Neuroscientific Prediction and the Intrusion of Intuitive Metaphysics

Neuroscientific Prediction and the Intrusion of Intuitive Metaphysics*

David Rose

Rutgers University

Wesley Buckwalter

University of Waterloo

Shaun Nichols

University of Arizona

* This is the penultimate draft of a manuscript to appear in *Cognitive Science*. Please cite the published version.

Abstract

How might advanced neuroscience—in which perfect neuro-predictions are possible—interact with ordinary judgments of free will? We propose that peoples' intuitive ideas about indeterminist free will are both imported into and intrude into their representation of neuroscientific scenarios and present six experiments demonstrating intrusion and importing effects in the context of scenarios depicting perfect neuro-prediction. In light of our findings, we suggest that the intuitive commitment to indeterminist free will may be resilient in the face of scientific evidence against such free will.

Keywords: free will; neuroscience; intrusion effect; importing; cultural transmission; compatibilism

1. Introduction

The achievement of scientific knowledge across the natural, life, social, and cognitive sciences has radically impacted the way we view the world. With these achievements however comes the question of how increased scientific knowledge interacts with fundamental concepts in social cognition that are also central to how we evaluate the world, such as evaluations of human agency, causation and free will. For example, many theorists have argued that our knowledge of the brain will one day advance to the point where the perfect neuroscientific prediction of all human choices is theoretically possible (Coyne, 2012; Harris, 2012). This sufficiently advanced brand of neuroscience would be able to predict what people will think and do with absolute certainty. At the same time, research in psychology suggests that people do not think human choices are deterministically caused by prior events, which suggests that people's actions are not thought to be perfectly predictable (e.g., Stillman, Baumeister & Mele, 2011; Monroe, Dillon & Malle, 2014; Nichols & Knobe, 2007; Nichols, 2012; Rose & Nichols, 2013; Sarkissian et al., 2010; Turri, Rose & Buckwalter, 2015; Turri, 2015; though see e.g., Murray & Nahmias, 2014; Nahmias, Morris, Nadelhoffer & Turner, 2006). Thus there seems to be a looming conflict between the worldview encouraged by advancements in scientific knowledge and commonsense notions of human choice.

This apparent conflict between future neuroscience and intuitive notions of choice is, of course, an instance of the much larger question about how our humanistic concerns will interact with our developing scientific knowledge. The theme is familiar throughout the history of science from the Copernican revolution to the theory of evolution. In the case of free will and neuroscience, one way to explore the issue is by examining people's reactions to imaginative neuroscientific scenarios. In a recent study by Nahmias, Shepard and Reuter (2014), for example,

researchers presented participants with scenarios involving neuroscientists that can "predict with 100% accuracy every single decision a person will make" and where "everything that any human thinks or does could be predicted ahead of time based on their earlier brain activity" (p. 514). They found that people overwhelmingly attribute free will to agents in these contexts even when their behavior is predicted by neuroscience with absolute certainty.

Nahmias and colleagues interpret such ascriptions as evidence for the compatibilist view that there is no inherent conflict between a perfectly predictive neuroscience and the common notion of free will. According to this view, people fully accept the conditions of neuro-prediction cases, and then, attribute free will to agents in those conditions. However Nahmias and colleagues also acknowledge the possibility that "many participants may be failing to understand or internalize relevant information from the scenarios" and speculate that if "participants did not attend to the fact that every decision could be predicted with 100% accuracy" or "could be predicted before the agent was even aware of making their decision" then it would significantly challenge the inference that people are broadly comfortable with the idea of perfect neural prediction (Nahmias et al., 2014, p. 512).

Despite these possibilities, the extent to which participants understand and internalize relevant features of neuro-prediction scenarios has not been measured directly. This paper presents six experiments that measure this directly. We provide evidence that participants *fill in* the scenarios in ways that undermine the inference that the ordinary notion of free will is compatible with the idea of perfect neural prediction.

There are, of course, many ways in which participants might fill in imaginative scenarios. In some cases, the details of such narratives are filled in with information from intuitive theories and other background assumptions. A powerful demonstration of filling in comes from the

cognitive science of religion, in which the intuitive views that people hold about human agency influence their comprehension of stories (e.g., Boyer, 1994; Barrett & Keil, 1996). To show this, researchers played short audio narratives involving God as an agent (Barrett & Keil, 1996). After a short delay, participants were asked to recall whether certain pieces of information were included in the narrative. Some of these items were not included in the narrative, but would naturally be imposed by an intuitive concept of human agency. One vignette, for instance, featured God responding to two prayers. Participants recalled that God finished responding to one prayer and then responded to the other, despite the fact that no such temporal sequencing was stated in the vignette. This suggests that participants' intuitive views about the temporality of agency informed their representation of the narrative. These findings are particularly striking since the same participants also explicitly affirmed elsewhere in the study that God could do two things at once.

Following Barrett and Keil's example, we hypothesize that participants' representations of neuroscientific scenarios are *filled in* with the intuitive views that people hold about human agency. It's useful to distinguish two kinds of filling in, what we'll call "importing" and "intruding". *Importing* occurs when participants fill in the scenario in ways that are *consistent* with the scenario, but the filling-in systematically goes beyond the information provided in the scenario. Of course, when participants read vignettes, importing will be a common occurrence. It becomes theoretically interesting when the imported information undermines the interpretation of the results. *Intruding* occurs when the filling in leads to a *mis*representation of the scenario.

Prior work on beliefs about indeterminist agency indicates at least one way in which filling-in might result in intrusion. This work suggests that the intuitive view of human agency that people hold is indeterministic. In other words, just as people naturally tend to assume a

temporal ordering in social cognition, they also naturally tend to assume that human decisions are undetermined. If people fill in scenarios depicting perfect neuro-prediction based on prior beliefs about indeterminist free will, then this could actually lead them to *reject* the notion that decisions are perfectly predictable in the way specified in the scenario. Thus it could be that when people ascribe free will in cases of perfect neuro-prediction, this might be accompanied by the intuition of indeterminist free will, resulting in an inaccurate representation of the scenario.

We develop a number of strategies to test whether this occurs in neuro-prediction scenarios. Previous work indicates that people tend to think that free choices issue from conscious awareness (e.g., Shepherd, 2012) and that when a person makes a free choice, they could have chosen otherwise even if all the determining conditions were the same (Nichols, 2012; Turri, 2015). The vignettes used by Nahmias and colleagues explicitly state that the action initiation is generated before conscious awareness and that the prediction is 100% accurate, which indicates that after initiation of the brain process, the agent could *not* have done otherwise. If attributions of indeterminist free will are intruding on the interpretation of the scenarios of perfect neuroscientific prediction, then we might expect that those who affirm free will in the neuroscientific scenarios will also be likely to attribute to the agent the ability to change her decision. One way to express indeterminist free will is to claim that the agent could have done otherwise despite the same determining conditions (e.g., prior brain states). A different way to express indeterminist free will is in terms of the possibility for the agent to make a different decision than the one predicted by her patterns of brain activity. The presence of these judgments would suggest that intrusion has occurred. Of course, we might naturally expect judgments about free will to regularly coincide with judgments about options to do otherwise in normal circumstances, just like we might expect temporal ordering to. Such judgments qualify as

"intrusion" in the present context however insofar as they lead to a misrepresentation of the conditions of perfect neuro-prediction stated in the story.

It is likely that *importing* also occurs in neuro-prediction stories. To investigate this issue we devise variations of neuro-prediction stories utilizing a technique in the free will literature known as "rollback" in which one imagines what would happen if we rolled back time to a certain point in the past and let events unfold again (van Inwagen, 2000; Nahmias et al., 2006). If people think that events might unfold differently, this suggests that they import indeterminist beliefs in their representations of the scenario. To anticipate our results, we find that importing of indeterminist free will occurs in these scenarios and provides further evidence for another kind of problematic filling in of neuro-prediction stories.

Experiment 1 demonstrates that those who affirm free will in Nahmias et al.'s neuro-prediction case do so while imposing indeterministic details contrary to those stated in the story. This indicates that people misrepresent instances of perfect neuro-prediction and thus that intrusion has occurred. Experiment 2 replicates this intrusion effect and demonstrates that it plays a mediating role in the comprehension of neuro-prediction stories. Experiments 3 and 4 find the same kind of intrusion effect in a different narrative context utilizing simplified cases with more minimally matched pairs. Experiment 5 again finds intrusion effects and demonstrates that they persists across different ways of probing participants that make the predictive nature of brain activity highly salient. Experiment 6 demonstrates that filling in also occurs as a result of importing an indeterminist view of choice when presented with an adapted case of perfect neuro-prediction. We discuss the implications of these findings for studying the psychology and philosophy of free will.

2. Experiment 1

2.1 Method

2.1.1 Participants

One hundred and fifteen people participated (aged 18-74 years, mean age = 36 years, 53 female, 100% reporting English as a native language). Participants were U.S. residents, recruited and tested online using Amazon Mechanical Turk and Qualtrics, and compensated $0.45 for approximately 2-3 minutes of their time. The same basic recruitment and testing procedures were used in all subsequent studies reported in the paper. Repeat participation was prevented. Thirteen participants were excluded from analysis for failing a comprehension question.

2.1.2 Materials and procedure

Participants were randomly assigned to one of two conditions with stimuli taken verbatim from Nahmias et al. (2014: Experiment 1). In the Neuro-Prediction condition, neuroscientists correctly predict that Jill will vote Green for Governor. In the Manipulation condition (in bold below), neuroscientists manipulate Jill into voting Black for Governor (see Nahmias et al., 2014: Appendix). After reading one of these two stories, participants were asked one comprehension question, six test questions, and completed a brief demographic questionnaire. These questions were presented in random order as the text of the story remained visible at the top of the screen:

1. (Comprehension) Which candidate does Jill vote for Governor? [Smith/Green/Black]
2. (Free Will) Jill's choice about who to vote for Governor _____ made freely. [was/was not]
3. (Activity Change) After the pattern of brain activity occurred, could Jill have voted for a different candidate for Governor instead of Green? [Yes/No]

4. (Aware Change) When Jill became aware that she was going to vote for Green [**Black**] as Governor, could she have voted for a different candidate for Governor instead of Green [**Black**]? [Yes/No]
5. (Different Pattern) If the pattern of brain activity which lead Jill to vote for Green [**Black**] as Governor had not occurred but a pattern of brain activity which leads to voting for a different candidate for Governor did occur, Jill would have definitely voted for a different candidate for Governor. [Yes/No]
6. (Brain Aware) Was Jill aware of who she would vote for as Governor when the pattern of brain activity occurred? [Yes/No]
7. (Change Mind) When Jill became aware that she was going to vote for Green [**Black**] as Governor, could she have changed her mind and voted for a different candidate for Governor instead of Green [**Black**]? [Yes/No]

2.2 Results

Assignment to condition affected responses to Free Will, χ^2 (1, N = 102) = 85.90, p < .001, Cramer's V = .918, with 94% of participants in Neuro-Prediction affirming free will and 100% of participants in Manipulation denying free will. Assignment to condition also significantly affected responses to Activity Change, χ^2 (1, N = 102) = 28.33, p < .001, Cramer's V = .527; Aware Change, χ^2 (1, N = 102) = 28.71, p < .001, Cramer's V = .531; Change Mind, χ^2 (1, N = 102) = 36.62, p < .001, Cramer's V = .599; but not in Different Pattern X^2 (1, N = 102) = 0.112, p = .81; and Brain Aware, χ^2 (1, N = 102) = 0.50, p = .53.

To evaluate whether free will attributions intruded on their representation of the scenarios, we reanalyzed responses from those participants who affirmed free will in Neuro-Prediction and those who denied free will in Manipulation conditions (Fig. 1). We find

significant differences in "yes" answers between Neuro-Prediction and Manipulation for Activity Change, 69%/12%, χ^2 (1, n = 98) = 28.85, p < .001, Cramer's V = .543; Aware Change, 81%/26%, χ^2 (1, n = 98) = 28.35, p < .001, Cramer's V = .538; and Change Mind, 84%/21% χ^2 (1, n = 98) = 38.44, p < .001, Cramer's V = .626. There were no differences in Different Pattern, 75%/79%, X^2 (1, n = 98) = .241, p = .624; and Brain Aware, 43%/38%, χ^2 (1, n = 98) = .278, p = .598.

----------Insert Figure 1 about here ----------

To assess the overall rate of intrusion effects in this study, we pooled responses across Activity Change, Aware Change and Change Mind in Neuro-Prediction conditions. The result is that 60% of those who affirmed free will displayed intrusion effects on all three of these probes. The overall rate of intrusion effects differed from chance rates, test proportion =.125, p < .001.

2.3 Discussion

Participants in the Neuro-Prediction case who affirmed free will were significantly more likely to display intrusion effects in Activity Change, Aware Change and Change Mind than those participants who denied free will in the Manipulation case. This suggests that participants in the Neuro-Prediction case who affirm free will represent the scenario in terms of their intuitive metaphysics of free will, rather than according to the explicit details of the story.

3. Experiment 2

Experiment 1 provides evidence that filling-in can occur due to intrusion. These results suggest that people misrepresent instances of perfect neuro-prediction, since agents can do otherwise

after a perfectly predictive brain pattern occurs. To demonstrate that intrusion is the best explanation of that effect, we will now investigate the causal relationship between free will judgments and the tendency to misrepresent neuro-prediction stories. According to this hypothesis, intuitive free will judgments cause people to misrepresent instances of perfect neuro-prediction.

3.1 Method

3.1.1 Participants

One hundred and twenty people participated (aged 18-68 years, mean age = 30 years, 30 females, 97% reporting English as a native language) in this study. Thirty-one participants were excluded from the analysis (17 failed a comprehension question, 14 were repeat participants).

3.1.2 Materials and procedure

Participants were randomly assigned to the Neuro-Prediction or Manipulation conditions used in Experiment 1. Participants were asked the same comprehension question used in Experiment 1, Free Will, and four additional test questions. Given that intrusion effects were displayed on Activity Change, Aware Change and Change Mind in Experiment 1, we again included those questions and added the following question:

> (Possibility) Even though Jill voted for Green as Governor, it was possible for her to decide to vote for a different candidate.

Participants indicated agreement on a 7-pt scale anchored with “strongly disagree”, “disagree”, “somewhat disagree”, “neither agree nor disagree”, “somewhat agree”, “agree” and “strongly agree”. The presentation of each test item was randomized.

3.2 Results

Assignment to condition affected responses to Free Will, $t(87)=10.885$, $p<.001$ $d=2.342$; Activity Change, $t(87)=2.238$, $p<.05$ $d=.492$; Aware Change, $t(87)=3.458$, $p<.01$ $d=.758$; Change Mind, $t(87)=3.616$, $p<.01$ $d=.798$; and Possibility $t(87)=3.754$, $p<.001$, $d=.829$ (see Table 1).

----------Insert Table 1 about here ----------

A mediation analysis was conducted to examine the relationship between Free Will and Possibility judgments (Fig. 2). Following the procedure outlined in Baron and Kenny (1986), we found that a regression model with Condition as a predictor of Possibility was significant, $t(87)=-3.754$, Beta=-.373, $p<.001$, a regression model with Condition as a predictor of Free Will was significant, $t(87)=-10.885$, Beta=-.759, $p<.001$, a regression model with Free Will as a predictor of Possibility was significant, $t(87)=5.945$, Beta=.537, $p<.001$, but that in a multiple regression model with both Condition and Free Will as predictors of Possibility, the effect of Condition on Possibility was no longer significant, $t(86)=.589$, Beta=.082, $p=.558$. Moreover, the reduction in the effect of Condition on Possibility when Free Will was included in the model is significant, $Z=-3.9853$, $p<.001$.

----------Insert Figure 2 about here ----------

Following Iacobucci, Saldanha and Deng (2007), and Rose and Nichols (2013) we also tested the alternative mediation model. A multiple regression model with both Condition and Possibility as

predictors of Free Will showed that Condition significantly predicted Free Will, $t(86)=-9.461$, Beta=-.649, $p<.001$, but that Possibility did not mediate the effect of Condition on Free Will.

3.3 Discussion

We replicated results from Experiment 1 using scalar measures finding that intrusion effects were displayed on Activity Change, Aware Change and Change Mind. We also found an intrusion effect using Possibility as a measure and found that Free Will mediates the effects of Condition on Possibility. In other words, intuitive free will judgments led people to represent the scenario in ways that are inaccurate and inconsistent with a perfectly predictive neuroscience. These results provide evidence that people's intuitive views about agency intrude into the representation of neuroscientific scenarios when making free will judgments.

4. Experiment 3

Given that participants display intrusion effects in the neuro-prediction case used by Nahmias et al., we now want to consider the extent of these effects in other narrative contexts. It might be objected that the present results only persist in neuro-prediction cases involving voting, and other acts naturally expected to correlate with prior commitments or values. It also might be objected that the results are due to overly complex stimuli that do not utilize minimally matched pairs. Experiment 3 addresses these concerns by using simpler stimuli less likely to be associated with values or prior commitments: pushing a button. Our hypothesis is that while the tendency to display intrusion effects may decrease, those who affirm free will in neuro-prediction cases will still display intrusion effects when representing cases involving very simple actions.

4.1 Method

4.1.1 Participants

One hundred and thirty-one people participated (aged 18-67 years, mean age = 33 years, 47 female, 98% reporting English as a native language) in this study. Twelve participants were excluded from the analysis (4 failed a comprehension question, 8 were repeat participants).

4.1.2 Materials and procedure

Participants were again assigned to one of two conditions—Neuro-Prediction or Manipulation—with stimuli featuring a new narrative context and a simpler decision: pushing a button. The vignettes were as follows, with the Manipulation condition marked in bold:

> Recent brain scanning experiments have shown that specific patterns of brain activity predict simple decisions several seconds before people are consciously aware of them. In the experiments, people are asked to press a button, and they can press the button with either their left or right hand. By measuring activity in part of the brain (the motor cortex), scientists can use brain scanners to predict with 100% accuracy that a person will push the button with their left hand before the person is consciously aware of their decision. [**The scientists can even alter a person's decision about which hand to use to push the button by altering the person's brain activity without the person being aware of it**.] ¶[1] Jill is the subject in one of these experiments. She is asked to push the button with either her left or right hand. [Before she is aware of deciding which hand to use, the neuroscientists can see, based on her brain activity, that she will push the button with her left hand. As predicted, Jill pushes the button with her left hand./**However, before Jill is aware of making any decision, the scientists alter Jill's brain activity so that she pushes the button with her left hand.**]

[1] This signifies a paragraph break on participants' screens.

After reading one of these two stories, participants were asked similar questions according to the same procedure used in Experiment 1:

1. (Comprehension) The scientists asked Jill to _____ [press a button/move her hand/move her leg].
2. (Free Will) Jill freely chose to push the button with her left hand. [Yes/No]
3. (Activity Change) After the pattern of brain activity occurred, could Jill have pushed the button with her right hand instead of her left hand? [Yes/No]
4. (Aware Change) When Jill became aware that she was going to push the button with her left hand, could she have pushed the button with her right hand instead of her left hand? [Yes/No]
5. (Different Pattern) If the pattern of brain activity which led Jill to push the button with her left hand had not occurred but a pattern of brain activity which leads to pushing the button with the right hand did occur, Jill would have definitely pushed the button with her right hand. [Yes/No]
6. (Brain Aware) Was Jill aware of which hand she would use to push the button when the pattern of brain activity occurred? [Yes/No]
7. (Change Mind) When Jill became aware that she was going to push the button with her left hand, could she have changed her mind and pushed the button with her right hand instead of her left hand? [Yes/No]

4.2 Results

Assignment to condition affected responses to Free Will, χ^2 (1, N = 119) = 93.33, $p < .001$, Cramer's V = .900, with 97% of participants in Neuro-Prediction affirming free will and 94% of participants in Manipulation denying free will. Assignment to condition also significantly

affected responses to Aware Change, χ^2 (1, N=119) = 7.92, p < .01, Cramer's V = .258; and Change Mind, χ^2 (1, N = 119) = 11.34, p < .01, Cramer's V = .309; but did not affect Activity Change, χ^2 (1, N = 119) = .928, p=.335; Different Pattern χ^2 (1, N = 119) = .690, p = .406; and Brain Aware, χ^2 (1, N = 119) = .948, p = .330.

To test whether the free will attributions that participants made were intruding on their interpretation of the scenarios, we again reanalyze responses from those participants who affirmed free will in Neuro-Prediction and those who denied free will in Manipulation conditions (Fig. 3). We find significant differences in "yes" answers between Neuro-Prediction and Manipulation for Aware Change, 51%/22%, χ^2 (1, n = 113) = 9.92, p < .01, Cramer's V = .296; and for Change Mind, 62%/28%, χ^2 (1, n = 113) = 13.41, p <.001, Cramer's V = .344. There were no differences in Activity Change, 24%/29%, χ^2 (1, n = 113) =.446, p=.495; Different Pattern, 87%/84%, χ^2 (1, n = 113) = .181, p = .671; or Brain Aware, 18%/10%, χ^2 (1, n = 113) = 1.43, p = .232.

----------Insert Figure 3 about here ----------

To assess the overall rate of intrusion effects in this study, we again pool responses across Aware Change and Change Mind in Neuro-Prediction conditions. We find that 49% of those who affirmed free will in Neuro-Prediction displayed intrusion effects on both of these probes and that the overall rate of intrusion effects differed from chance, test proportion = .25, p < .001.

4.3 Discussion

We found that participants continue to display intrusion effects when confronted with a case of perfect neuro-prediction, which replicates the results of Experiment 1. Although decreasing the complexity of the action to the simple pushing of a button narrowed the extent to which intrusion effects occurred, it did not eliminate them. Moreover, we continued to find that those who denied free will in Manipulation were significantly less likely to display intrusion effects.

5. Experiment 4

Having shown that intrusion effects persisted in the button-pressing case, we now wish to investigate whether this result replicates using scalar measures as in Experiment 2, and whether free will judgments continue to mediate the effect of condition on possibility judgments in these circumstances as predicted by the intrusion hypothesis. This is tested in Experiment 4.

5.1 Method

5.1.1 Participants

Sixty-five people participated (aged 19-69 years, mean age = 33 years, 13 female, 97% reporting English as a native language) in this study. Twelve participants were excluded from the analysis (3 failed a comprehension question, 9 were repeat participants).

5.1.2 Materials and procedure

Participants were randomly assigned to one of two conditions—the Neuro-prediction and Manipulation conditions—using the same vignettes as in Experiment 3. Participants were asked the same comprehension question as used in Experiment 3, a test question about free will, and three additional test questions. Given that intrusion effects were displayed on Aware Change and Change Mind in Experiment 3, we again included those questions and added the following question:

(Possibility) Even though Jill pushed the button with her left hand, it was possible for her to decide to push the button with her right hand.

Participants answered on the same agreement scale, and the presentation of each test item was randomized according to the procedure used in Experiment 2.

5.2 Results

Assignment to condition affected responses to Free Will, $t(51)=5.286$, $p<.001$ $d=1.471$; Aware Change, $t(51)=2.316$, $p<.05$ $d=.643$; Change Mind, $t(51)=2.692$, $p<.05$ $d=.745$; and Possibility $t(51)=3.744$, $p<.001$ $d=1.040$ (see Table 2).

----------Insert Table 2 about here ----------

A mediation analysis was conducted to examine the relationship between Free Will and Possibility judgments (Fig. 4). We found that a regression model with Condition as a predictor of Possibility was significant, $t(51)=-3.744$, Beta=-.464, $p<.001$, a regression model with Condition as a predictor of Free Will was significant, $t(51)=-5.286$, Beta=-.595, $p<.001$, a regression model with Free Will as a predictor of Possibility was significant, $t(51)=5.048$, Beta=.577, $p<.001$, but that in a multiple regression model with both Condition and Free Will as predictors of Possibility, the effect of Condition on Possibility was no longer significant, $t(50)=-1.325$, Beta=-.187, $p=.191$. Moreover, the reduction in the effect of Condition on Possibility when Free Will was included in the model is significant, $Z=-2.7634$, $p<.01$.

----------Insert Figure 4 about here ----------

We again also evaluated the alternative model in which Possibility mediates the effect of Condition on Free Will. A multiple regression model with both Condition and Possibility as predictors of Free Will showed that Condition significantly predicted Free Will, $t(50)=-3.583$, Beta=-.417, $p<.001$ but that Possibility did not mediate the effect of Condition on Free Will.

5.3 Discussion

We replicated the results from Experiment 3, finding that intrusion effects were displayed on Aware Change and Change Mind. We also found intrusion effects for Possibility, and found evidence that Free Will mediates the effects of Condition on Possibility. This occurred even though the decision in this case was decreased in complexity, from voting to the simple pushing of a button. Free will judgments again led people to represent the scenario in ways that are inconsistent with a perfectly predictive neuroscience, such as that it's possible for Jill to push the button with her right hand. These results continue to provide evidence that intrusion of indeterministic metaphysics influences the representation of neuroscientific scenarios.

6. Experiment 5

Participants in prior experiments judge that brain activity may not be perfectly predictive of behavior, indicating that intrusion has occurred. One objection to this however is that participants are not attending to the fact that these behaviors happen despite the same pattern of predictive brain activity. For instance, perhaps they answer that agents in the stories could act otherwise because doing so would result in another pattern of activity than the one neuroscientists originally detected in the story, which would have perfectly predicted that

behavior. This reading is complex and would still violate the conditions of the thought experiment. Nonetheless, in this experiment we investigate whether participants think that agents can act otherwise, despite the *same* pattern of perfectly predictive brain activity to the contrary.

6.1 Method

6.1.1 Participants

Sixty-eight people participated (aged 20-56 years, mean age = 32 years, 23 female, 93% reporting English as a native language) in this study. No participants were removed from analysis.

6.1.2 Materials and procedure

Participants were randomly assigned to either the Neuro-Prediction or Manipulation condition used in Experiment 4. After seeing one of these conditions participants were asked the following four questions:

> (Free Will) Jill freely chose to push the button with her left hand.
>
> (After State) After activation of the brain state, Jill could still choose to use either hand to push the button.
>
> (Final Prediction) Once the scientists make their final prediction, Jill must use her left hand to push the button.
>
> (Agent General) Someone with the exact same brain activity as Jill could still decide to push the button with their right hand.

Participants evaluated these items on the same agreement scale, and the presentation of each test item was randomized according to the procedure used in Experiment 2.

6.2 Results

Assignment to condition affected responses to Free Will, $t(66) = 6.77$, $p < .001$, $d=1.632$; After State, $t(66) = 3.38$, $p<.001$, $d=.822$; Final Prediction, $t(66) = -3.76$, $p<.001$, $d=.923$; Agent General $t(66) = 4.90$, $p<.001$, $d=1.189$, (see Table 3).

----------Insert Table 3 about here ----------

A mediation analysis was conducted to examine the relationship between Free Will and After State judgments (Fig. 5). We found that a regression model with Condition as a predictor of After State was significant, $t(66)=-3.38$, Beta=-.384, $p<.001$, a regression model with Condition as a predictor of Free Will was significant, $t(66)=-6.77$, Beta=-.640, $p <.001$, a regression model with Free Will as a predictor of After State was significant, $t(66)=5.68$, Beta=.573 $p<.001$, but that in a multiple regression model with both Condition and Free Will as predictors of After State, the effect of Condition on After State was no longer significant, $t(65)=-.223$, Beta=-.029, $p=.824$. Moreover, the reduction in the effect of Condition on After State when Free Will was included in the model is significant, $Z=-3.5344$, $p<.001$.

----------Insert Figure 5 about here ----------

We also evaluated the alternative model in which After State mediates the effect of Condition on Free Will. A multiple regression model with both Condition and After State as predictors of Free

Will showed that Condition significantly predicted Free Will, $t(65)=-5.38$, Beta=-.493, $p<.001$, but that After State did not mediate the effect of Condition on Free Will.

A mediation analysis was conducted to examine the relationship between Free Will and Agent General judgments (Fig. 6). We found that a regression model with Condition as a predictor of Agent General was significant, $t(66)=-4.90$, Beta=-.516, $p<.001$, a regression model with Condition as a predictor of Free Will was significant, $t(66)=-6.77$, Beta=-.640, $p<.001$, a regression model with Free Will as a predictor of Agent General was significant, $t(66)=5.64$, Beta=.570, $p<.001$, but that in a multiple regression model with both Condition and Free Will as predictors of Agent General, the effect of Condition on Agent General was no longer significant, $t(65)=-1.99$, Beta=-.256, $p=.051$. Moreover, the reduction in the effect of Condition on Agent General when Free Will was included in the model is significant, $Z=-2.8347$, $p<.001$.

----------Insert Figure 6 about here ----------

We again also evaluated the alternative model in which Agent General mediates the effect of Condition on Free Will. A multiple regression model with both Condition and Agent General as predictors of Free Will showed that Condition significantly predicted Free Will, $t(65)=-4.55$, Beta=-.471, $p<.001$, but that Agent General did not mediate the effect of Condition on Free Will.

Lastly, we found that a regression model with Condition as a predictor of Final Prediction was significant, $t(66)=3.76$, Beta=-.420, $p<.001$, but that a model with Free Will as a predictor of Final Prediction was not significant $t(66)=-1.92$, Beta=-.230, $p=.06$. Moreover, the effect of

Condition on Final Prediction was not mediated by Free Will nor was the effect of Condition on Free Will mediated by Final Prediction.

6.3 Discussion

The results from Experiment 5 continue to suggest that people fill in neuro-prediction stories. Specifically, they indicate that agents could "choose to use either hand" and "push the button with their right hand" in the presence of brain patterns that were supposed to perfectly predict that "she will push the button with her left hand".

7. Experiment 6

We've provided a range of evidence that participants fill in details of neuro-prediction scenarios. Experiments 1-5 suggest that this leads participants to misrepresent perfect neuro-prediction scenarios through intrusion. But recall from Section 1 that we distinguished between two ways in which participants might fill in: intruding or importing. We now consider whether importing occurs in response to these kinds of cases, and whether this would still undermine the inference that people are comfortable with neuro-prediction.

Our strategy will be to investigate importing in the context of a rollback case. Rollback cases are scenarios in which the universe is recreated and everything unfolds as it did in the original universe (Nahmias et al., 2006). These cases are theoretically interesting because while rollback is intended to make determinism "as salient to participants as possible without being misleading" (Nahmias et al., 2006, p. 37) they are still consistent with indeterminism. For this reason, they also provide for a good test of whether importing of indeterministic metaphysics occurs in neuro-prediction cases.

In addition to testing for importing, the rollback cases were also designed to address another potential objection. The objection is that people affirm "free will" statements in perfect neuro-prediction cases because they interpret these statements to mean that the relevant agent was not coerced.[2] Assuming that the concept of coercion is compatible with both determinist and indeterminist metaphysics, this "no coercion" or lightweight reading of "free will" challenges the claim that people are filling-in by either intruding or importing a particular metaphysics. We test this possibility in rollback cases by using an adapted probe to measure indeterminism through the assignment of probabilities. This circumvents the "no coercion" interpretation because the question is simply about the probability of the outcome, not whether the agent had "free will".

7.1 Method

7.1.1 Participants

One hundred and ten people participated (aged 19-65 years, mean age = 32 years, 38 female, 95% reporting English as a native language) in this study. Five participants were excluded from the analysis (3 failed a comprehension check, 2 were repeat participants).

7.1.2 Materials and procedure

Participants were randomly assigned to either the Neuro-Prediction or Manipulation condition used in Experiment 4, but with the following addition to the end of the text (manipulation condition marked in bold):

> Now imagine that the universe is recreated with everything the same right before Jill decides which hand to push the button with. The scientists will conduct the same

---

[2] We thank a referee for raising this point.

> procedures again that will allow them to predict **[interfere with]** which hand Jill will push the button with. In this recreated universe, Jill will decide again which hand she will use to push the button.

After seeing one of these conditions participants were asked:

> (Chance) In the recreated universe, what do you think the chances are that Jill will end up pushing the button with her right hand?

In response, participants were asked to enter a value ranging from 0% to 100%.

7.2 Results

Mean percentage ratings to the *Chance* item were 33% in Neuro-Prediction and 54% in Manipulation. These responses were statistically different, $t(103)=-3.153$, $p<.01$. But the key question is whether people are interpreting the scenario deterministically or indeterministically and, more specifically, whether they are more inclined to interpret the Neuro-Prediction scenario indeterministically. We created an *Indeterminism* measure by coding *Chance* responses to test for these interpretations. On this measure, two responses indicate a deterministic interpretation: a response of 0% indicates that it is determined that Jill will not end up using her right hand to push the button (i.e., there is no chance that she will use her right hand) while a response of 100% indicates that it is determined that Jill will use her right hand to use the button (i.e., there is no chance that she will not use her right hand). All other responses indicate the indeterministic response that there is at least some chance that Jill could use either hand to push the button in the recreated universe. Assignment to condition significantly affected *Indeterminist* scores, χ^2 (1, N = 105) = 4.82, $p < .05$, Cramer's V = .214, with 69% of participants in Neuro-Prediction giving an indeterministic response, and 48% of participants in Manipulation giving an indeterministic response.

7.3 Discussion

The results from Experiment 6 suggest that people fill in rollback cases of perfect neuro-prediction. Participants are significantly more likely to import indeterminism into the prefect neuro-prediction scenario in contrast to a manipulation case. Participants import indeterminist free will into rollback variations of neuro-prediction stories even when deterministic readings of these cases are made highly salient, which while technically consistent with rollback, continues to undermine the inference that people are comfortable with perfect neuro-prediction. These results also undermine the "no coercion" interpretation of "free will" statements. The rollback scenario asked only about the probability of an outcome and did not invite a contrast with coercion. Despite this, we find that participants assign non-extreme probabilities when considering whether Jill will push the button with her other hand in the rollback universe, which suggests that people import indeterminism into the representation of the scenario.

8. General discussion

Intuitive metaphysics shapes cultural transmission (Sperber, 1994), scientific understanding (de Cruz & de Smedt, 2007), and the elaboration of religious representations (Boyer, 1994). In some cases, intuitive metaphysics is used to systematically fill in narratives (Barrett & Keil, 2006). Our experiments expand on this effect and demonstrate that it also extends to agentive evaluations of free will. We tested for intruding and importing effects in recent work by Nahmias, Shepard, and Reuter featuring narrative cases of futuristic neuroscientific prediction. Like Nahmias and colleagues, we found that participants overwhelmingly attribute free will in cases of perfect neuro-prediction. However we also found that participants' intuitive metaphysics of free will intrudes into their representation of these perfect neuro-prediction scenarios. Though the explicit description in the scenarios of action initiation being generated before conscious

awareness and the description of the prediction being 100% accurate implies that the agent in the scenarios could *not* have done otherwise, we found that participants who affirmed free will tended to say that the agent could have changed her mind after becoming aware of what she was going to do. We also found that affirming free will caused participants to say that agents could act otherwise despite an activation of a perfectly predictive brain state. Moreover, we provided evidence that even when asked about mere probabilities in rollback cases, people continue to import an indeterministic view of choice into their interpretation of the scenario. The presence of intruding and importing effects suggest that people are imposing an indeterminist notion of free will onto the situation, despite the fact that the situation is explicitly described in terms of perfect predictability. Thus we doubt that people are broadly comfortable with the idea of perfect neuro-prediction or that it is fully compatible with commonsense notions of free will.

The presence of these effects in ordinary judgments of free will has implications for experimental work on free will. One key dispute is whether the ordinary view of free will is such that people view causal determinism as compatible with free will and moral responsibility or whether people view causal determinism as incompatible with free will and moral responsibility. Though this research has largely suggested that participants tend to think that free will and moral responsibility are incompatible with causal determinism (e.g., Nichols & Knobe, 2007; Nichols, 2012; Rose & Nichols, 2013; Sarkissian et al., 2010), some research has been taken to suggest that participants tend to think free will and moral responsibility are compatible with causal determinism (e.g., Murray & Nahmias, 2014; Nahmias et al., 2006). Assuming our findings extend to cases where causal determinism is explicitly stipulated, those who seek to provide support for intuitive compatibilism may be faced with a considerable difficulty. In our studies—and unlike the mistaken recollections found by Barrett and Keil—we found that filling in effects

occurred even though the text of the vignettes remained at the top of the screen during testing. Moreover, filling in effects occurred in cases involving very simple actions, as in Experiment 3, suggesting that it may be difficult to eliminate them entirely. Taken together, it may require considerable efforts to ensure that filling in effects are eliminated. But insofar as it takes considerable effort to eliminate the presence of filling in effects such as intruding and importing in the context of cases which explicitly stipulate causal determinism, it's unclear whether research which finds support for compatibilism after successfully eliminating importing effects would provide convincing support for the view that people are intuitive compatibilists.

With developing scientific knowledge we're confronted with the question of how this knowledge will interact with humanistic concerns. Though more advanced scientific knowledge may challenge such concerns, it might not displace them. For instance, though the theory of evolution arguably challenges a perspective on reality whereby nature is infused with agency and purpose (e.g., Bloom, 2007; Kelemen, 2012), it doesn't entirely displace that perspective. Work in developmental psychology suggests that young children are strongly inclined toward viewing nature as being infused with agency and purpose (e.g., Kelemen, 1999a,1999b, 2004; Kelemen & DiYanni, 2005). This default perspective on reality is not entirely displaced with a more mature, scientifically informed perspective on the world but rather is masked in adulthood (e.g., Kelemen & Rosset, 2009; Kelemen, Rottman & Seston, 2013; Lombrozo, Kelemen & Zaitchik, 2007). A study by Kelemen et al. (2013), for instance, found that laypeople, professional scientists and professionals in humanities each accept illegitimate teleological statements (e.g., "The sun radiates heat because warmth nurtures life.") when placed under time pressure. Moreover, they found that background scientific knowledge did not predict the extent to which participants were willing to accept teleological statements (see also Rose, 2015 for further discussion). Far from

displacing this default perspective on the world, some work suggests that this perspective interferes with the acquisition of scientific knowledge, serving as one of the main obstacles to acquiring an adequate understanding of natural selection (see Galli & Meinardi, 2011; and Kelemen, 2012 for an overview). For instance, students tend to think that "a personified 'Mother Nature'" responded to animals' functional needs by "generating or conferring the functional part with a view to preserving the animal's survival" (Kelemen, 2012, p. 71; see also e.g., Gregory, 2009; Kampourakis & Zogza, 2008; and Moore et al., 2002).

Our finding that intuitive metaphysics both intrudes and is imported into the representation of neuroscientific scenarios is perhaps best viewed as a further demonstration of the resilience of humanistic concerns in social cognition. While people may reflectively endorse the theory of evolution, the acquisition of this scientific knowledge does not displace a default perspective in which the world is infused with agency and purpose. Similarly, our scientific knowledge may become so advanced that perfect neuroscientific predictions can be given for all of human behavior. Though people may come to reflectively accept that neuroscience could perfectly predict behavior, arguably not even the highest level of neuroscientific knowledge is enough to displace the natural default view of indeterministic human decision-making. Rather just as the default view of nature as being infused with agency and purpose continues to reside alongside a reflective endorsement of the theory of evolution, the default view of indeterminist free will may reside along the reflective endorsement of perfect neuro-prediction.

Acknowledgements

We thank Alisabeth Ayars, Josh Knobe, Myrto Mylopoulos, Eddy Nahmias, Gualtiero Piccinini, Jason Shepherd, John Turri, two anonymous reviewers, and audiences at the Society for Philosophy and Psychology for helpful comments on earlier versions of this paper. This research was supported by a Banting Fellowship awarded through the Social Sciences and Humanities Research Council of Canada.

References

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

Barrett, J. L., & Keil, F. C. (1996). Conceptualizing a Nonnatural Entity: Anthropomorphism in God Concepts. *Cognitive Psychology* 31, 219--247.

Bloom, P. (2007). Religion is Natural. *Developmental Science,* 10, 147--151.

Boyer, P. (1994). *The naturalness of religious ideas: A cognitive theory of religion*. Berkeley: University of California Press.

Coyne, J. (2012). Why you don't really have free will. *USA Today* (01/01/12).

De Cruz, H., & De Smedt, J. (2007). The role of intuitive ontologies in scientific understanding–the case of human evolution. *Biology & Philosophy*, 22 (3), 351--368.

Galli, L., & Meinardi, E. (2011). The Role of Teleological Thinking in Learning the Darwinian Model of Evolution. *Evolution Education Outreach,* 4, 145--152.

Greene, J. D., & Cohen J. D. (2004) For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London B*, 359, 1775--1785.

Gregory, T. R. (2009). Understanding Natural Selection: Essential Concepts and Common misconceptions. *Evolution: Education and Outreach*, 2, 156--175.

Harris, S. (2012). *Free will*. New York: Free Press.

Iacobucci, D., Saldanha, N., and Deng, X. (2007). A Mediation on Mediation: Evidence That

Structural Equation Models Perform Better Than Regressions. *Journal of Consumer Psychology*, 17 (2), 140--154.

Kampourakis, K. & Zogza, V. (2008). Students' Intuitive Explanations of the Causes of Homologies and Adaptations. *Science and Education*, 17, 27--47.

Kelemen, D. (1999a). The Scope of Teleological Thinking in Preschool Children. *Cognition*, 70, 241--272.

Kelemen, D. (1999b). Why are Rocks Pointy? Children's Preference for Teleological Explanations of the Natural World. *Developmental Psychology*, 35, 1440--1452.

Kelemen, D. (2004). Are Children "Intuitive Theists"? Reasoning about Purpose and Design in Nature. *Psychological Science,* 15, 295--301.

Kelemen, D. (2012). Teleological Minds: How Natural Intuitions About Agency and Purpose Influence Learning About Evolution. In K. S. Rosengren, S. K. Brem, E. M. Evans & G.

M. Sinatra (Eds.), *Evolution challenges: Integrating research and practice in teaching and learning about evolution*. 66–92. Oxford: Oxford University Press.

Kelemen, D. & DiYanni, C. (2005). Intuitions about Origins: Purpose and Intelligent Design in Children's Reasoning about Nature. *Journal of Cognition and Development,* 6, 3--31.

Kelemen, D., & Rosset, E. (2009). The Human Function Compunction: Teleological Explanation in Adults. *Cognition*, 111, 138--143.

Kelemen, D., Rottman, J. & Seston, R. (2013). Professional Physical Scientists Display Tenacious Teleological Tendencies: Purpose-Based Reasoning as a Cognitive Default. *Journal of Experimental Psychology: General*, 142, 1074--1083.

Lombrozo, T., Kelemen, D., & Zaitchik, D. (2007). Inferring Design: Evidence for a Preference for Teleological Explanation in Patients with Alzheimer's Disease. *Psychological Science*, 18, 999--1006

Monroe, A. E., Dillon, K. D., & Malle, B. F. (2014). Bringing free will down to Earth: People's psychological concept of free will and its role in moral judgment. *Consciousness and Cognition, 27,* 100--108.

Moore, R., Mitchell, G., Bally, R., Inglis, M., Day, J., & Jacobs, D. (2002). Undergraduates Understanding of Evolution: Ascription of Agency as a Problem for Student Learning. *Journal of Biological Education*, 36, 65--71.

Murray, D., & Nahmias, E. (2014). Explaining away incompatibilist intuitions. *Philosophy and Phenomenological Research*, 88, 434--467.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2006). Is incompatibilism intuitive? *Philosophy and Phenomenological Research*, 73, 28--53.

Nahmias, E., Shepard, J., & Reuter, S. (2014). It's OK if 'my brain made me do it': People's intuitions about free will and neuroscientific prediction. *Cognition,* 133, 502--516.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs,* 41, 663--685.

Nichols, S. (2012). The Indeterminist Intuition. *The Monist*, 95 (2), 290--307.

Pereboom, D. (2001). *Living without free will*. Cambridge: Cambridge University Press.

Rose, D. (2015). Persistence through function preservation. *Synthese*, 192, 97--146.

Rose, D., & Nichols, S. (2013). The lesson of bypassing. *Review of Philosophy and Psychology* 4, 599--619.

Sarkissian, H., Chatterjee, A., De Brigard, F., Knobe, J., Nichols, S., & Sirker, S. (2010). Is belief in free will a cultural universal? *Mind & Language* 35, 346--358.

Shepherd, J. (2012). Free will and consciousness: Experimental studies. *Consciousness and cognition*, 21 (2), 915--927.

Sperber, D. (1994). The modularity of thought and the epidemiology of representations. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture*, 39--67. Cambridge University Press.

Strawson, G. (1986). *Freedom and belief*. Oxford: Clarendon Press.

Stillman, T., Baumeister, R., & Mele, A. (2011). Free will in everyday life: Autobiographical accounts of free and unfree action. *Philosophical Psychology* 24 (3), 381--394.

Turri, J. (2015). Exceptionalist naturalism: Human agency and the causal order. Unpublished Manuscript.

Turri, J., Rose, D., & Buckwalter, W. (2015). Choosing and refusing: Doxastic voluntarism and folk psychology. Unpublished Manuscript.

van Inwagen, P. (2000). Free Will Remains a Mystery. *Philosophical Perspectives,* 14, 1--20.

Table 1

Means and standard deviations for measures in Neuro-Prediction and Manipulation conditions in Experiment 2.

	Neuro-Prediction	Manipulation
Free Will	5.71(1.28)	2.47(1.48)
Activity Change	4.42(1.95)	3.47(1.91)
Aware Change	4.73(1.82)	3.35(1.82)
Change Mind	4.84(1.87)	3.38(1.79)
Possibility	4.87(1.87)	3.38(1.72)

Table 2

Means and standard deviations for measures in Neuro-Prediction and Manipulation conditions in Experiment 4.

	Neuro-Prediction	Manipulation
Free Will	5.80(1.44)	3.32(1.90)
Aware Change	4.80(1.80)	3.60(1.93)
Change Mind	4.68(1.79)	3.32(1.86)
Possibility	5.52(1.73)	3.53(2.08)

Table 3

Means and standard deviations for measures in Neuro-Prediction and Manipulation conditions in Experiment 5.

	Neuro-Prediction	Manipulation
Free Will	5.94 (1.47)	3.00 (2.08)
After State	4.63 (2.07)	3.00 (1.89)
Final Prediction	3.71 (2.32)	5.55 (1.60)
Agent General	5.23 (1.90)	3.12 (1.64)

Figures and Legends

Fig. 1: Experiment 1. Percentage of participants who affirmed free will in Neuro-Prediction and denied free will in Manipulation displaying intrusion effects on each measure.

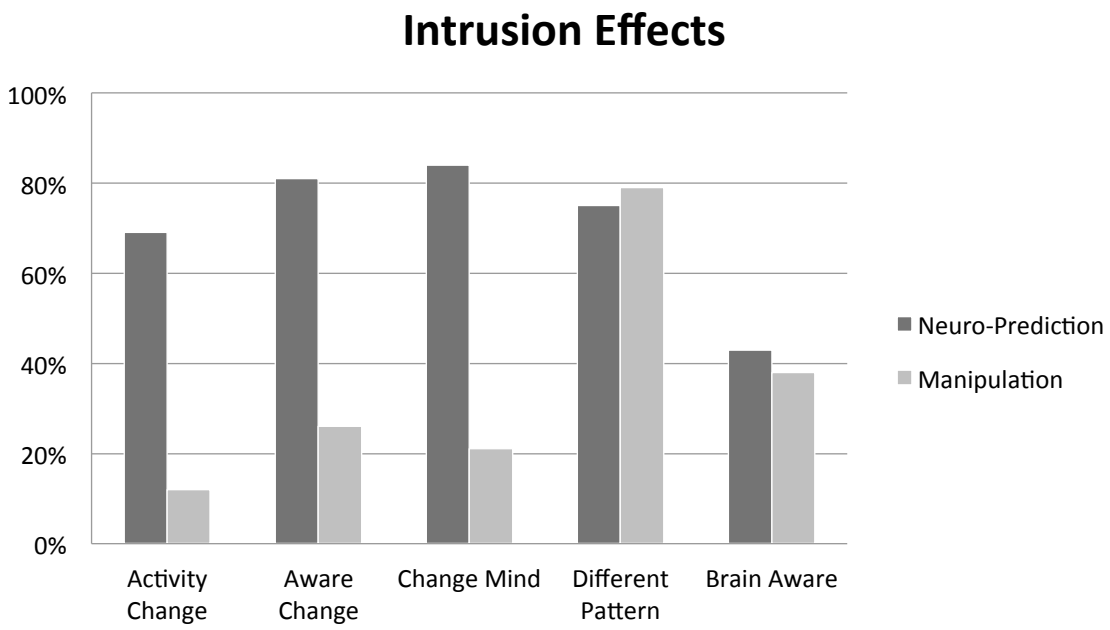

Fig. 2: Experiment 2. Standardized regression coefficients for the relationship between Condition and Possibility mediated by Free Will.

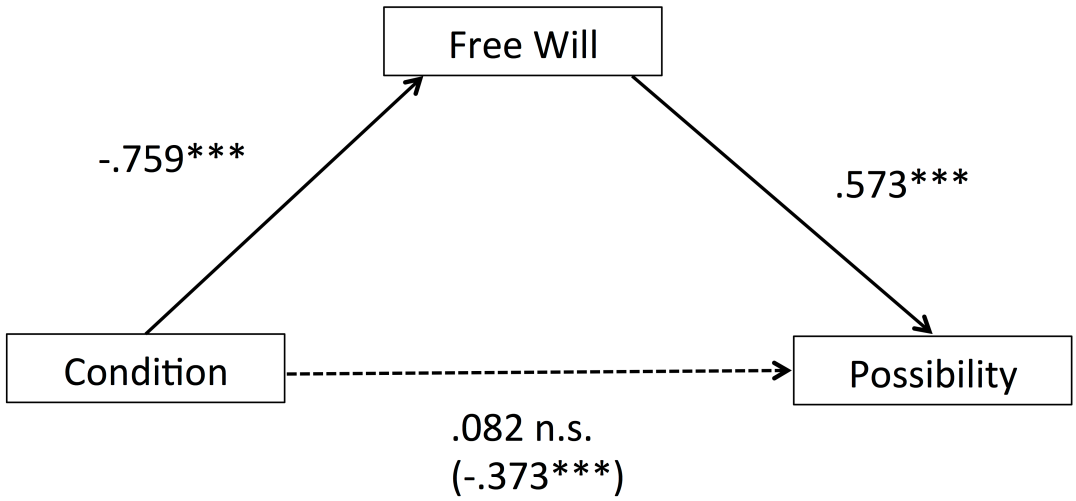

Fig. 3: Experiment 3. Percentage of participants who affirmed free will in Neuro-Prediction and denied free will in Manipulation displaying intrusion effects on each measure.

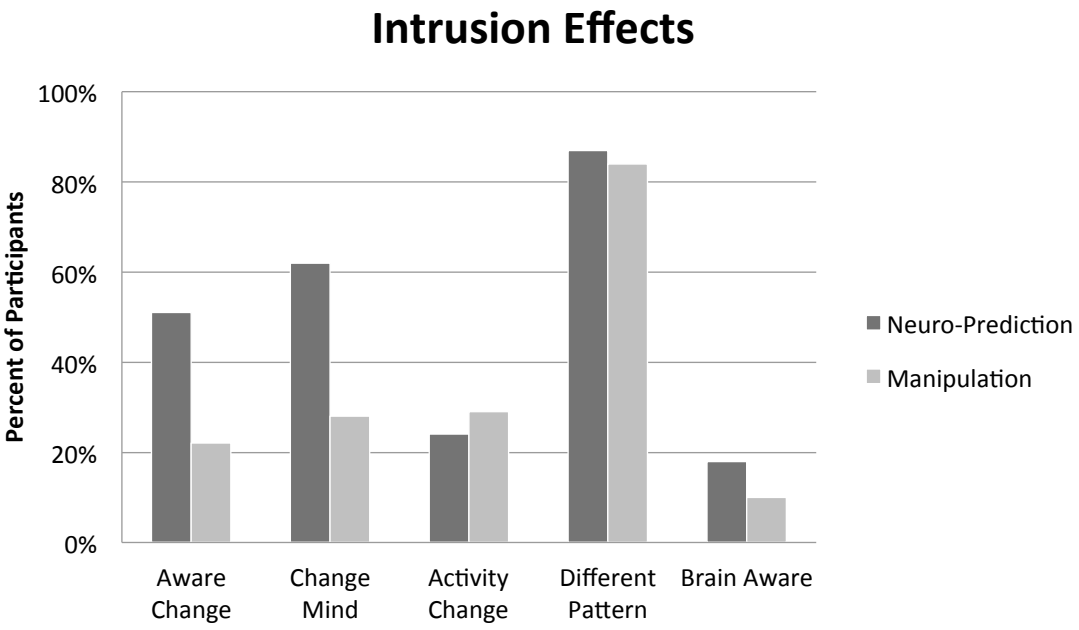

Fig. 4: Experiment 4. Standardized regression coefficients for the relationship between Condition and Possibility mediated by Free Will.

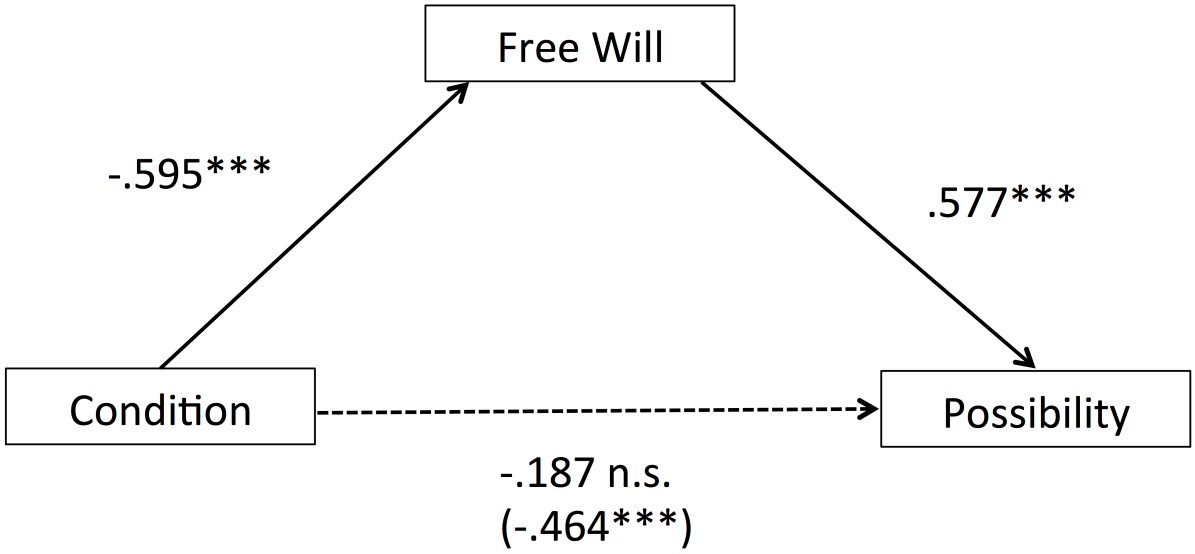

Fig. 5: Experiment 5. Standardized regression coefficients for the relationship between Condition and After State mediated by Free Will.

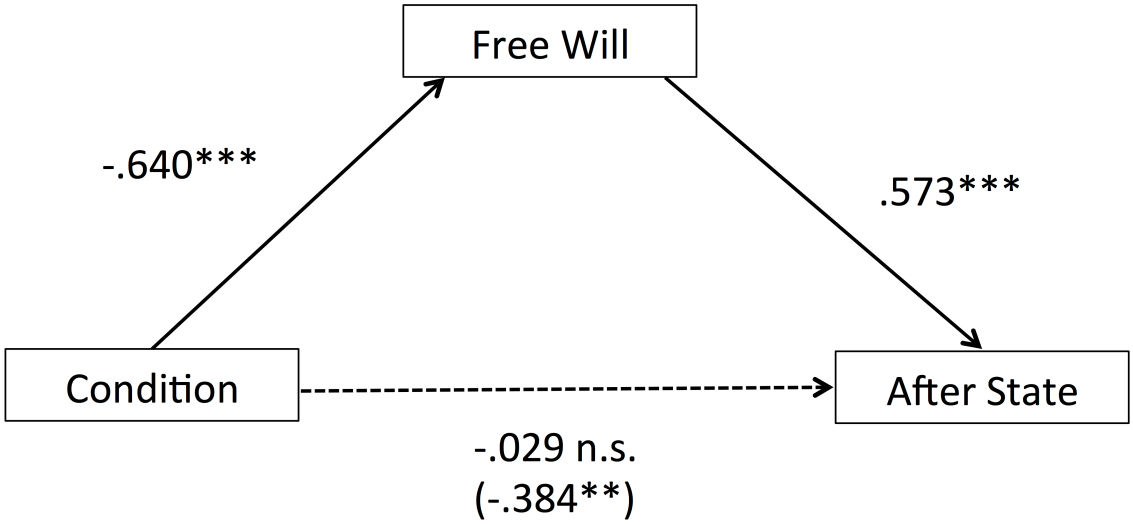

Fig. 6: Experiment 5. Standardized regression coefficients for the relationship between Condition and Agent General mediated by Free Will.