothesis, recall, is that when subjects take responsibility to be undermined in deterministic scenarios, this is largely because they take agents' beliefs, desires and decisions to play no role in bringing about actions, i.e. because they take agents' deliberative or agential capacities to be *bypassed*. Some early evidence for this hypothesis came from studies by Nahmias, Coates and Kvaran (2007), where subjects were quite willing to attribute moral responsibility when deterministic causation of actions were described in psychological terms, but more reluctant when it was described in neurological terms. In the latter sort of scenario, but not in the former, it would be possible for subjects to conclude that ordinary psychological processes were bypassed.[2] The Bypass Hypothesis might also seem to explain why subjects in studies using the Nichols and Knobe paradigm, though prone to understand determinism as involving bypassing, would be less prone to make the mistake when considering a concrete case, and especially one describing the agent's motivation. After all, our ordinary understanding of such cases takes those to involve deliberative capacities.

Apart from having some initial plausibility, the Bypass Hypothesis is potentially highly significant. Since it is generally agreed that determinism does not imply that agents' beliefs,

[2] For my preferred explanation of this phenomenon, in terms of the Explanation Hypothesis, see Björnsson and Persson 2013: 626–32.)

desires and decisions are bypassed, it would be clear that incompatibilist folk intuitions are based on a mistake. Consequently, to the extent that pre-theoretical hunches and commitments account for stable intuitions and commitments among philosophers, incompatibilist theories of responsibility would also clearly rest on a mistake.

To more directly test the Bypass Hypothesis, N&M (Nahmias and Murray 2010; Murray and Nahmias 2012) conducted a survey where subjects were randomly assigned to one of four conditions: the two conditions of the Nichols and Knobe experiment and two further conditions in which subjects read descriptions of deterministic Universe C, descriptions that N&M hypothesized would be less likely to give rise to bypass misinterpretations of determinism, one involving an abstract description on human agency, and another involving an agent, Jill, who steals a necklace.[3] All in all, then, there were two abstract and two concrete conditions.

After having read one of the four vignettes, subjects were asked to indicate agreement with the statements below on a 1-to-6 scale (*strongly disagree*, *disagree*, *somewhat disagree*, *somewhat agree*, *agree*, *strongly agree*). The three first statements attribute free will, moral responsibility, or desert of blame; the latter four are meant to measure bypass judgments, saying that agents' beliefs, desires and decisions have no effect or that agents have no control over what they do. Subjects assigned to the abstract conditions read the first version of each statement; subjects assigned to concrete conditions read the version in parentheses:

RESPONSIBILITY

> MORAL RESPONSIBILITY: 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].)
>
> FREE WILL: 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].)

BYPASS

> 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.)

[3] The description of determinism was fetched from Nahmias, Morris, Nadelhoffer and Turner 2006.

> 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.)

Mean scores on both RESPONSIBILITY and BYPASS measures were calculated for each subject. Scores on different measures in each group were strongly internally consistent, with each statement contributing to that consistency, suggesting that each group of questions tracked one factor. The results were striking. First, there was a strong negative correlation between the RESPONSIBILITY and BYPASS mean scores.[4] Second, mediation analysis revealed that differences in responsibility scores between the two abstract conditions were largely predicted by differences in BYPASS score. This is exactly what you would expect if reluctance to attribute responsibility were largely explained by subjects' bypass interpretations of the deterministic scenarios.

To further test the bypass hypothesis, Nahmias and Murray conducted a second study, where they tried to directly manipulate bypass scores. They supplemented the deterministic scenarios with what we might call "throughpass"-statements, meant to explicitly rule out the bypass interpretations of determinism. The third paragraph of the Nichols & Knobe scenarios was modified as follows:

> The key difference, then, is that in Universe A every decision is completely caused by what happened before the decision. This does *not* mean that in Universe A people's mental states (their beliefs, desires, and decisions) have no effect on what they end up doing, and it does *not* mean that people are not part of the causal chains that lead to their actions. Rather, people's mental states *are* part of the causal chains that lead to their actions, though their mental states are always completely caused by earlier things in the causal chain that happened before them—*given* that the past happened the way it did, each decision *has to happen* the way it does. By contrast, in Universe B, decisions are

[4] $r(247) = -0.734$. Shepherd (2012) finds a similarly strong correlation using the same responsibility and bypass statements and slightly different scenarios.

> not completely caused by the past, and each human decision *does not have to happen* the way that it does given what happened in the past.

The other two scenarios were modified in a similar fashion. As predicted, this provided significantly lower BYPASS scores in the abstract conditions, but did not meaningfully affect the correlation between RESPONSIBILITY and BYPASS scores.[5] This seems to further strengthen the Bypass Hypothesis.

5. Some worries about the bypassing results

A number of worries can be raised about what is actually tested by the N&M bypass statements. One worry is that the last BYPASS statement employs the notion of "control", a notion that is notoriously contested in the debate about compatibilism and very closely linked to notions of responsibility: just as it is contested whether determinism rules out free will and moral responsibility, it is contested whether it leaves us with the control required for free will and responsiblity. In light of this, the NO CONTROL measure would seem to belong with the responsibility measures rather than the bypass measures. Whether this undermines N&M's results depends on whether it would change the relevant relation between responsibility and bypass scores. On the one hand, one might think that the inclusion of what many take to be a component of responsibility into the bypassing measure will illegitimately strengthen the correlation between the two measures. On the other hand, one might expect the removal of the NO CONTROL measure to make little difference to the N&M conclusion, as N&M report that the BYPASS scale would remain strongly internally consistent if NO CONTROL scores were removed (Nahmias and Murray 2010: 213, n. 16). To resolve this uncertainty, a study without this possible confound would be helpful.

Another question concerns the effect of the explicit throughpass statements added in N&M's second study: while responsibility scores were generally higher in this study than in the first, and bypass scores lower, the effect size was quite modest (Murray and Nahmias 2012, Appendix, Table 1). Given how prominently the throughpass statements figured in the vignettes, it is puzzling that many subjects would continue to misunderstand determinism in this way.

A third worry concerns what subjects have in mind when they agree that agents' beliefs, desires and decisions "have no effect" on what they do. Perhaps subjects understand the "no effect" statements as saying that beliefs, desires and decisions have no causal influence, direct or indirect, on actions. If so, subjects agreeing with these statements really do understand core

[5] $r(292) = 0.724$. In this second study, questions about the concrete cases were answered by subjects that had already answered the abstract questions, making a comparison between answers to the concrete cases problematic.

features or practical reasoning as being bypassed, thus indicating a misunderstanding of determinism. But another way of understanding talk about whether something has an effect on what happens is in terms of whether it provides some *independent input* into what happens.

This is obviously not how we should understand most talk of what has or does not have an effect on what happens. Still, such interpretations might be particularly salient to subjects who not only take determinism to imply that beliefs, desires and decisions provide no independent input in the relevant sense, but for whom this is a particularly striking fact. But subjects who take this to be a particularly salient fact are likely to understand events in the deterministic universe using the abstract explanatory model provided by the deterministic scenario, a model in which agents' motivational structures are at most intermediary variables. By the Explanation Hypothesis, those subjects will also take agents' responsibility to be undermined in that universe.

If this is right, we can straightforwardly account for the negative correlation between responsibility attributions and bypass judgments while assuming that subjects have a perfectly adequate understanding of determinism: those who (i) take responsibility to be undermined by determinism tend to be the same people who (ii) understand "no effect" statements as saying that beliefs, desires and decisions play no independent role in determining action, and then (iii) agree with these statements on the (controversial but nevertheless widely accepted) assumption that determinism implies that there is no independent agential input.

A neat feature of this alternative account of bypass judgments is that it can explain why the addition of explicit throughpass statements to the deterministic scenarios in N&M's second experiment had only a small effect on BYPASS and RESPONSIBILITY judgments. Since such statements explicitly mentioned the explanatory role of beliefs, desires and decisions, they increased the relative salience of folk-psychological explanatory models in which beliefs, desires and decisions figure as independent variables. By the proposed account of bypass judgments, this would decrease subjects' tendency to understand "no effect" statements as saying that beliefs, desires and decisions provide no independent input, and, by the Explanation Hypothesis, increase responsibility attributions. However, for subjects who are sufficiently taken by the deterministic explanatory model, this effect will be limited.

At this point, both the worries about the Bypass Hypothesis and the proposed alternative account of N&M's data based on the Explanation Hypothesis are of course speculative. More evidence is needed.

6. Experiment 1: BYPASS and THROUGHPASS

To resolve worries about the interpretation of N&M's results and to begin assessing the alternative hypothesis, I attempted a replication of N&M's first study, with some changes. First, I restricted my attention to the two Nichols and Knobe conditions (abstract and concrete), because

the difference in responsibility judgments between those conditions was particularly robust. Second, I removed the CONTROL statement from the BYPASS statements and modified DECISIONS, WANTS and BELIEVES to say that the agents decisions, or what they want or believe, has no effect on "what they do", as opposed to on "what they end up being caused to do". Finally, I added a question asking to what extent subjects agreed with an explicit throughpass statement:

> ABSTRACT THROUGHPASS: In Universe A, when earlier events cause an agent's action, they typically do so by affecting what the agent believes and wants, which in turn causes the agent to act in a certain way.
>
> CONCRETE THROUGHPASS: When earlier events caused Bill's action, they did so by affecting what he believed and wanted, which in turn caused him to act in a certain way.

If BYPASS statements are understood as intended by N&M, it seems that we should expect BYPASS and THROUGHPASS scores to be strongly negatively correlated, as the two THROUGHPASS statements explicitly assign a causal role to the agent's believes and desires. By contrast, no such negative correlation should be expected on the hypothesis that subjects interpret BYPASS statements as saying that beliefs, desires, or decisions have *no independent effect* on actions, as neither THROUGHPASS statement implies that the agent has such an independent effect. (If anything we might expect a positive correlation, as talk about actions being caused by earlier events might itself suggest a lack of independent effects.)

171 subjects were recruited through Amazon's Mechanical Turk and randomly assigned to either the concrete or the abstract condition where they answered RESPONSIBILITY and BYPASS questions presented in randomized order. 155 subjects passed a simple accuracy test and were included in further analysis.[6] Composite scores for RESPONSIBILITY and BYPASS were calculated, taking the mean of answers to each of the statements in the group.[7] The correlation between BYPASS and RESPONSIBILITY scores was roughly in line with those obtained by N&M: $r = -0.632$. (Figure 2 provides a graphical representation of the correlation, where size of bubble centered at a point indicates number of subjects at that point, ranging from 1 to 17.) Apparently, the correlation cannot be explained away with reference to the particular way that BYPASS statements had been formulated and the inclusion of a CONTROL statement.

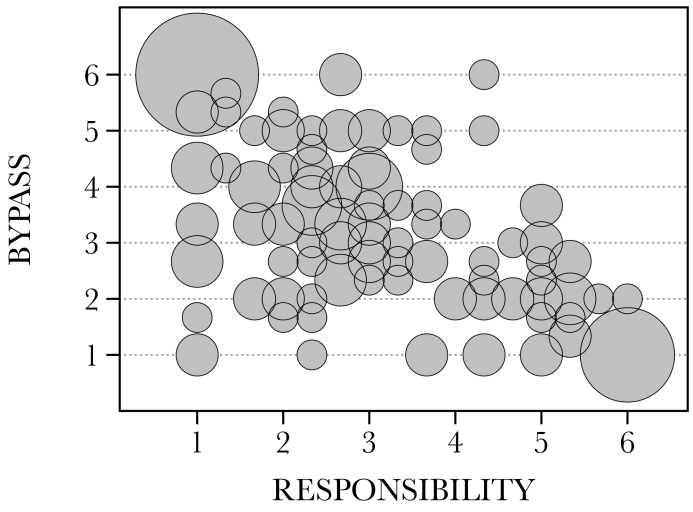
[6] Subjects were asked for agreement with the claim that in Universe B, decisions are not completely caused by what happens before (disagreement indicates inaccuracy).

[7] Cronbach's alfa > .853 for both measures.

Other results from the N&M study were not replicated, however. A mediation analysis treating ABSTRACT (the abstract / concrete variation) as the independent variable, RESPONSIBILITY as the dependent variable, and BYPASS as a mediator indicated a significant effect of ABSTRACT on RESPONSIBILITY mediated by BYPASS scores, but also a highly significant direct effect, accounting for 47% of the total effect (95% Confidence Interval: 14 to 69%).[8] Even on the assumption that BYPASS mediates the effect on RESPONSIBILITY to some degree or other, then, it seems that at least a substantial part of what explained intuitions of undermined responsibility in the abstract condition is independent of bypass interpretations of determinism.[9] At the very least, the Bypass Hypothesis does not seem to tell the full story.

Even more important, however, is the relation between scores on THROUGHPASS and BYPASS. THROUGHPASS scores were roughly what one might expect given an adequate understanding of Universe A, with 122 of 155 subjects answering *slightly agree*, *agree*, or *strongly agree*, with a mean well over midline (M = 4.38; CI(95%): 4.15 to 4.62). But rather than being strongly negatively correlated with BYPASS scores, as one would expect on the Bypass Hypothesis, these scores displayed a highly significant (albeit quite weak) *positive* correlation (r = .250, p = .002). Moreover, there was no meaningful correlation between THROUGHPASS and RESPONSIBILITY (r = -.043, p = .591). Notably, many subjects give quite high scores on both THROUGHPASS and BYPASS and a majority of those who gave the lowest RESPONSIBILITY scores gave the highest THROUGHPASS scores, as revealed in Figure 3 and 4. Pending reasons to think that subjects' agreement with THROUGHPASS statements should not be taken at face value, this strongly suggests that the

Figure 2: RESPONSIBILITY, BYPASS



[8] Percentile bootstrap confidence intervals calculated with Hayes PROCESS macro for SPSS (see Hayes 2013).

[9] Notably, my mediation analysis used the abstract / concrete variation of the N&K vignettes as independent variable, whereas N&M used the variation between two abstract conditions. This might account for the difference in outcome between the two studies, as the concrete action used in the N&K case involves an extreme moral transgression that might trigger emotional reasoning. (Below, we consider later studies, comparing determinism / indeterminism but using the new BYPASS statements; they too displayed a significant direct effect on RESPONSIBILITY, independent of BYPASS.)

Bypass Hypothesis is mistaken. Apparently, the reason that people do take responsibility to be undermined in these scenarios is not that they take beliefs, desires and decisions to lack causal influence on actions.

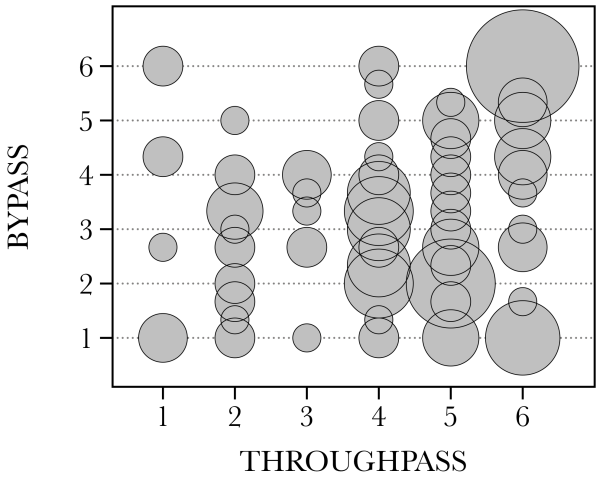


Figure 3: THROUGHPASS, BYPASS

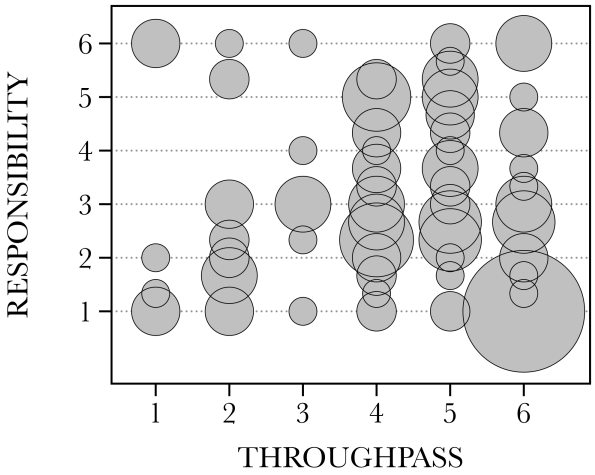


Figure 4: THROUGHPASS, RESPONSIBILITY

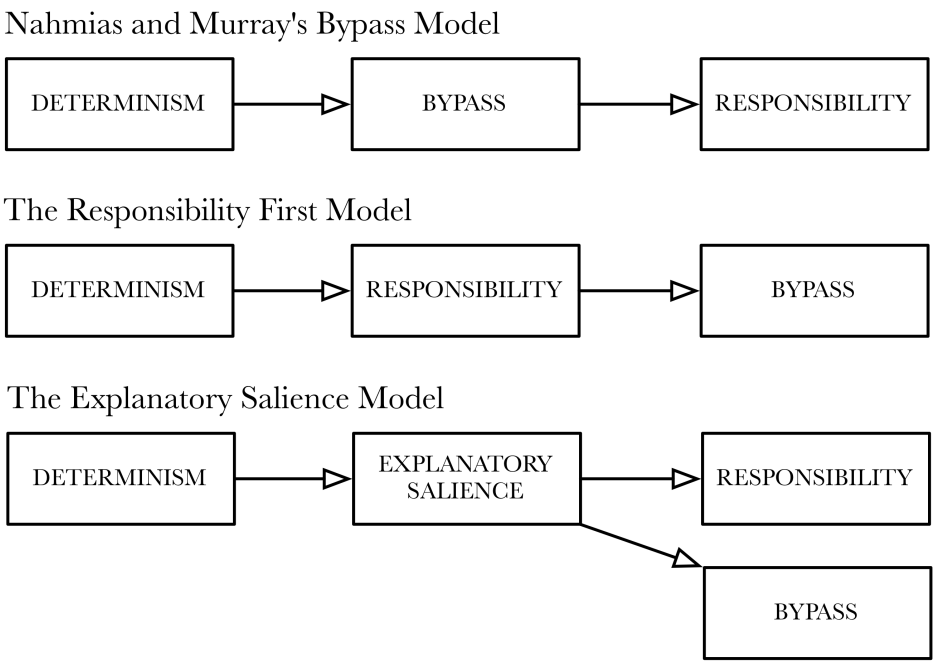
All this seems to fit with the alternative hypothesis sketched in the previous section: agreement with BYPASS statements is negatively correlated with RESPONSIBILITY attributions because subjects who take determinism to undermine responsibility are particularly likely to interpret "no effect" statements as saying that beliefs, desires or decisions provide no *independent* input into what happens. If this explanation of the correlation between BYPASS and RESPONSIBILITY is correct, a significant part of the total effect of the abstract / concrete conditions on BYPASS scores should be predicted by RESPONSIBILITY scores. This is indeed what we see: a mediation analysis treating RESPONSIBILITY as a possible mediator of the effect on BYPASS tells us that over 60% of the total effect was mediated by RESPONSIBILITY (CI(95%): 38 to 98%). Though we should want a replication of these results (this is one of the objectives of Experiment 2, reported below), the data from Experiment 1 suggest that the correlation between BYPASS and RESPONSIBILITY can be explained without reference to the Bypass Hypothesis.

7. Two more problems for the Bypass Hypothesis

While differences in RESPONSIBILITY scores between the two abstract conditions were significantly predicted by the differences in BYPASS scores in N&M's studies, prediction is not causation. Since RESPONSIBILITY and BYPASS scores were strongly correlated, it could well be that differences in BYPASS scores between the two abstract conditions are explained by differences in RESPONSIBILITY scores, rather than the other way around. What I have been proposing above represents a third alternative: the effect of ABSTRACT on both RESPONSIBILITY and BYPASS

judgments depends on whether one's most salient explanatory model represents the agent's motivational or deliberative structures as dependent or independent variables. The correlation itself is compatible with all these causal models. (See Figure 5: boxes symbolize variables and arrows indicate direction of causation.)

Figure 5: Three BYPASS, RESPONSIBILITY models



In a forthcoming paper, David Rose and Shaun Nichols set out to use causal modeling techniques to determine which of the two first models is correct. To this end, they conducted a version of the N&M bypass study: all subjects read the descriptions of Universes A and B and were then randomly assigned to either of two conditions, being asked to indicate levels of agreement with either RESPONSIBILITY and BYPASS statements concerning Universe A, or with corresponding statements concerning Universe B. Rose and Nichols found that on prominent ways of comparing statistical models, the Responsibility First Model fit the data much better than the Bypass Model.[10] Rose and Nichols' experiment thus seems to provide strong further reason to think that the Bypass Model gets the causal relation between RESPONSIBILITY and BYPASS scores wrong. (This comparison does not tell us how the responsibility first model compares to the Explanatory Salience model, as we have no direct measurement of the postulated EXPLANATORY SALIENCE variable.)

[10] E.g. p-values (probability of data given the model, higher scores better) for Responsibility First: p=.3421; for Bypass: p=.0013.

A second further problem for the Bypass Hypothesis is that while subjects tend to agree that in deterministic scenarios, beliefs, desires and decisions have no effect on action, they do not seem to think the same about ordinary causes of non-actions. This was first discovered in a study by Joshua Knobe (forthcoming), and replicated in a follow-up study by Rose and Nichols (forthcoming). In the latter study, subjects were asked to indicate level of agreement with one of the following statements about a Universe A type scenario:

> PRACTICAL REASONING: In this universe, when people make decisions, what they think and want has no effect on what actions they end up performing.
>
> THEORETICAL REASONING: In this universe, when people solve math problems the numbers they add has no effect on the answers they end up giving.
>
> PHYSICAL EVENT: In this universe, the earth's shaking has no effect on whether trees fall over.

Agreement with PRACTICAL REASONING was significantly higher than agreement with THEORETICAL REASONING, which was significantly higher than agreement with PHYSICAL EVENT. Together, the studies by Knobe and by Rose and Nichols strongly suggest that what leads some subjects to make bypassing judgments is their way of understanding human decision making or the relation between such decision making and determinism, not a general misunderstanding of determinism. The question, though, is what it is about the understanding of human reasoning that prompts bypassing judgments.

8. Why varying bypass judgments?

Trying to explain why bypass judgments are restricted to human agency in particular, Rose and Nichols suggests that subjects

> … tend to think of decisions as fundamentally indeterminist such that if determinism is true, people really don't make decisions. If that's right, the bypassing questions might provide people with a way of expressing their view that decisions don't occur under determinism. (Rose and Nichols *forthcoming*, §4)

To test their suggestion, they asked the subjects of the study recounted in the previous section one further question, corresponding to the bypass question they were asked (answers in parentheses):

> PRACTICAL REASONING: In this universe, people make decisions. (YES: 53%; NO: 47%)
>
> THEORETICAL REASONING: In this universe, people add numbers. (YES: 88%; NO: 12%)

PHYSICAL EVENT: In this universe, trees fall over. (YES: 100%; NO: 0%)

Interestingly, many subjects were reluctant to say that people in a deterministic universe make decisions, and some reluctant to say that they add numbers. Mediation analysis and comparison of causal models also suggested that the effect of kinds of reasoning on bypass judgments was mediated by its effect on reluctance to attribute decisions or adding. Judging from these results, it seems that when subjects deny that what an agent thinks, wants or decides has any effect on what she ends up doing, they do so because they think that the agent neither thinks, or wants, nor decides.

The connection might seem antecedently very plausible: if one thinks that no one makes decisions, say, one might naturally agree with the claim that decisions have no effect on what people do. Moreover, it might seem plausible that subjects understand decisions to involve the exercise of free will, and thus that subjects who take determinism to undermine free will also take it to undermine decisions. But acceptance of BELIEVES and WANTS statements were even more strongly negatively correlated with responsibility scores than were acceptance of the DECISIONS statement, and it seems much less natural to think that determinism or lack of free will undermines the existence of desires or beliefs. Moreover, the explanation seems to conflict with subjects' agreement with THROUGHPASS statements in Experiment 1. Since such statements explicitly postulate the existence of beliefs and desires, one would think that subjects keen to express the thought that people do not really believe or want things would reject such statements inasmuch as they would accept the corresponding "no effect" statements. However, since there was no such correlation, it seems prima facie unlikely that Rose and Nichols' explanation generalizes. At the very least, we need to further explore the relation between non-existence and bypass judgments, focusing not only on decisions, as well as the relation between responsibility judgments and judgments of non-existence.

It would also be good to compare the non-existence explanation of BYPASS judgments with an explanation building on our working hypothesis about the negative correlation between BYPASS and RESPONSIBILITY judgments. The explanation, recall, was the following: Subjects generally conceive of human deliberation as providing independent input into causal systems—i.e. relying on explanatory models in which the agent's decision, beliefs or desires are independent variables. Deterministic scenarios introduce abstract explanatory models in which aspects of human agency are seen as dependent or intermediary variables. Subjects who are particularly taken by this model when introduced to a Universe A type scenario will tend to (a) take responsibility to be undermined (given the Explanation Hypothesis) and (b) interpret "no effect" statements as saying that aspects of human deliberation have *no independent* effect on human action (rather than saying

more strongly that they play *no* causal role in producing action), and thus tend to accept those statements. Hence the negative BYPASS-RESPONSIBILITY correlation.

Suppose that this general explanation of BYPASS judgments is correct. Then we should expect the BYPASS judgments to primarily concern phenomena that subjects antecedently expect to provide independent causal input: it is with respect to those phenomena that the deterministic explanatory model represents surprising explanatory relations, and so is likely to grab hold of the subjects' attention. This condition is satisfied for the case of human agency, but less so for theoretical reasoning, and much less so for non-agential events like trees falling over as a result of earthquakes.[11] Hence the restriction of BYPASS judgments to human agency.

This proposed explanation suggests a further prediction. We might expect people who do not think that human deliberation provides independent causal input—people who think that determinism is true—to find deterministic causation of action less out of the ordinary. On the current proposal, they would thus be less likely to be in the grip of abstract deterministic causal models, and so less likely to see responsibility as undermined in deterministic scenarios, and less likely to make bypass judgments about agency in such scenarios.

The data from Experiment 1 seem to fall in line with this prediction. In that experiment, subjects had been asked whether they think that our world is most like Universe A or B, and subjects answering "A" did indeed provide higher RESPONSIBILITY and lower BYPASS scores (RESPONSIBILITY M = 3.95 (A) vs. 2.92 (B) and BYPASS: M = 2.83 (A) vs. 3.42 (B)).[12] Unfortunately, the number of A-subjects was very small (14 of 155 subjects) and only the variation of RESPONSIBILITY scores was found to be significant in a one-way analysis of variance (ANOVA) ($F = 5.64$; $p = .019$ for RESP. vs. $F = 1.74$; $p = .189$ for BYPASS). A better assessment of this prediction requires further studies.

9. Experiment 2: responsibility, bypass, and non-existence

Experiment 2 had three purposes. One was to further test the prediction that subjects who already take human deliberation to be determined will be less prone to make bypass judgments about deterministic scenarios. This prediction would be most directly tested in an experiment assigning subjects to deterministic and indeterministic conditions, as in the Rose and Nichols

[11] Recall that 90% of subjects in Nichols and Knobe's (2007) study thought that our universe is more like Universe B, where human decision-making is not determined by prior events. Similarly, a cross-cultural study involving subjects from India, Columbia, Hong Kong and the United States found that between 68% and 85% of subjects (university students) thought that our universe was more like Universe B (Sarkissian et al. 2010). In line with this, the experiments of Deery et al. 2013 indicate that subjects tend to take the phenomenology of free choice to be incompatible with determinism.

[12] Similar results were found in a study by Shepherd (2012: 922).

experiment recounted in section 7, rather than to abstract and concrete conditions, or different abstract conditions. A second purpose was to explore the relation between responsibility and non-existence judgments, and attempt to replicate the effects of determinism on judgments of non-existence. The lack of negative correlation between THROUGHPASS and BYPASS judgments already suggests that non-existence judgments cannot do the explanatory work required by Rose and Nichols, but it would be helpful to ask a more straightforward existence question involving not only decisions, but also beliefs. The third purpose was to see whether the independence of THROUGHPASS scores from BYPASS and RESPONSIBILITY scores discovered in Experiment 1 would hold up with a Universe A/B (determinist/indeterminist) variation rather than abstract/concrete variation as the independent variable.

122 subjects were recruited through Amazon's Mechanical Turk. After reading descriptions of Universe A and B and being asked which universe they thought was most like ours, they were randomly assigned to either of two conditions, being asked to indicate levels of agreement either with RESPONSIBILITY and BYPASS statements concerning Universe A, or with corresponding statements concerning Universe B; the statements were the same as those used for the abstract condition in Experiment 1, with the added A/B variation. In addition, subjects were asked to answer the following question about the existence of deliberation in their assigned universe:

> DELIBERATION: In Universe A (Universe B), does it happen that people believe things about their situation and make decisions based on these beliefs?

Unlike THROUGHPASS statements used in Experiment 1, this question asks whether agents make (certain kinds of) decisions, and unlike Rose and Nichols' existence question, this question concerns not only decisions, but also beliefs. Finally, I added reference to decisions in the explicit THROUGHPASS statement:

> THROUGHPASS: In Universe A (Universe B), when earlier events cause an agent's action, they typically do so by affecting what the agent believes and wants, which in turn causes the agent to decide and act in a certain way.

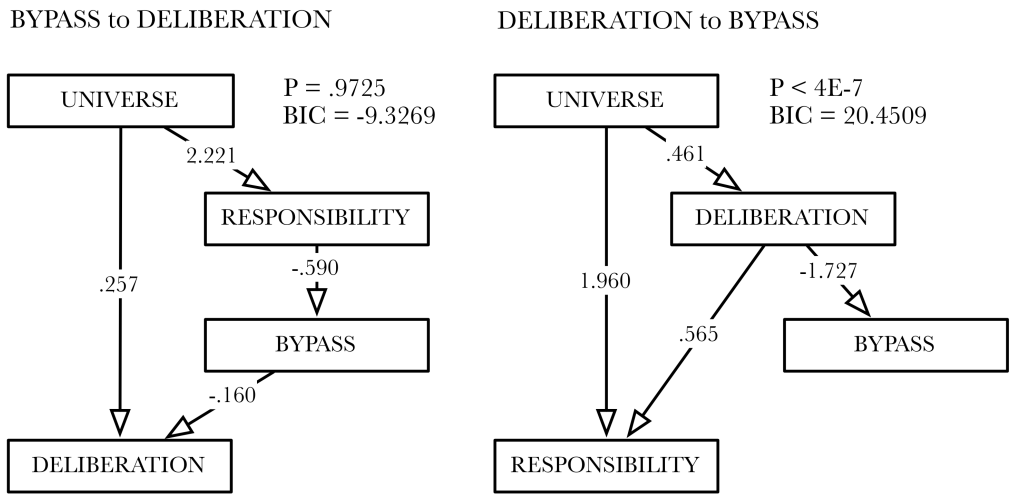
After purging the data set from answers from subjects who failed the accuracy test, analysis of data was based on answers from 109 subjects.

The correlation between RESPONSIBILITY and BYPASS was almost exactly as strong as in Experiment 1 ($r = -.622$, $p < .000$). There was no significant correlation between THROUGHPASS and BYPASS ($r = -.038$, $p = .693$), and a weak but highly significant *negative* correlation between THROUGHPASS and RESPONSIBILITY ($r = -.269$, $p = .005$). A mediation analysis treating UNIVERSE (i.e. whether statements concerned Universe A or B) as independent variable, RESPONSIBILITY as dependent variable, and BYPASS as a proposed mediator indicated that 77% of

total effect of independent on dependent variable was direct (CI(95%): 59 to 90%), suggesting that bypass interpretations play at most a partial role in explaining incompatibilist intuitions. By contrast, there was virtually no direct effect of UNIVERSE on BYPASS in a model treating RESPONSIBILITY as mediator.[13] In line with this, the Responsibility First model fit the data *much* better than the Bypass model, which did not fit the data at all.[14] All this provides extraordinarily strong support for our earlier conclusion: subjects' tendencies to withhold responsibility attributions to agents in deterministic scenarios do not stem from tendencies to understand determinism as implying bypassed agency. Instead, bypass intuitions are explained by intuitions of undermined responsibility, or, as I have suggested, by a condition closely associated with those intuitions.

The continued lack of negative correlation between THROUGHPASS and BYPASS provided some further evidence against Rose and Nichols' non-existence hypothesis. To more directly determine what role attributions of deliberation play in responsibility and bypass judgments, I compared a wide variety of causal models of the relation between UNIVERSE and the dependent variables with respect to BIC scores, one commonly used measure for model choice.[15] Figure 6

Figure 6: Two RESPONSIBILITY, BYPASS, DELIBERATION models



Means for Universe A:
UNIVERSE: 0 (A=0, B=1); DELIB.: .8542 (Yes=1, No=0); RESP.: 4.806; BYPASS: 2.681

[13] Total effect of UNIVERSE on BYPASS = 1.2757; CI(95%): .7734 to 1.7780. Direct effect of UNIVERSE on BYPASS = -.0735; CI(95%): -.7090 to .5620. Indirect effects of UNIVERSE on BYPASS through RESPONSIBILITY = 1.3492; CI(95%): .8650 to 1.8652.

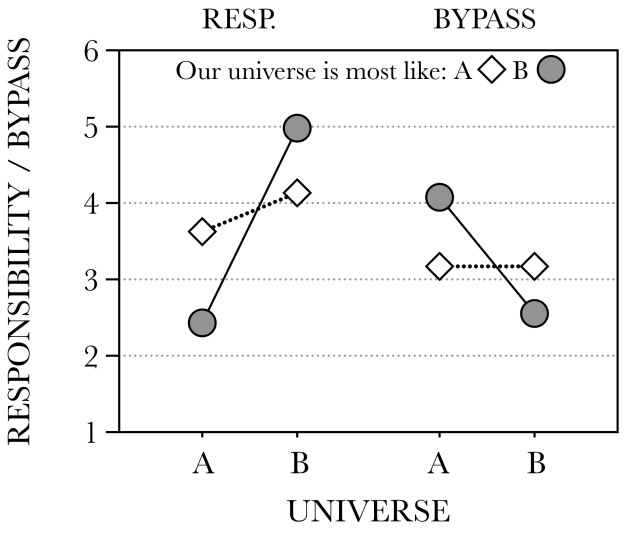
[14] Responsibility First: $p = .8170$; Bypass: $p < 5 \times 10^{-14}$.

[15] Lower score is better. For the motivation behind BIC ("Bayesian Information Criterion"), see e.g. Wagenmakers 2007.

displays both the best scoring model, *BYPASS to DELIBERATION*, and the best model in which the effect of UNIVERSE on BYPASS is entirely mediated by DELIBERATION, *DELIBERATION to BYPASS* (coefficients on arrows indicate what effect a one unit change in value of one variable has on the "downstream" variable). While *BYPASS to DELIBERATION* was a very good fit with data, *DELIBERATION to BYPASS* did not fit at all. This strongly suggests that subjects do not make BYPASS judgments because they think that there is no deliberation in a deterministic universe.[16]

Finally, to test whether subjects' prior beliefs in determinism would affect BYPASS and RESPONSIBILITY judgments, as predicted by the Explanation Hypothesis, the interaction between such beliefs (the "BELIEF" variable) and the UNIVERSE condition was explored using two-way ANOVAs.[17] For BYPASS, I found the expected significant effect of UNIVERSE on BYPASS ($F = 5.13$; $p = .026$; partial eta squared = .047), no significant effect of BELIEF ($F = .192$; $p = .662$; partial eta squared = .002), but a significant interaction effect, as predicted ($F = 5.13$; $p = .026$; partial eta squared = .047). For RESPONSIBILITY, I found the expected highly significant effect of UNIVERSE ($F = 34.55$; $p = .000$; partial eta squared = .248), no significant effect of BELIEF ($F = .446$; $p = .506$; partial eta squared = .004), but the predicted significant interaction effect ($F = 15{,}420$; $p = .000$; partial eta squared = .128). (See Figure 7.) Judging from this, BELIEF makes a significant difference to the effect of determinism on both BYPASS and RESPONSIBILITY. All this seems to support the account provided by the Explanation Hypothesis: since those who find deterministic causation of agency out of the ordinary will more likely be in the grip of the abstract deterministic explanatory model, they are more likely both to take responsibility to be undermined, and to interpret "no effect" statements as saying

Figure 7: BELIEF, UNIVERSE interaction



[16] This conclusion assumes that I have considered the best model in line with the non-existence hypothesis. To be sure not to miss the best models, I used two algorithms for model search in Tetrad IV, HBSMS and GES. For Tetrad, see http://www.phil.cmu.edu/projects/tetrad. For the principles behind GES, see Chickering 2002.

For further confirmation of this negative result, see n. 18.

[17] Since the number of A-subjects was again low, not all standard assumptions of ANOVAs are satisfied, and the numbers should be taken as suggestive rather than probative.

that agential states provide no independent causal input into the actions performed.[18]

10. Concluding remarks

If correct, the Bypass Hypothesis would provide a powerful error theory for incompatibilist intuitions among lay people, potentially also undermining the credibility of philosophers' incompatibilist intuitions. But while Nahmias and Murray's studies were suggestive, other experiments strongly indicate both that (i) subject's disagreement with RESPONSIBILITY statements are not explained by their acceptance of BYPASS statements, and that (ii) subjects do not interpret the BYPASS statements in the way intended. We have also seen strong experimental reasons to reject Rose and Nichols non-existence hypothesis: subjects do not seem to make BYPASS judgments because they take determinism to rule out the existence of beliefs, desires, and decisions.

More constructively, I have suggested that the negative correlation between BYPASS and RESPONSIBILITY judgments might be explained given the independently motivated Explanation

---

[18] To deal with three minor lingering worries, I conducted another study on the pattern of Experiment 2, with two minor changes (N=136 after 9 subjects had been removed for failing accuracy test). First, the following two statements were substituted for DELIBERATION to see whether simpler existence statements like those used by Rose and Nichols might trigger the sort of judgments responsible for the results in their study:

DECIDES: In Universe A (B), people make decisions.

BELIEVES: In Universe A (B), people believe things about their situation.

Answers were given on a 6-point Likert scale. The results were essentially the same as in Experiment 2: compared to the highest-scoring model, BIC scores were much worse (≈30 points higher) for the best model where BYPASS was entirely mediated by either DECIDES, BELIEVES, both DECIDES and BELIEVES, or the mean of the two. Moreover, the correlation between UNIVERSE and BELIEVES was weak and barely significant ($r = .171$, $p = .047$).

Second, I rephrased THROUGHPASS to make it explicit that decisions were not bypassed in the causation of action:

> THROUGHPASS: In Universe A, when earlier events cause an agent's action, they typically do so by affecting what the agent believes and wants, which affects what the agent decides to do, which in turn determines how the agent acts.

Again, this reformulation made no meaningful difference: there was still no significant correlation between THROUGHPASS and BYPASS scores.

Finally, Experiment 3 replicated the role of BELIEF as a moderator of the effect of UNIVERSE on BYPASS and RESPONSIBILITY. Collapsing the results from both studies (N=245), the interaction effect for BYPASS was highly significant ($p = .000$, partial eta squared = .063), as was that for RESPONSIBILITY ($p = .000$, partial eta squared = .096).

Hypothesis and the assumption that subjects who take responsibility to be undermined also interpret bypass statements in a certain non-literal way. This suggestion found support not only in the failure of alternative hypotheses, but also in the lack of correlation between BYPASS and THROUGHPASS judgments, and in the interaction between beliefs in determinism and BYPASS and RESPONSIBILITY judgments. If the proposed explanation is correct, it might still support a compatibilist error theory for incompatibilist intuitions (as I argue elsewhere), but the mistake involved will be much more subtle than that of taking determinism to imply bypassed agency.[19] Incompatibilism is probably a mistake, but not that simple a mistake.

References

Björnsson, Gunnar ms: 'Illusions of Undermined Responsisibility', manuscript

Björnsson, Gunnar 2011: 'Joint Responsibility Without Individual Control: Applying the Explanation Hypothesis'. In *Compatibilist Responsibility: Beyond Free Will and Determinism*. van den Hoven, Jeroen, van de Poel, Ibo and Vincent, Nicole (eds.), Springer, pp. 181–99.

Björnsson, Gunnar and Persson, Karl 2013: 'A Unified Empirical Account of Responsibility Judgments'. *Philosophy and Phenomenological Research*, 87, pp. 611-39.

Björnsson, Gunnar and Persson, Karl 2012: 'The Explanatory Component of Moral Responsibility'. *Nous*, 46, pp. 326–54.

Björnsson, Gunnar and Persson, Karl 2009: 'Judgments of Moral Responsibility: A Unified Account'. *Society for Philosophy and Psychology*, 35th Annual Meeting, Bloomington, IN, http://philsci-archive.pitt.edu/4633/.

Chickering, David Maxwell 2002: 'Optimal Structure Identification With Greedy Search'. *Journal of Machine Learning Research*, 3, pp. 507-54.

Deery, Oisín, Bedke, Matthew S. and Nichols, Shaun 2013: 'Phenomenal Abilities: Incompatibilism and the Experience of Agency'. In *Oxford Studies in Agency and Responsibility*. Shoemaker, David (ed.) Oxford University Press, pp. 126-50.

Hayes, Andrew F 2013: *Introduction to Mediation, Moderation, and Conditional Process Analysis: A regression-Based Approach*. New York: Guilford Press.

Knobe, Joshua, forthcoming, "Free Will and Scientific Vision" In *Current Controversies in Experimental Philosophy,* Edouard Machery & Elizabeth O.'Neill (eds.), Routledge.

[19] Work on this chapter was supported by grants from the John Templeton Foundation and Riksbankens Jubileumsfond. Their views are not necessarily reflected by the opinions express in this chapter. Many thanks to Josh Knobe, David Rose, Eddy Nahmias, Al Mele and Stefano Cossara for comments on a previous version.

Mele, Al, forthcoming, "Free Will and Substance Dualism: The Real Scientific Threat to Free Will?"

Murray, Dylan and Nahmias, Eddy 2012: 'Explaining Away Incompatibilist Intuitions'. *Philosophy and Phenomenological Research*, http://dx.doi.org/10.1111/j.1933-1592.2012.00609.x

Nahmias, Eddy, Coates, D. Justin and Kvaran, Trevor 2007: 'Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions'. *Midwest Studies in Philosophy*, 31, pp. 214–42.

Nahmias, Eddy, Morris, Stephen G., Nadelhoffer, Thomas and Turner, Jason 2006: 'Is Incompatibilism Intuitive?' *Philosophy and Phenomenological Research*, 73, pp. 28–53.

Nahmias, Eddy and Murray, Dylan 2010: 'Experimental Philosophy On Free Will: An Error Theory For Incompatibilist Intuitions'. In *New Waves in Philosophy of Action*. Aguilar, Jesús, Buckareff, Andrei and Frankish, Keith (eds.), Palgrave Macmillan pp. 189-216.

Nichols, Shaun and Knobe, Joshua 2007: 'Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions'. *Noûs*, 41, pp. 663-85.

Rose, David and Nichols, Shaun, forthcoming, 'The Lesson of Bypassing'. In *Review of Philosophy and Psychology*.

Sarkissian, Hagop, Chatterjee, Amita, Brigard, Felipe De, Knobe, Joshua, Nichols, Shaun and Sirker, Smita 2010: 'Is Belief in Free Will a Cultural Universal?'. Mind & Language, 25, pp. 346-58.

Shepherd, Joshua 2012: 'Free Will and Consciousness: Experimental Studies'. *Consciousness and Cognition*, 21, pp. 915–27.

Wagenmakers, Eric-Jan 2007: 'A Practical Solution to The Pervasive Problems of P Values'. *Psychonomic Bulletin & Review*, 14, pp. 779-804.