Journal of
Cognition
and
Neuroethics

The Importance of Correctly Explaining Intuitions: Why Pereboom's Four-Case Manipulation Argument is Manipulative

Jay Spitzley
Georgia State University

Biography

Jay Spitzley is pursuing a Masters of Arts degree in philosophy at Georgia State University. His research interests include moral psychology, action theory, experimental philosophy, and neuroethics. He received his BA at the University of Michigan and can be contacted at jspitzley1@student.gsu.edu.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation

Spitzley, Jay. 2015. "The Importance of Correctly Explaining Intuitions: Why Pereboom's Four-Case Manipulation Argument is Manipulative." *Journal of Cognition and Neuroethics* 3 (1): 363–382.

The Importance of Correctly Explaining Intuitions: Why Pereboom's Four-Case Manipulation Argument is Manipulative

Jay Spitzley

Abstract

Recent empirical findings have shown that intuitions are significantly influenced by subtle and seemingly irrelevant factors. In light of these findings, I argue that before making claims about what best explains intuitions regarding thought experiments, one must acknowledge the effects that certain psychological influences have on intuitions. To demonstrate how problematic it can be to ignore these covert factors, I discuss Derk Pereboom's four-case manipulation argument. While Pereboom claims that intuitions regarding his argument for incompatibilism reliably track relevant features of the four cases, I argue instead that these intuitions are likely driven by order effects motivated by unconscious psychological influences and that these order effects put significant pressure on Pereboom's argument.

Keywords

Intuition, moral judgment, order effects, moral responsibility, Derk Pereboom, manipulation

I. Introduction

It has become common in philosophy to use intuitions about thought experiments and hypothetical cases to bolster one's argument. While we like to think intuitions are reliable, recent empirical research shows that this isn't always the case. Experimental philosophers who contribute to the "negative" program of experimental moral philosophy have discovered that intuitions are not universally held.[1] Rather, judgments vary according to ethnicity (Weinberg et al. 2001), gender (Buckwalter and Stich 2011), and linguistic background (Vaesen et al. 2013). Further research shows that intuitions are unreliable in an additional sense. That is, intuitions and moral judgments are significantly influenced by trivial and rationally irrelevant factors of hypothetical cases, such as the order in which information is presented (Weigmann et al. 2012; Schwitzgebel and Cushman 2012), the way in which the information is worded (Petrinovich and O'Neill 1996), the emotional

[1] Experimental philosophy's "negative program," generally seeks to challenge the usefulness in appealing to philosophical intuitions as a method of uncovering justified beliefs (Alexander, Mallon, and Weinberg 2014).

status of the reader (King and Hicks 2011; He et al. 2013; Guiseppe et al. 2012), and even the clean smell of Lysol (Tobia et al. 2013).

While the variation of intuitions across demographics has led some to argue that intuitions about these hypothetical cases should not be used as evidence for philosophical views (Weinberg 2008; Sinnott-Armstrong 2008), I will focus on problems that result from our intuitions being unreliable in the other sense. Specifically, I address problems that arise from features and psychological influences we are largely unaware of driving our intuitions and moral judgments. Given the sway such factors have on intuitions about hypothetical cases and thought experiments, I argue one must proceed cautiously when presenting an argument that relies on an explanation for what features of a case motivate intuitions about that case. Furthermore, I argue that failure to consider these psychological influences (some of which may be entirely unconscious) as alternative explanations for what drives intuitions can undermine one's argument.

Despite overwhelming evidence that humans are bad at knowing what influences their judgments (King and Hicks 2011; Mlodinow 2012; Li et al. 2008), and that even philosophers are susceptible to unconscious psychological influences (Schwitzgebel and Cushman 2012; Tobia 2013), philosophers frequently assume they know what drives intuitions. To demonstrate the importance of taking this new evidence into account for philosophical debate, I discuss Derk Pereboom's (2014) four-case manipulation argument. The success of his argument hinges on knowing what motivates intuitions about the four cases he presents the reader. I argue that by neglecting to consider an alternative explanation for what drives intuitions, namely, order effects, Derk Pereboom leaves open a serious objection to his argument.

II. Pereboom's Four-Case Manipulation Argument

In an attempt to demonstrate that the compatibilist conditions for moral responsibility are insufficient on the grounds that determinism, when properly understood, is incompatible with moral responsibility, Derk Pereboom (2014) presents a manipulation argument. Pereboom attempts to show that even in cases when all compatibilist requirements for free will and moral responsibility are met, agents can still lack moral responsibility. To achieve these aims, Pereboom presents four cases.

Each case involves an agent, Plum, who is causally determined by factors beyond his control to kill another agent, White. Additionally, in each case Plum satisfies all purported compatibilist requirements for free will and moral responsibility.[2] In Case 1, Plum's mental states are manipulated via radio-like technology by a team of neuroscientists in such a way that he reasons egoistically and decides to kill White. In Case 2, Plum is just like an ordinary human being except that neuroscientists manipulate him in the beginning of his life in such a way that he will later reason egoistically and kill White. In Case 3, the training practices of Plum's community causally determine that he reasons egoistically such that he kills White. Last, Pereboom presents Case 4, wherein Plum is an ordinary human being in a deterministic universe and Plum's egoistic decision to kill White is causally determined by the past and laws of nature. Again, in all four cases Plum satisfies *all* purported compatibilist requirements for free will and moral responsibility *and* Plum's actions are causally determined by factors outside of his control.[3] Pereboom claims: "The salient factor that can plausibly explain why Plum is not responsible in all of the cases is that in each he is causally determined by factors beyond his control to decide as he does. This is therefore a sufficient, and I think also the best, explanation for his non-responsibility in all of the cases" (2014, 79).

Given this presentation, whether Pereboom's argument successfully poses a problem for compatibilist accounts of free will and moral responsibility depends on a few conditions being met. First, readers must not be confused about the causal nature of determinism. Second, readers must truly understand that Plum meets all compatibilist requirements for moral responsibility. Third, readers must find Plum intuitively not morally responsible. Last, since Pereboom is attempting to show both that the compatibilist conditions for free will are insufficient and that determinism is incompatible with free will and moral responsibility, a single feature of these cases – that Plum's actions are causally determined by factors outside his control – needs to explain why it is that individuals intuitively find Plum not morally responsible. If this intuition is the result of any other aspects of the argument, then Pereboom's argument fails because something independent of the features of determinism would best explain why people judge that Plum lacks moral responsibility. Given that correctly explaining intuitions about Plum is

[2] Pereboom asserts that in all four cases Plum satisfies the requirements which Hume (1739/1978), Harry Frankfurt (1971), John Fischer and Mark Ravizza (1998), Jay Wallace (1994), and Alfred Mele (1995; 2006) have argued are necessary for an agent to be considered morally responsible.

[3] While it may be impossible for both manipulation to occur and for manipulated agents to meet all compatibilist requirements for moral responsibility (Demetriou 2010), for the purposes of this paper, I will assume these features are compatible with one another.

vital for the success of Pereboom's argument, Pereboom (2014) presents the four-case manipulation argument as an argument for the best explanation.[4]

There is reason to believe that readers are easily confused about what determinism entails (Murray and Nahmias 2014), that readers fail to understand manipulated agents as having all of the necessary compatibilist requirements for moral responsibility (Sripada 2011), and that readers don't actually get the intuition that Plum lacks moral responsibility (Feltz 2013). While these are significant problems for Pereboom's argument, I will focus my attention only on the problem that arises from neglecting to respect other factors that may influence intuitions of non-responsibility. I argue Pereboom's presentation of the four-case argument likely leads to certain, largely unconscious, psychological influences driving intuitions that Plum is not morally responsible. Since the effects of these unconscious psychological influences lead to order effects, I argue that order effects can provide a plausible, and likely better, explanation for why readers get the intuition that Plum is not morally responsible. Pereboom's argument is credibly threatened and potentially undermined by neglecting to ascertain the presence and impact of such influences.

It is important to note that I am not offering a hard-line response and arguing that Plum actually *should* be considered morally responsible in all four cases (McKenna 2008), nor am I taking a soft-line response and arguing that there is a relevant dissimilarity between two of the cases which allows us to consider Plum not morally responsible in Case 1 but morally responsible in Case 4 (Demetriou 2011, Waller 2013). Rather, I take a stance similar to Mele (2005) and call into question Pereboom's explanation for why we find it intuitive that Plum is not morally responsible. We must be certain that the cases are presented in such a way that our intuitions actually track features relevant to the debate before arguing about what intuitions are appropriate for each case. [5]

III. Order Effects as an Alternative Explanation

In this section, I argue that order effects serve as a plausible alternative explanation for what likely motivates judgments of Plum's non-responsibility in the four-case

[4] It has been argued that the four-case manipulation argument can be best employed without understanding it as an argument for what best explains intuitions (Mele 2005). I will address this objection later in this paper.

[5] Kadri Vihvelin has recently made a similar point and argued that using certain intuitions and thought experiments where is not clear what is being described or when we do not all agree about the verdict, such as manipulation cases, is not helpful for advancing the free will debate.

argument. After providing evidence that the order in which Pereboom's four cases are presented affects judgments about whether Plum is morally responsible, I will discuss specific features and psychological mechanisms which likely lead to order effects occurring in the four-case argument.

Alex Weigmann, Yasmina Okan, and Jonas Nagel (2012) demonstrated that the order in which trolley dilemmas are presented significantly influences judgments of moral permissibility.[6] After presenting participants with five variations of the trolley dilemma, which differed only in what the life-saving action was, they found that the order in which the cases were presented drastically influenced responses to each scenario. [7] Weigmann et al. concluded, "judgments would be most likely transferred if the initial rating was strongly negative" (2012, 825). That is, when readers had a strongly negative judgment towards the first case, this judgment was likely to affect judgments of later cases. This highly negative first case resulted in consistently more negative judgments of moral permissibility relative to judgments of these cases presented on their own. Given that readers have strongly negative reactions to Case 1 in Pereboom's four-case manipulation argument (Feltz 2013), I argue it is highly likely that the order in which these cases are presented by Pereboom has an effect on judgments of Plum's level of moral responsibility in later cases much in the same way Wigmann et al. observed order affected judgments about the trolley dilemmas.

While one might assume the experienced agnostic philosopher would not be affected by the order in which cases are presented, Schwitzgebel and Cushman (2012) found that with respect to moral principles, order of presentation influenced the judgments

[6] Trolley dilemmas are scenarios where a trolley train is out of control and on track to run over multiple workers. However, someone has the option of choosing to sacrifice the life of one person to save the multitude.

[7] The potentially life-saving actions were: pressing a switch that will redirect the train that is out of control to a parallel track where one person will be run over; redirecting an empty train that is on a parallel track onto the main track to stop the train, running over a person that is on the connecting track; redirecting a train with a person inside that is on a parallel track onto the main track to stop the train; pushing a button that will open a trap door that will let a large person on top of a bridge fall and stop the train; push the large person from the bridge to stop the train.

of philosophers *more* than it did non-philosophers![8] Furthermore, this effect persists among philosophers self-reporting familiarity, expertise, stability, and specialization in ethics (Schwitzgebel and Cushman forthcoming). Not only does this finding suggest that philosophers need to take the salience of order effects seriously, it provides reason for philosophers to take these effects more seriously than others. If it turned out that order effects better explain why we find Plum not morally responsible in Case 4, then Pereboom would fail to provide the best explanation for these intuitions and his argument would be unsuccessful.

Agency-Detection Mechanism

A psychological mechanism that likely guides intuitions and contributes to the effect that order has on judgments regarding Pereboom's four cases is an agency-detection mechanism. Scott Atran (2006) argues that human evolution has naturally selected for an innate and overly sensitive mechanism for detecting agents and agential properties. While this mechanism often beneficially and accurately identifies agents, Atran argues that it also causes humans to wrongly attribute agential properties to nearly any complex or uncertain situation or design. For example, Atran believes this overly sensitive mechanism explains why people often see faces in the clouds and are quick to believe in supernatural beings. This mechanism would become active in Case 1 and correctly lead us to attribute agential properties to the causal determinants of Plum's actions (i.e., the neuroscientists). However, an agency-detection mechanism would likely remain active in later cases when Pereboom replaces these agents with the complex structure of causal determinism, which, importantly, contains no agential properties. If this mechanism remained active, then readers would (perhaps unconsciously) attribute agential properties to the causal determinants of Plum's actions in Case 4. Such attributions would, thereby, alter judgments of Case 4 by confusing the reader about the nature of determinism.[9]

[8] Pereboom (2014, 81) states, "…the manipulation argument aims to persuade the natural compatibilist and the agnostic their resistance to incompatibilism is best given up." While it is extremely important to properly recognize who Pereboom's intended audience *is,* who Pereboom's audience *ought to be,* and to what degree such an audience actually exists, I do not have room to adequately address these concerns in this paper.

[9] In an unpublished manuscript, Neil Levy makes a similar argument, claiming that Pereboom's four-case manipulation argument only succeeds insofar as it activates an agency-detection mechanism which causes the readers to see determinism in agential terms.

If this overly sensitive agency detection mechanism does, in fact, influence intuitions about Case 4, then the order in which Pereboom presents these cases has an effect on judgments of Plum's non-responsibility. Furthermore, this alternative explanation for intuitions would undermine Pereboom's goal of getting readers to properly understand the causal nature of determinism. Since determinism, and therefore Case 4, does not involve agents or agential properties which influence Plum, it would be misguided for intuitions about Case 4 to be influenced by agency. If intuitions about Plum in Case 4 are motivated by an agency-detection mechanism responding to agency in earlier cases, as I argue they are, then these intuitions are unreliable and cannot be used to motivate Pereboom's argument.

Intent

While the mere presence of agents in Case 1 might cause readers to judge Plum not morally responsible in Case 4, the intent of these agents also appears to contribute to the order effects. Phillips and Shaw (forthcoming) investigated how third-party intent (the intent of agents who causally determine how another agent acts but nonetheless are not involved in the action themselves) influences judgments of moral responsibility. First, they found that the presence of third-party intent does reduce judgments of blame.[10] Second, third-party intent *only* influenced judgments when the agent's actions perfectly match with the intended action. Third, their results suggest that intent affects judgments of moral responsibility by altering the reader's causal perception. If Pereboom's four-case argument successfully alters one's causal perception only because third-party intent is present in earlier cases, then judgments of earlier cases are influencing judgments of later cases, and order effects are thereby produced. If intuitions of Plum's non-responsibility are the result of order effects, then we have an alternative explanation for these intuitions that is deeply problematic for Pereboom's argument.

To see why third-party intent altering judgments would be problematic, consider that according to Pereboom, many people don't see determinism as ruling out the possibility of moral responsibility because they misunderstand the true nature of determinism. To remedy these misconceptions, "the manipulation cases are formulated so as to correct for inadequacy in the extent to which we take into account hidden deterministic causes in our intuitions about ordinary cases" (2014, 95). That is, manipulation cases are intended to expose to us the true causal nature of determinism and they attempt to alter how one

[10] These findings are consistent with Robyn Waller's (2013) argument that intent is a relevant difference between cases and affects judgments of moral responsibility.

perceives the causal implications of determinism. Phillips and Shaw's research suggests that manipulation cases can succeed in altering one's causal perception *only* when third-party intent is present and matches the action performed. Therefore, according to Phillips and Shaw's assessment, if a change in causal perception occurs, it must be because readers understand there to be third-party intent present which matched the action. While Pereboom is clearly attempting to change the reader's causal perception, it would be mistaken to alter perceptions by getting readers to understand determinism as having any intent (or, for that matter, any other agential properties) since compatibilists and incompatibilists agree this is the wrong way to conceive of determinism. This suggests that Pereboom elicits the desired intuitions by confusing readers about the true nature of determinism.

While the concern outlined above is certainly problematic for Pereboom's argument, it is worth noting that in order for my argument to succeed, intent doesn't necessarily need to confuse readers about the true nature of determinism. Rather, I merely need to demonstrate that the intent, along with other unconscious psychological influences, lead to order effects influencing judgments and that these order effects explain intuitions of non-responsibility better than the mere fact that Plum's actions are causally determined by factors over which he has no control.[11]

Emotional Responses

Another psychological influence that likely motivates order effects in Pereboom's four-case argument is emotional engagement with features present in Case 1. The first case of the four-case argument involves agential intent, an abnormal bodily violation (brain manipulation), and an abnormal social violation (manipulation). Reading vignettes that contain intent, abnormal bodily violations, and abnormal social violations have been shown to elicit emotional responses (Giner-Sorolla 2011; Haidt 2003). Also, engaging emotionally with such vignettes has been shown both to be correlated with particular moral judgments (Greene 2001), as well as to influence moral judgments (Haidt 2003; Guiseppe et al. 2012) even when these emotions are primed non-consciously and

[11] In a response to Mele's criticisms, Pereboom (2014, 82) argues even if these intentional agents, "were replaced by force fields or machines that randomly form in space that have the same deterministic effect on Plum as the manipulators do, the intuition that Plum is not morally responsible persists." While I remain skeptical of this claim, it is interesting that Pereboom chooses not to make this replacement and he only mentions such a possibility after priming the reader with cases involving intentional agents.

automatically (Valdesolo and DeSteno 2006).[12] Furthermore, responding emotionally to a vignette has been shown to affect judgments and behavior continually for some time after reading the vignette (Plaisier and Konijn 2013; He et al. 2013).

In light of such evidence, it seems very likely that readers of Pereboom's four-case argument would have a strongly negative emotional response to Case 1 and that this highly negative response would influence judgments regarding Case 4. Insofar as one's emotions are negatively responding to agential intent, body violations, or social violations, and not to the fact that Plum's actions are causally determined by factors he has no control over, emotional engagement serves as a plausible confounding variable for what explains judgments. That is, if our intuitions about Plum are the result of responding to emotional-priming factors that are irrelevant to determinism, then it isn't a feature of determinism that drives moral judgments, as Pereboom argues. Since features of Case 1 are known to elicit emotional reactions, it seems likely that emotional engagement with features present in Case 1 influence judgments of later cases, thus leading to order effects taking place. These order effects, again, serve as an alternative explanation for intuitions of Plum's non-responsibility in Case 4 and thereby threaten the success of Pereboom's four-case manipulation argument.

In summary, given Pereboom's presentation of his four-case manipulation argument, it is likely that features only present in earlier cases (agents, third-party intent, abnormal body and social violations) are initiating certain unconscious psychological mechanisms that drive judgments of Case 4, thus resulting in order effects. There may be additional psychological influences that drive order effects which I have not discussed. For example, intuitions could also be swayed by one's own demands for consistency across cases, readers having intuitions of non-responsibility simply because Pereboom makes suggestions about what intuitions readers ought to have, or readers agreeing with Pereboom because he is understood to be some kind of authority figure on what one ought to think about these cases. If *any* such influences, either collectively or on their own, better explain why we (or "agnostic" readers) find Plum intuitively not morally responsible, then Pereboom's argument is unsuccessful. Therefore, Pereboom, like anyone else attempting to make claims about what drives intuitions, needs to take unconscious psychological influences seriously. As I have now demonstrated, neglecting

[12] Haidt (2001) argues that in most circumstances, emotional engagement is the primary cause of moral reasoning. While this may or may not be the case, for my argument to work, it only needs to be the case that emotional engagement influences judgments of Pereboom's four cases.

to acknowledge seemingly irrelevant influences, such as order effects, can undermine one's entire argument.

IV. Objections

Thus far, I have argued that by failing to recognize salient and largely unconscious psychological influences that have been shown to affect intuitions, Pereboom's four-case manipulation argument likely does not elicit judgments about moral responsibility in a way that is required to support the argument. More specifically, I have argued that the intuition that Plum is not morally responsible is not likely best explained by the fact that Plum's actions are causally determined by factors outside of his control. Rather, these intuitions are more plausibly explained by the presence of order effects that are driven by certain psychological influences which readers are largely unaware of, such as an agency-detection mechanism, third-party intent, and highly negative emotional engagement. I will now entertain objections to my argument.

Order Effects Are Intended

First, one might be tempted to object to my argument by saying something like the following: "Of course order effects sway intuitions in Pereboom's favor. The whole point of the four-case argument is to lead people to understand that the factors that undermine moral responsibility in Case 1 undermine responsibility in Case 4 as well. Therefore, the emotional responses and initial judgments about Case 1 *should* transfer over and influence intuitions about Case 4 so that we think of these cases in the same way and with the same types of attitudes."

In response to this objection, I would first point out that insofar as Pereboom's four-case manipulation argument is to be understood as an argument to the best explanation, the argument only works if Pereboom's explanation is actually the best. Therefore, if the fact that Plum's actions are being determined by factors outside his control is *not* what drives intuitions, then the argument simply doesn't work. Mele (2005; 2008) has argued that readers would judge Plum not morally responsible even if the causation in these cases was indeterministic, and this would show that determinism is not what motivates intuitions about the four cases. If Mele is right and deterministic causation isn't what drives intuitions, then these judgments must be sensitive to other factors within these cases. I presented a few likely candidates for which features of these cases influence intuitions regarding Case 1: the presence of agents, third-party intent, and emotionally responding to manipulation. Furthermore, I provided reason to believe that if the factors I discuss are what motivate intuitions about Case 1, then it's highly likely that order effects will take place as a result and intuitions of non-responsibility will remain consistent across cases. Therefore, order effects driven by psychological influences that attend to features present in Case 1 serve as a confounding variable for the success of Pereboom's argument if these order effects better account for what motivates the intuition that Plum is not morally responsible.

As a second response to this objection, I'd point out that if order effects are supposed to take place and we are supposed to understand Case 1 and Case 4 in roughly the same way, then Pereboom is likely confusing the reader about the true causal nature of determinism. As discussed earlier, if the intuition that Plum is not morally responsible in Case 4 is residually influenced by the presence of agents or third-party intent in Case 1, then the intuitions about Case 4 are misguided since determinism has no agential properties or intentions.

If it turns out that intuitions about Case 1 are solely, or at least primarily, motivated by the fact that Plum's actions are causally determined by factors outside his control, and if after reading the four-case argument readers are not at all confused about determinism, then judgments regarding Case 4 being influenced by order effects would not be problematic. However, as I have now argued, it seems extremely unlikely that judgments are best explained by the single feature Pereboom addresses, given the many other features present in Case 1 that are known to engage psychological mechanisms that lead to order effects and alter judgments of later cases. Furthermore, it seems plausible that readers are conflating features such as agency and intent with determinism in Case 4, thus confusing the reader about the true nature of determinism. Work in experimental philosophy has provided evidence of such confusion (Murray and Nahmias 2014; Sripada 2011).

Explaining Intuitions Is Unimportant

A second objection to my argument is that by presenting Pereboom's four-case argument as an argument to the best explanation, I am misrepresenting it. Thus far, I have been assuming that Pereboom's explanation for intuitions about the four cases is a central feature of his argument. Nonetheless, it's possible that one can conclude Plum is not morally responsible without offering any explanation at all for what drives these intuitions. In response to this objection, I argue that this alternative understanding of the four-case manipulation argument, besides running counter to Pereboom's stated intentions, is extremely problematic.

In Pereboom's most recent presentation of his four-case argument, he argues,

> It's highly intuitive that Plum is not morally responsible in Case 1, and there are no differences between Cases 1 and 2, 2 and 3, and 3 and 4 that can explain in a principled way why he would not be responsible in the former of each pair but would be in the latter. We are thus driven to the conclusion that he is not responsible in Case 4. The salient factor that can plausibly explain why Plum is not responsible in all of the cases is that in each he is causally determined by factors beyond his control to decide as he does. This is therefore a sufficient, and I think also the best, explanation for his non-responsibility in all of the cases. (2014, 79)

This passage might lead one to assume Pereboom's argument is similar to other manipulation arguments, which can very roughly be formulated in the manner below. I will refer to this formulation as MA.

(P1) Plum is not morally responsible in Case 1.

(P2) There are no differences between cases that are relevant to moral responsibility.

(C) Therefore, Plum is not morally responsible in Case 4, and since Plum in Case 4 is no different from any agent in a deterministic universe, no agents in a deterministic universe are morally responsible either.

MA seems to get the conclusion Pereboom desires without employing any premises that explain intuitions. While one *could* present Pereboom's argument in a way that does not make use of his explanation for intuitions, I would argue that this understanding of Pereboom's argument would be problematic.

Though there may be other problems with this kind of formulation, I will focus my attention on the fact that it draws a conclusion about moral responsibility directly from an intuition about moral responsibility: Plum *is intuitively* not morally responsible in Case 1. Therefore, Plum *is* not morally responsible in Case 1. This reasoning is what motivates P1 of MA. Nonetheless, if this move is permitted then compatibilists could simply employ the same reasoning and argue that because they find persons in deterministic universes *intuitively* morally responsible, then these agents must actually be morally responsible

(King 2013). Furthermore, if such reasoning is permitted, then debates about free will and moral responsibility would be reduced to a battle of intuitions instead of being won via philosophical argumentation. While this reduction is undesirable and would likely be unfruitful, one might argue this is what Pereboom has in mind. For instance, in his response to McKenna's criticisms of the four-case argument, Pereboom (2005, 242) suggests we "let the intuitions fall where they may."

Appealing to intuitions without any explanation for what drives these intuitions may be useful if virtually all readers have the same intuition about the cases presented. However, this universality doesn't seem to occur with Pereboom's four-case manipulation argument (Feltz 2013) or similar cases involving manipulation or determinism (Murray and Nahmias 2014; Nichols and Knobe 2007; Sripada 2011). Given that intuitions about these cases are not uniform, the only ways to avoid a stalemate is to explain what drives intuitions about P1 or simply provide a separate, substantive philosophical argument which justifies P1.

I assume that Pereboom intends to avoid such a stalemate and the related methodological issues which arise from understanding his argument to be formulated similar to MA. There is good reason to consider Pereboom's explanation for what drives intuitions as a significant aspect of his four-case argument, since Pereboom himself explains this is how the argument ought to be understood in a footnote. He states,

> Al Mele (2006) argues that a manipulation argument against compatibilism need not be cast as an argument to the best explanation. I doubt that this is so. True, the argument can be represented without a best-explanation premise, but such a representation will not reveal its real structure. By analogy, the teleological argument for God existence can be represented as a deductive argument, but its real structure is an argument to the best explanation for biological order in the universe. The fact that the real structure of a manipulation argument against compatibilism is an argument to the best explanation becomes clear when one considers compatibilist objections to it—that, for, example, the non-responsibility intuitions can be accounted for by manipulation of a certain sort and not by causal determination. (2015, 79-80)

Here Pereboom makes it clear that his argument is one in which the explanation of intuitions is paramount. Furthermore, he states that the way one should object to his argument is by providing an alternative explanation for what causes intuitions of Plum's non-responsibility. This is exactly what I have attempted to do in this paper.

As a final note, I'd point out that the claims Pereboom and myself make about what best explains intuitions are empirical claims. It's possible to manipulate the features of these cases and determine what does and does not motivate intuitions. Furthermore, we can test whether, after reading Pereboom's four-case argument, readers correctly understand the true causal nature of determinism. If it turns out intuitions of Plum's non-responsibility are directly driven by the fact that Plum's actions are causally determined by factors outside his control and, if after reading all four cases, readers understand exactly what determinism entails, then Pereboom's argument would successfully avoid my criticisms in this paper. I doubt, however, that this is what we would find and hope to investigate these matters empirically in the future.

V. Conclusion

The goal of this paper was to demonstrate that arguments which appeal to intuitions about thought experiments and hypothetical cases must acknowledge the many psychological influences that subtly motivate intuitions. I argued that influences, such as order effects, can affect judgments to the extent that arguments which employ these cases are unsuccessful. Without ensuring that our intuitions are tracking relevant features of an argument, intuitions regarding thought experiments will likely be unreliable and, therefore, fruitless for the purposes of philosophical discussion. To exemplify these concerns, I presented Derk Pereboom's four-case manipulation argument. I have provided evidence that suggests intuitions about these four cases can better be explained by order effects than by recognizing that Plum's actions are causally determined by factors outside of his control. Since it may be the case that what best explains intuitions of Plum's non-responsibility across all four cases is not that Plum's actions are causally determined by factors outside his control, order effects serve as a plausible alternative explanation for what drives intuitions. If what drives intuitions about Pereboom's hypothetical cases are factors irrelevant to causal determinism, as I argue is the case, then by failing to correctly identify what motivates intuitions about his four cases, Pereboom's argument is unsuccessful.

My suggestion to consider alternative psychological explanations, such as order effects, when explaining what motivates intuitions does not solve the potential problem of unreliability that arises as a result of intuitions differing across demographics. However, I have provided evidence that intuitional unreliability, in the sense that intuitions are sensitive to trivial features of hypothetical cases and thought experiments, is problematic when one's explanation for what motivates these intuitions is incorrect. One must take seriously the fact that intuitions are influenced by many seemingly irrelevant factors when attempting to use thought experiments or hypothetical cases to provide support for an argument. Just as a good scientist considers all confounding variables before claiming to know the cause of a certain event, philosophers must address potential confounding factors for intuitions.

References

Alexander, Joshua, Ronald Mallon, and Jonathan Weinberg. 2014. "Accentuate the Negative." In *Experimental Philosophy* Volume 2, Edited by Joshua Knobe and Shaun Nichols, 31–50. New York: Oxford University Press.

Atran, Scott. 2006. "Religion's Innate Origins and Evolutionary Background." In *The Innate Mind: Culture and Cognition*, edited by Peter Carruthers, Stephen Laurence, and Stephen Stich, 302–317. Oxford: Oxford University Press.

Buckwalter, Wesley, and Stephen Stich. 2011. "Gender and the Philosophy Club." *The Philosophers' Magazine* 52: 60–65.

Dennett, Daniel C. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge: The MIT Press.

Demetriou, Kristin. 2010. "The Soft-Line Solution to Pereboom's Four-Case Argument." *Australasian Journal of Philosophy* 88 (4): 595–617.

Feltz, Adam. 2013. "Pereboom and premises: Asking the right questions in the experimental philosophy of free will." *Consciousness and Cognition* 22 (1): 53-63.

Fischer, John Martin, 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." *Journal of Philosophy* 68 (1): 5–20.

Giner-Sorolla, Roger, and Pascale Sophie Russell. 2011. "Moral anger, but not moral disgust, responds to intentionality." *Emotion* 11 (2): 233–240.

Greene, Joshua D. 2011. "An fMRI Investigation of Emotional Engagement in Moral Judgment." *Science* 293 (5537): 2105–2108.

Guiseppe, Ugazio, Claus Lamm, and Tania Singer. 2012. "The role of emotions for moral judgments depends on the type of emotion and moral scenario." *Emotion* 12 (3): 579–590.

Haidt, Jonathan. 2001. "The Emotional Dog And Its Rational Tail: A Social Intuitionist Approach To Moral Judgment." *Psychological Review* 108 (4): 814–834.

Haidt, Jonathan. 2003. "The moral emotions." In *Handbook of Affective Sciences*, edited by Richard J Davidson, Klaus R Sherer, and H. Hill Goldsmith 852–870. Oxford: Oxford University Press.

Haji, Ishtiyaque. 1998. *Moral Accountability*. New York: Oxford University Press.

Haji, Ishtiyaque. 2009. *Incompatibilism's Allure: Principal Arguments for Incompatibilism*. Peterborough ON: Broadview Press.

He, J.; X. Jin, M. Zhang, X. Huang, R. Shui, and M. Shen. 2013. "Anger and selective attention to reward and punish children." *Journal of Experimental Child Psychology* 115 (3): 389-404.

Hume, David. (1739) 1978. *A Treatise of Human Nature*. Oxford: Oxford University Press.

King, Laura A., and Joshua A Hicks. 2011. "Subliminal mere exposure and explicit and implicit positive affective responses." *Cognition & Emotion* 25 (4): 726–729.

King, Matt. 2013. "The Problem With Manipulation." *Ethics* 124 (1): 65–83.

Levy, Neil. Unpublished manuscript. "Manipulating the Reader: Manipulation Arguments and Agency Detection."

Li, Wen, Richard E. Zinbarg, Stephan G. Boehm, and Ken A. Paller. 2008. "Neural and Behavioral Evidence for Affective Priming from Unconsciously Perceived Emotional Facial Expressions and the Influence of Trait Anxiety." *Journal of Cognitive Neuroscience* 20 (1): 95–107.

McKenna, Michael. 2008. "A hard-line reply to Pereboom's four-case manipulation argument." *Philosophy and Phenomenological Research* 77 (1): 142–159.

Mele, Alfred. 1995. *Autonomous Agents*. New York: Oxford University Press.

Mele, Alfred. 2005. "A critique of Pereboom's 'four-case argument' for incompatibilism." *Analysis* 65 (1): 75-80.

Mele, Alfred. 2006. *Free Will and Luck*. New York: Oxford University Press.

Mele, Alfred. 2008. "Manipulation, Compatibilism, and Moral Responsibility." *The Journal of Ethics* 12 (3–4): 263–286.

Mlodinow, Leonard. 2012. *Subliminal: how your unconscious mind rules your behavior*. New York: Pantheon Books.

Murray, Dylan, and Eddy Nahmias. 2014. "Explaining Away Incompatibilist Intuitions." *Philosophy and Phenomenological Research* 88 (2): 434–467.

Nichols, Shaun, and Joshua Knobe. 2007. "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions." *NOUS* 41 (4): 663–685.

Pereboom, Derk. 2005. "Defending Hard Incompatibilism." *Midwest Studies in Philosophy* 29 (1): 228–247.

Pereboom, Derk. 2001. *Living without free will*. Cambridge: Cambridge University Press.

Pereboom, Derk. 2014. *Free will, agency, and meaning in life*. Oxford: Oxford University Press.

Petrinovich, Lewis, and Patricia O'Neill. 1996. "Influence of Wording and Framing Effects on Moral Intuitions." *Ethology and Sociobiology* 17 (3): 145–171.

Phillips, Jonathan, and Shaw, Alex. Forthcoming. "Manipulating Morality: Third-Party Intentions Alter Moral Judgments by Changing Causal Reasoning."

Plaisier, Xanthe S., and Konijn, Elly A. 2013. "Rejected by peers—Attracted to antisocial media content: Rejection-based anger impairs moral judgment among adolescents." *Developmental Psychology* 49 (6): 1165-1173.

Schwitzgebel, Eric, and Fiery Cushman. 2012. "Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers." *Mind & Language* 27 (2): 135–153.

Schwitzgebel, Eric, and Fiery Cushman. Forthcoming. "Professional Philosophers' Susceptibility to Order Effects and Framing Effects in Evaluating Moral Dilemmas."

Sinnott-Armstrong, Walter. 2008. "Framing Moral Intuition." In *Moral Psychology, Vol 2. The Cognitive Science of Morality: Intuition and Diversity*, 47–76. Cambridge, MA: MIT Press.

Sripada, Chandra. 2011. "What makes a manipulated agent unfree?" *Philosophy and Phenomenological Research* 85 (3): 1–31.

Tobia, Kevin P., Gretchen B. Chapman, and Stephen Stich. 2013. "Cleanliness is Next to Morality, Even for Philosophers." *Journal of Consciousness Studies* 20 (11–12): 195–204.

Todd, Patrick. 2012. "Defending (a modified version of) the Zygote Argument." *Philosophical Studies* 164 (1): 189–203.

Valdesolo, Piercarlo, and David DeSteno. 2006. “Manipulations Of Emotional Context Shape Moral Judgment.” *Psychological Science* 17 (6): 476–477.

Vaesen, Krist, Martin Peterson, and Bart Van Bezooijen. 2013. “The Reliability of Armchair Intuitions.” *Metaphilosophy* 44 (5): 559–578.

Vihvelin, Kadri. “How Not to Think about Free Will.” *Journal of Cognition and Neuroethics* 3 (1).

Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments.* Cambridge: Harvard University Press.

Waller, Robyn. 2013. “The Threat of Effective Intentions to Moral Responsibility in the Zygote Argument.” *Philosophia* 42 (1): 209–222.

Weigmann, Alex, Yasmina Okan, and Jonas Nagel. 2012. “Order effects in moral judgment.” *Philosophical Psychology* 25 6: 813–836.

Weinberg, Jonathan, Shaun Nichols, and Stephen Stich. 2001. “Normativity and epistemic intuitions.” *Philosophical Topics* 29 (1-2): 429–460.

Weinberg, Jonathan, Stacey Swain, and Joshua Alexander. 2008. “The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp.” *Philosophy and Phenomenological Research* 77 (1): 138–55.