Natural Compatibilism, Indeterminism, and Intrusive Metaphysics

Thomas Nadelhoffer, David Rose, Wesley Buckwalter, & Shaun Nichols

[forthcoming in *Cognitive Science*]

1. Natural Compatibilism vs. Natural Incompatibilism

According to proponents of *natural incompatibilism*, most people judge that determinism[1] threatens our traditional picture of human agency (see, e.g., Cover & O'Leary-Hawthorne, 1996; Ekstrom, 2000; Kane, 1999; Nichols, 2015; O'Connor, 2000; Pereboom, 2001; Searle, 1984; Smilansky, 2003; Strawson, 1986). According to proponents of *natural compatibilism*, most people intuitively believe that free will and/or moral responsibility are compatible with determinism (see, e.g., Ayer, 1954; Dennett, 1984; Fischer & Ravizza, 1998; Lycan, 2003; Nahmias, 2014; Nowell-Smith, 1949; Stace, 1952; Wolf, 1990). Given the ubiquity of these competing appeals to common sense in the free will literature, it is important to determine which is right. But until experimental philosophers started exploring the relevant folk intuitions in a controlled and systematic way, there was a dearth of evidence concerning which view has empirical support.

In some of the earliest work on this front, Nahmias, Morris, Nadelhoffer, and Turner (2005; 2006) used three different descriptions of determinism and found that a significant

[1] Determinism is the view that every event (X) is necessitated by past events (P) together with the laws of nature (L). More formally, it has the following form: Necessarily (if PL, then X). If determinism is true, then at any moment, given the past and the laws, there is one and only one possible outcome. Indeterminism is just the denial of determinism. If indeterminism is true, at any given moment, more than one outcome is possible even if you keep the past and the laws fixed.

majority of participants (typically 65%-85%) judged that agents in deterministic scenarios act of their own free will and are morally responsible. Consider, for instance, their supercomputer scenario, whereby a supercomputer can predict the future with 100% accuracy based on deductions from the "laws of nature and the current state of everything in the world" (Nahmias et al., 2005, 559).[2] Despite the fact that determinism was stipulated in the scenario, 76% of participants judged that Jeremy robbed the bank of his own free will and 83% judged that he was morally responsible. These findings were consistent across cases involving neutral actions (e.g., going jogging), positively valenced actions (e.g., saving a child from a burning building), and negatively valenced actions (e.g., robbing a bank). Nahmias and colleagues also replicated these effects in two additional studies utilizing further deterministic scenarios and concluded that the case for natural compatibilism was strong (2006).

Nichols and Knobe (2007) ran some follow-up studies to explore the psychological mechanisms that generate intuitions about moral responsibility.[3] Participants were randomly assigned to either an abstract condition that describes both a deterministic Universe (A) and an indeterministic Universe (B) or a concrete condition that describes these two universes but also describes a person in Universe A, Bill, who murders his wife and family to be with his secretary. Participants were first asked which one of these two universes was more like their own. Nearly all participants (90%) answered 'Universe B'. Then, participants in the abstract condition were asked whether it was possible for a person in Universe A to be "fully responsible for their actions."

[2] For complete details concerning this scenario, see §2.

[3] By focusing narrowly on moral responsibility, Nichols and Knobe (2007) couldn't speak directly to people's free will beliefs. While it is intuitive that people's beliefs about free will and moral responsibility should roughly track one another, there is some recent evidence that the two can come apart—e.g., Fidgor and Phelan (2015).

Participants in the Concrete Condition were asked "Is Bill fully morally responsible for killing his wife and children?" Whereas 72% of participants gave the *compatibilist* response that Bill *is* fully morally responsible in Universe A in the concrete condition, in the abstract condition 84% gave the *incompatibilist* response that it is *not* possible for people in Universe A to be fully morally responsible. These findings appear to challenge the claim that people's intuitions are *robustly* compatibilist. Instead, whether people are inclined to give compatibilist answers may depend less on the presence (or absence) of determinism and more on the moral features of the vignettes and questions. Whereas people tend to display compatibilist leanings when asked to make judgments concerning the responsibility of specific agents, when they are asked instead to think about responsibility in the abstract, their intuitions trend towards incompatibilism.

Debate about whether natural compatibilism or natural incompatibilism is best supported by the extant findings continues (see Bear & Knobe, 2016; Bourgeois-Gironde, Cova, Bertoux, & Dubois, 2012; Cova & Kitano, 2014; Deery, Davis, & Carey, 2014; Feltz, Cokely, & Nadelhoffer, 2009; Feltz & Milan, 2013; Mandelbaum & Ripley, 2012; May, 2014; Murray & Nahmias, 2014; Nadelhoffer et al., 2014; Nadelhoffer, Murray, & Murray, 2019; Nadelhoffer, Yin, & Graves, 2019; Nahmias, 2006; 2011; Nahmias, Coates, & Kvaran, 2007; Nahmias & Murray, 2011; Nichols, 2006a; Rose, Buckwalter, & Nichols, 2017; Rose & Nichols, 2013; Sinnott-Armstrong, 2008; Turri, 2017a; 2017b; Woolfolk, Doris, & Darley, 2006). Most of the research that has explored this issue has utilized vignette-based designs whereby participants are presented with deterministic scenarios and then asked whether the agents in the scenarios are free and responsible.

But in order for the results to provide support for natural compatibilism, it is imperative that participants have adequately understood and held fixed the determinism that is built into the scenarios. After all, a compatibilist is someone who thinks that we can be free and responsible *even if determinism is true*. The "even if" part requires that participants adequately understand the implications of determinism. Thus, from the standpoint of experimental design, putting natural compatibilism to the test requires that researchers ensure that participants' intuitions are sufficiently responsive to the deterministic nature of the scenarios.[4]

Let's call this the tracking problem. There are at least three ways participants could fail to adequately track determinism. First, they could fail to understand or comprehend the determinism altogether—e.g., they could incorrectly identify the scenario as indeterministic. Second, they could correctly identify the scenario as deterministic but fail to comprehend the implications of determinism. Third, they could comprehend the determinism and its implications but fail to hold this fixed in their mind while thinking about the scenario and the follow up questions. In each of these cases, this would represent a failure to do what the experimenter asked—namely, make judgments about agency and responsibility in the face of determinism. There are multiple reasons why someone might fail to track determinism—motivated cognition, researcher effects from bad materials, or failure of acceptance, belief, imagination, inference, or reading comprehension.

The depth of the tracking problem was recently highlighted by Rose et al. (2017), who took as their starting point the evidence that most people do not believe that human actions are

[4] This is not to suggest that natural incompatibilists don't have a similar burden. After all, they, too, have to ensure that people are not misunderstanding the deterministic features of the scenarios. The work on so-called "bypassing intuitions" suggests that it is possible for people who give seeming incompatibilist responses fail to understand the implications of determinism (Murray & Nahmias, 2014; Nahmias & Murray, 2011; cf. Chan, Deutsch, & Nichols, 2016; Rose & Nichols, 2013). Because our present focus is on the evidence that has been used to support natural compatibilism, we are not going to investigate the attempts that have been made to explain away the support for natural incompatibilism.

deterministically caused (Monroe, Dillon, & Malle, 2014; Nichols, 2012; Nichols & Knobe, 2007; Rose & Nichols, 2013; Sarkissian et al., 2010; Stillman, Baumeister, & Mele, 2011). Given that folk metaphysics about human agency is largely indeterministic, Rose and colleagues suggested that this may interfere with people's ability to understand and hold fixed the deterministic features of scenarios. After all, whenever one reads a fictional scenario, some details must be filled in by the reader. Might it be that people tend to fill in indeterministic details when it comes to deterministic scenarios? To shed light on this issue, Rose and colleagues drew the following distinction between 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 (2017, 484).

For present purposes, intrusion is the more germane phenomenon. For if people's metaphysical indeterminism intrudes on how they think through deterministic scenarios, we won't be able to read anything off about whether their intuitions provide support for natural compatibilism.

So, in a series of studies, Rose and colleagues used the neuro-deterministic vignettes from Nahmias et al. (2014) with the addition of some more fine-grained comprehension questions designed specifically to detect the potential influence of intrusive metaphysics. They found indeterministic intrusion effects across six studies which suggests that many participants who appeared, on the surface, to give compatibilist answers were failing to adequately track the determinism built into the scenarios. Rose and colleagues concluded that to the extent that natural compatibilists have failed to adequately address the tracking problem, they have failed to provide evidence for their preferred theory about the nature of ordinary intuitions about free will and responsibility.

In what follows, we extend the research by Rose et al. (2017) by revisiting two of the classic scenarios that have been used to support natural compatibilism—namely, the supercomputer and rollback scenarios from Nahmias et al. (2005; 2006). If it turns out that intrusion effects are prevalent when using these scenarios, this would potentially undercut two prominent findings for natural compatibilism. If participants are failing to track the deterministic features of these scenarios, then it cannot be inferred whether or not they find free will and responsibility to be compatible with determinism. Our prediction was that this is what we would find. This was borne out by our findings which show that indeterministic intrusion is common in people who judge that free will and responsibility are compatible with determinism in classic scenarios. As we will argue, this challenges some of the key evidence used to support natural compatibilism.

2. Current Research

The purpose of this research is to test for indeterministic intrusion effects while controlling for factors that have arisen in the literature on folk intuitions about free will and responsibility. First, given the influence of the work by Nahmias et al. (2005; 2006), we wanted to base our study on their aforementioned supercomputer and rollback scenarios. Second, we wanted to include scenarios involving negatively valenced actions as well as positively valenced actions since past research suggests that the effect created by negatively valenced actions may have an outsized influence on people’s intuitions (Nichols & Knobe 2007). Third, and relatedly, we wanted to include both concrete and abstract scenarios since past research suggests that there is an asymmetry in how people respond to each type of case (Nichols & Knobe 2007; Roskies & Nichols 2008). Finally, following Shepherd (2015) and Turri (2017c), we wanted to include scenarios involving agents as well as robots—predicting that the former scenarios would elicit more indeterministic intrusion because agents are usually deemed to be more indeterministic than robots, making the latter an illustrative contrast class.

Our primary prediction was that we would find intrusion effects across all conditions. We also predicted that the effects would be more pronounced in the agent conditions than in the robot conditions, which would suggest that the indeterminism at play is something focused more narrowly on human agency (rather than the structure of the universe more generally). We thought that if our findings supported these predictions, this would problematize the findings that have used to support natural compatibilism. For if the intrusion of an indeterministic metaphysics occurs when people think about human agency, then researchers must be careful to control for this tendency when studying free will judgments.

3. Methods

3.1. Participants

To contribute to a better understanding of indeterministic intrusion effects, we ran a new study using a between-subject design. While we recruited 1,500 participants from Amazon Mechanical Turk (MTurk), only 1,384 completed the study. Participants had to satisfy three conditions in order to qualify to take our Mturk Human Intelligence Task (HIT): (a) they needed to have successfully completed at least 500 HITS in the past, (b) they needed a HIT success completion rate of at least 97%, and (c) they needed to be located in the United States. On average, it took participants 15 minutes and 10 seconds to complete the study. They were paid $1 for participating. In addition to excluding participants who did not complete the study (*n* = 112), we also excluded participants who failed a comprehension check (*n* = 88).[5] This left us with *N* = 1,296 participants (M_{age} = 35.22 years, SD = 11.65, $range_{age}$ = [18 - 77], 48% females, 78% Caucasian). This study was approved by the Institutional Review Board of the College of Charleston.

3.2. Materials and Design

Participants were randomly assigned to one of twelve conditions. The factors that varied between cases included Scenario (Supercomputer, Rollback), Action (Good, Bad), Case Type (Abstract, Concrete), and Entity (Agent, Robot). Case Type wasn't fully crossed due to the fact that the abstract versions don't specify any particular actions undertaken in that universe. We thus begin by presenting the cases for the concrete variations and then the cases for the abstract variants.

Beginning with the concrete versions, here is the Bad Action version of the Supercomputer case, featuring an agent:

> *Supercomputer*: Imagine that in the next century we discover all the laws of nature, and we build a supercomputer which can deduce from these laws of nature and from the current state of everything in the world exactly what will be happening in the world at any future time. It can look at everything about the way the world is and predict

[5] We say more about the comprehension question we used and why we used it in §3.2.

> everything about how it will be with 100% accuracy. Suppose that such a supercomputer existed, and it looks at the state of the universe at a certain time on March 25th, 2150 A.D., twenty years before Jeremy Hall is born. The computer then deduces from this information and the laws of nature that Jeremy will definitely rob Fidelity Bank at 6:00 PM on January 26th, 2195. As always, the supercomputer's prediction is correct; Jeremy robs Fidelity Bank at 6:00 PM on January 26th, 2195.[6]

In the Good Action version, robbing was replaced with saving a child from a burning building. And to vary the Entity involved in the case, Jeremy was replaced with a robot. That said, here is the Bad Action version of the Rollback case, featuring an agent:

> *Rollback:* Imagine there is a universe that is re-created over and over again, starting from the exact same initial conditions and with all the same laws of nature. In this universe the same conditions and the same laws of nature produce the exact same outcomes, so that every single time the universe is re-created, everything must happen the exact same way. For instance, in this universe a person named Jeremy decides to rob a bank at a particular time, and every time the universe is re-created, Jeremy decides to rob a bank at that time.[7]

As in the Supercomputer case, in the Good Action version, robbing was replaced with saving a child. And again the Entity was varied by replacing Jeremy with a robot.

[6] This scenario was taken directly from Nahmias et al. (2005; 2006).

[7] This scenario was taken directly from Nahmias et al. (2005; 2006). There were two differences: First we used Jeremy rather than Jill so as not to introduce gender as a confound. Second, we had Jeremy rob a bank so that the severity of the action was the same in Supercomputer and Rollback (here again, to avoid a potential confound).

The abstract versions varied in whether the entity described in the scenario was an agent or robot and in whether they described determinism in terms of a supercomputer or rollback. Here is the Abstract Supercomputer case:

> *Abstract Supercomputer:* Imagine that in the next century we discover all the laws of nature, and we build a supercomputer which can deduce from these laws of nature and from the current state of everything in the world exactly what will be happening in the world at any future time. It can look at everything about the way the world is and predict everything about how it will be with 100% accuracy. Suppose that such a supercomputer existed, and it looks at the state of the universe at a certain time on March 25th, 2150 A.D., twenty years before a person is born. The computer then deduces from this information and the laws of nature what this person will do at any given time.

And here is the abstract rollback case:

> *Abstract Rollback*: Imagine there is a universe that is re-created over and over again, starting from the exact same initial conditions and with all the same laws of nature. In this universe the same conditions and the same laws of nature produce the exact same outcomes, so that every single time the universe is re-created, everything must happen the exact same way. For instance, a person in this universe will perform the same actions, at the same times, every time this universe is re-created.

After reading one of the twelve cases, participants were asked the following (in fixed order):[8]

[8] We borrowed this "suspension of belief" paradigm from Nahmias et al. (2005; 2006). They adopted this conditional approach in the event that some participants would not find the scenarios to be possible. We followed their lead in our present studies.

> *Possible:* Do you think this scenario is possible? [Yes/No]
>
> *Explain:* Please briefly explain your answer.

Next, participants read the following instructions:

> Regardless of how you answered the previous question, imagine such a supercomputer actually did exist and actually could predict the future, including Jeremy's robbing the bank (and assume Jeremy does not know about the prediction) [or imagine such a universe actually did exist and that every time it is recreated everything must happen the same way].
>
> Now please note your level of agreement (or disagreement) with the following statements. Remember that the statements are specifically about the scenario you just read.

Participants in all conditions then rated their agreement on a standard 7-pt Likert scale (1 = strongly disagree and 7 = strongly agree) for eleven different test statements,[9] all of which were presented in random order.[10] Here are the versions of the test statements used in the concrete version of the supercomputer case where Jeremy robbed a bank:

> *Free Will:* Jeremy robbed the bank of his own free will.

[9] We made appropriate wording adjustments depending on the case (e.g., the items were worded differently for Supercomputer and Rollback since the background conditions are not the same).

[10] In Nahmias et al. (2005; 2006), they asked questions about free will and moral responsibility—namely, "Do you think that, when Jeremy robs the bank, he acts of his own free will [is morally blameworthy for robbing the bank]?" They also asked a question about whether the agents in the scenarios had the ability to "choose to do otherwise." In all three cases, answer choices were dichotomous. Given our interest in a wider variety of issues and in more fine-grained data, we included more statements and we didn't give participants a forced choice—we used a 7-pt Likert instead.

Freely Decide: Jeremy freely decided to rob the bank.

Determined: Jeremy was determined to rob the bank.

No Option: Jeremy had no other option than to rob the bank.

Avoid: Jeremy can avoid doing what the computer predicts he will do.

Ability: Jeremy has the ability to change his mind about robbing the bank.

Responsible: Jeremy is fully morally responsible for robbing the bank.

Ultimately: It was ultimately up to Jeremy to decide to rob the bank.

Slight Chance: There was at least a slight chance that Jeremy would not rob the bank as the computer predicted he would.

Do Otherwise: Even though Jeremy actually did what the computer predicted he would do, it was possible for Jeremy to do something else instead at the time.

Scenario Determinism: According to the scenario, the supercomputer can deduce from the laws of nature and from the current state of everything in the world exactly what will be happening in the world at any future time.[11]

Also included in each condition, were three further test questions that utilized different scales:

[11] This item is modelled on the comprehension question used in Nahmias et al. (2005; 2006). It is simply a restatement of deterministic features directly described in the scenarios. Following their lead, we, too, used Scenario Determinism as a comprehension check. The overwhelming majority of participants (93%) correctly identified the scenarios as deterministic (that is, they somewhat agreed, agreed, or strongly agreed with this statement). So, we only excluded 88 out of 1,296 participants from our analysis. Nevertheless, as we will see, intrusion effects were common. This suggests that while most participants correctly identified the very features of the scenarios that made them deterministic, many still either failed to comprehend the broader implications of determinism or failed to hold these implications constant.

Praise/Blame: Jeremy is _________ for his decisions and actions. (1=very praiseworthy, 7=very blameworthy)

Reward/Punish: Jeremy deserves _______ for his decisions and actions. (1=a substantial reward, 7=a substantial punishment)

Chance: What do you think the chances are that Jeremy will do something different than what the computer predicts he will do? (Slider scale ranging from 0=very unlikely to 100=very likely)

Once participants provided their responses to these dependent variables, they were presented with Part 1 and Part 2 of The Free Will Inventory (FWI: Nadelhoffer et al., 2014)—which is a scale designed to measure beliefs about free will, determinism, dualism, and related constructs. Because we included FWI for exploratory purposes, we ultimately decided not to analyze the results as part of the present paper.[12]

4. Results

Since Scenario Determinism was designed to serve as a comprehension check, we removed participants who gave a rating between 1 and 4. This choice was made to cohere with the analytic strategy used by prior researchers (see Nahmias et. al., 2005; 2006) and did not affect any of the central findings reported below. Eighty-eight people were removed resulting in *N* = 1,296 final participants. Next we created three new variables: Free Will, Intrusion and Blame. The full set of variables we will analyze is in Table 1.

[12] The complete data set and supplemental materials (including stimuli, dependent variables for all vignettes, and analyses not included in the paper) can be found at: https://osf.io/4z6r2/

Table 1: Items for used for data analysis

Variable	**Item(s)**	**Cronbach's Alpha**
Free Will	Free Will: Jeremy robbed the bank of his own free will. Freely Decide: Jeremy freely decided to rob the bank. Ultimately: It was ultimately up to Jeremy to decide to rob the bank.	.923
Intrusion	No Option: Jeremy had no other option than to rob the bank.[13] Avoid: Jeremy can avoid doing what the computer predicts he will do. Ability: Jeremy has the ability to change his mind about robbing the bank. Slight Chance: There was at least a slight chance that Jeremy would not rob the bank as the computer predicted he would.	.906

[13] Since disagreement on this indicates intrusion while on all other intrusion items, agreement indicates intrusion, this was reverse coded.

	Do Otherwise: Even though Jeremy actually did what the computer predicted he would do, it was possible for Jeremy to do something else instead at the time.	
Praise/Blame	Praise/Blame: Jeremy is _________ for his decisions and actions. Reward/Punish: Jeremy deserves _______ for his decisions and actions.	.928
Responsibility[14]	Responsible: Jeremy is fully morally responsible for robbing the bank.	n/a
Determined	Determined: Jeremy was determined to rob the bank.	n/a
Chance[15]	Chance: What do you think the chances are that Jeremy will do something different than what the computer predicts he will do?	n/a

Since our design was not a full factorial design, we will begin by analyzing responses in the concrete versions before presenting responses in the abstract versions.

[14] This was not combined with the items for Praise/Blame since this item has a different scale.

[15] This was not combined with the items for Intrusion since this item used a different scale.

We conducted six separate 2(Case: Supercomputer, Rollback) x 2(Entity: Agent, Robot) x 2 (Action: Good, Bad) ANOVAs, one for each dependent measure. See Table 1.

Free Will: There was a main effect of Case, F(1, 854)=72.007, p<.001, ηp2=.078 and Entity, F(1, 854)=367.846, p<.001, ηp2=.301 but no effect of Action, F(1, 854)=.404, p=.524, ηp2=.000. There was a two-way interaction between Case and Entity, F(1, 854)=6.475, p<.05, ηp2=.008 and between Entity and Action F(1, 854)=7.814, p<.01, ηp2=.009 but no two-way interaction between Case and Action, F(1, 854)=2.236, p=.135, ηp2=.003. The three-way interaction between Case, Entity and Action was also significant, F(1, 854)=7.796, p<.01, ηp2=.009.

Intrusion: There was a main effect of Case, F(1, 854)=195.818, p<.001, ηp2=.187 and Entity, F(1, 854)=128.996, p<.001, ηp2=.131 but no effect of Action, F(1, 854)=1.755, p=.186, ηp2=.002. None of the two- or three-way interactions were significant: Case and Entity, F(1, 854)=1.274, p=.259, ηp2=.001; Case and Action F(1, 854)=.654, p=.419, ηp2=.001; Entity and Action, F(1, 854)=2.090, p=.149, ηp2=.002; Case, Entity and Action, F(1, 854)=2.622, p.106, ηp2=.003.

Praise/Blame: There was a main effect of Action, F(1, 854)=1485.696, p<.001, ηp2=.635, but no effect of Case, F(1, 854)=2.389, p=.123, ηp2=.003 or Entity, F(1, 854)=2.250, p<=.134, ηp2=.003. There was a two-way interaction between Entity and Action, F(1, 854)=191.163, p<.001, ηp2=.183 and Case and Action, F(1, 854)=25.279, p<.001, ηp2=.029, but no interaction between Case and Entity F(1, 854)=1.502, p=.221, ηp2=.002. The three-way interaction between Case, Entity and Action was not significant, F(1, 854)=.114, p=.736, ηp2=.000.

Responsibility: There was a main effect of Case, F(1, 854)=15.469, p<.001, ηp2=.018, Entity, F(1, 854)=306.709, p<.001, ηp2=.264 and Action, F(1, 854)=4.247, p<.05, ηp2=.005. There was a two-way interaction between Entity and Action, F(1, 854)=14.072, p<.001, ηp2=.016 but no interaction between Case and Entity F(1, 854)=.888, p=.346, ηp2=.001 or Case and Action, F(1, 854)=1.806, p=.179, ηp2=.002. The three-way interaction between Case, Entity and Action was not significant, F(1, 854)=.220, p=.639, ηp2=.000.

Determined: There was a main effect of Entity, F(1, 854)=24.395, p<.001, ηp2=.028 but no effect of either Case, F(1, 854)=.603, p=, ηp2=.001 or Action, F(1, 854)=.649, p=.421, ηp2=.001. There were no significant two-way interactions: Case and Entity, F(1, 854)=2.375, p=.124, ηp2=.003; Entity and Action F(1, 854)=.118, p=.731, ηp2=.000; Case and Action, F(1, 854)=.013, p=.908, ηp2=.000. The three-way interaction between Case, Entity and Action was also not significant, F(1, 854)=3.646, p=.056, ηp2=.004.

Chance: There was a main effect of Case, F(1, 854)=89.950, p<.001, ηp2=.095 and Entity, F(1, 854)=20.888, p<.001, ηp2=.024 but no effect of Action, F(1, 854)=.256, p=.613, ηp2=.000. There were no significant two-way interactions: Case and Entity, F(1, 854)=.016, p=.900, ηp2=.000; Entity and Action F(1, 854)=.034, p=.854, ηp2=.000; Case and Action, F(1, 854)=.337, p=.562, ηp2=.000. The three-way interaction between Case, Entity and Action was also not significant, F(1, 854)=.001, p=.970, ηp2=.000.

Figs. 1 and 2 visualize the findings—with the primary focus on differences between the Agent and Robot cases—for both the Supercomputer and Rollback cases. Since Chance utilized a 0-100 pt. scale, these findings are shown in Fig. 3.

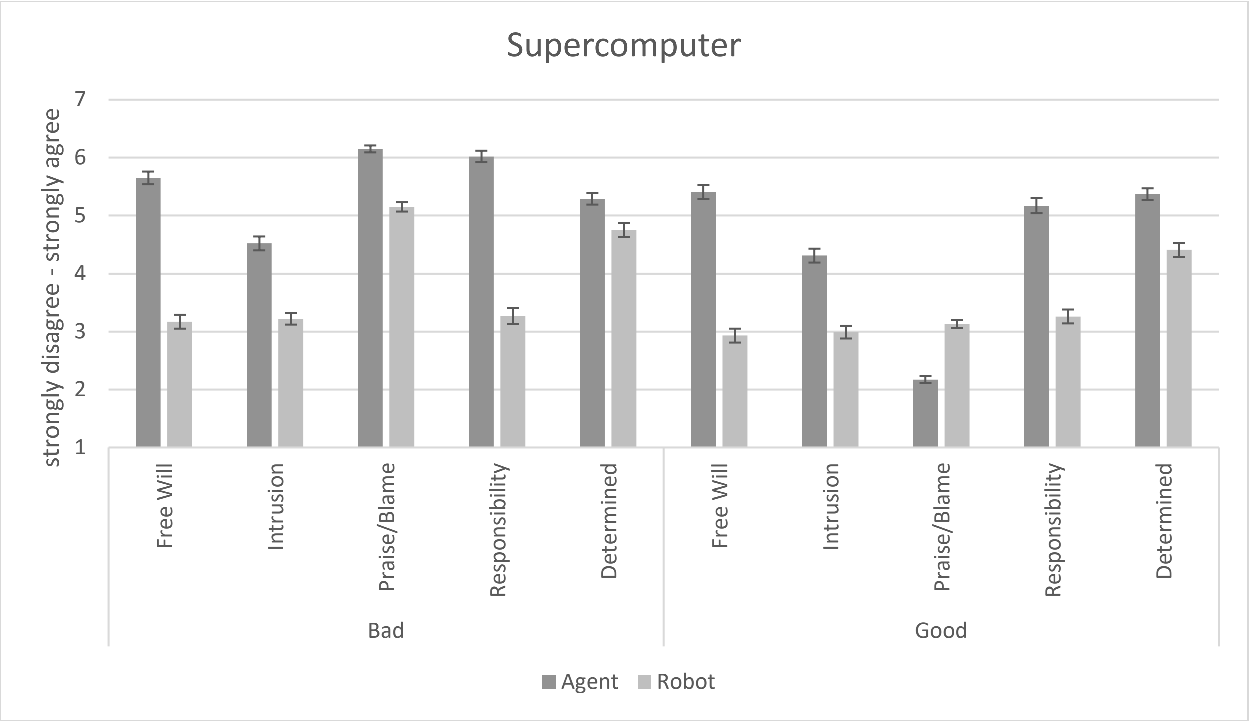

Fig. 1: Agent and Robot results for Supercomputer concrete cases with 95% confidence intervals.

1 = strongly disagree; 7 = strongly agree

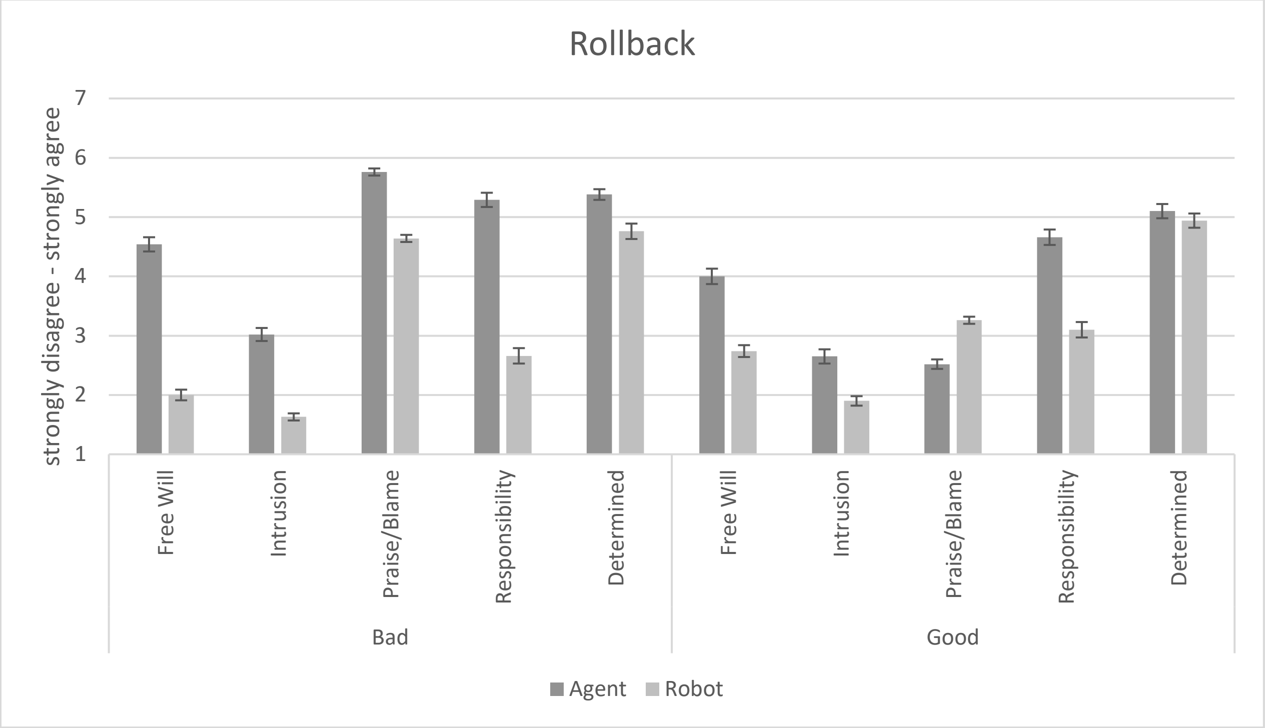

Fig. 2: Agent and Robot results for Rollback concrete cases. 1 = strongly disagree; 7 = strongly agree

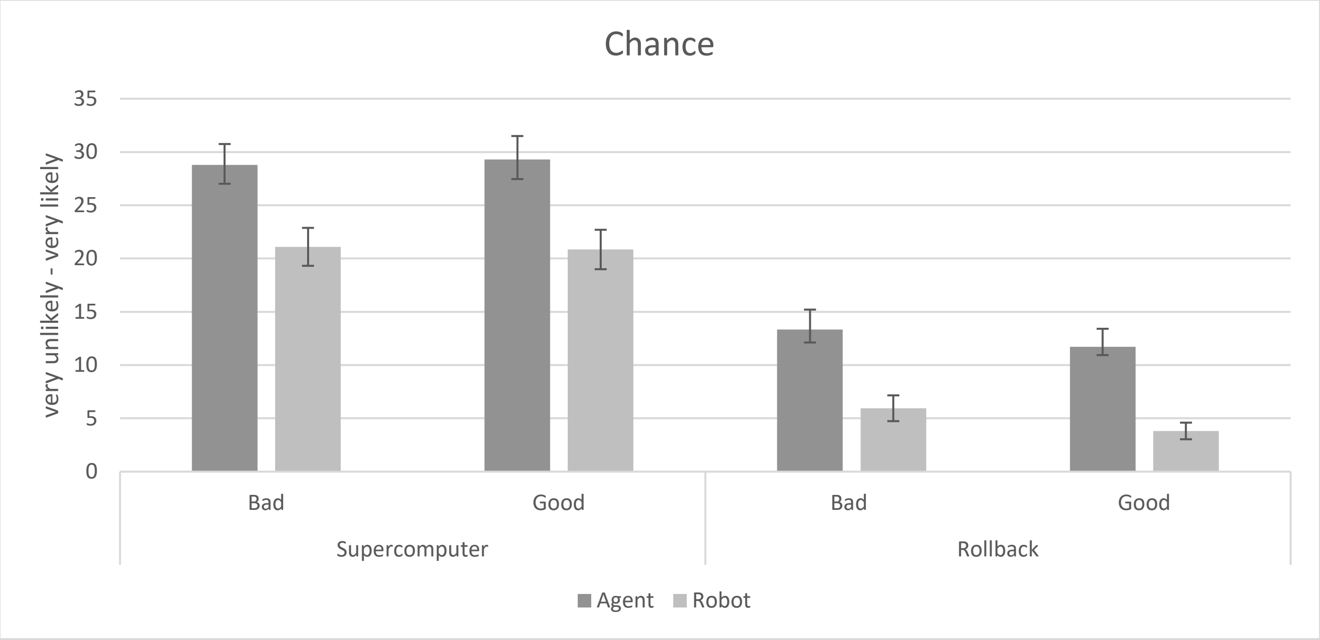

Fig. 3: Agent and Robot results for concrete cases. 0 = very unlikely; 100 = very likely

Having analyzed the concrete cases, we next turned our attention to the abstract cases. The first thing worth noting on this front is that in the abstract conditions, praise/blame and reward/punish questions were not included and so will be excluded from analysis. We conducted five separate 2(Case: Supercomputer, Rollback) x 2(Entity: Agent, Robot) ANOVAs, one for each dependent measure (excluding blame). See Table 1.

Free Will: There was a main effect of Case, F(1, 430)=76.701, p<.001, ηp2=.151 and Entity, F(1, 430)=221.877, p<.001, ηp2=.340, as well as a two-way interaction between Case and Entity, F(1, 430)=22.295, p<.001, ηp2=.049.

Intrusion: There was a main effect of Case, F(1, 430)=102.110, p<.001, ηp2=.192 and Entity, F(1, 430)=85.357, p<.001, ηp2=.166, as well as a two-way interaction between Case and Entity, F(1, 430)=8.709, p<.01, ηp2=.049.

Responsibility: There was a main effect of Case, F(1, 430)=23.581, p<.001, ηp2=.052 and Entity, F(1, 430)=225.451, p<.001, ηp2=.453, and a two-way interaction between Case and Entity, F(1, 430)=18.057, p<.001, ηp2=.040.

Determined: There was a main effect of Case, F(1, 430)=7.660, p<.01, ηp2=.151 and Entity, F(1, 430)=20.633, p<.001, ηp2=.046, but no two-way interaction between Case and Entity, F(1, 430)=.026, p=.872, ηp2=.000.

Chance: There was a main effect of Case, F(1, 430)=63.218, p<.001, ηp2=.128 and Entity, F(1, 430)=50.234, p<.001, ηp2=.105, as well as a two-way interaction between Case and Entity, F(1, 430)=8.661, p<.01, ηp2=.020.

Fig. 4 presents a visualization of the differences between the agent and robot versions in the abstract cases. Again since chance ratings were made on a different scale, these are shown in Fig. 5.

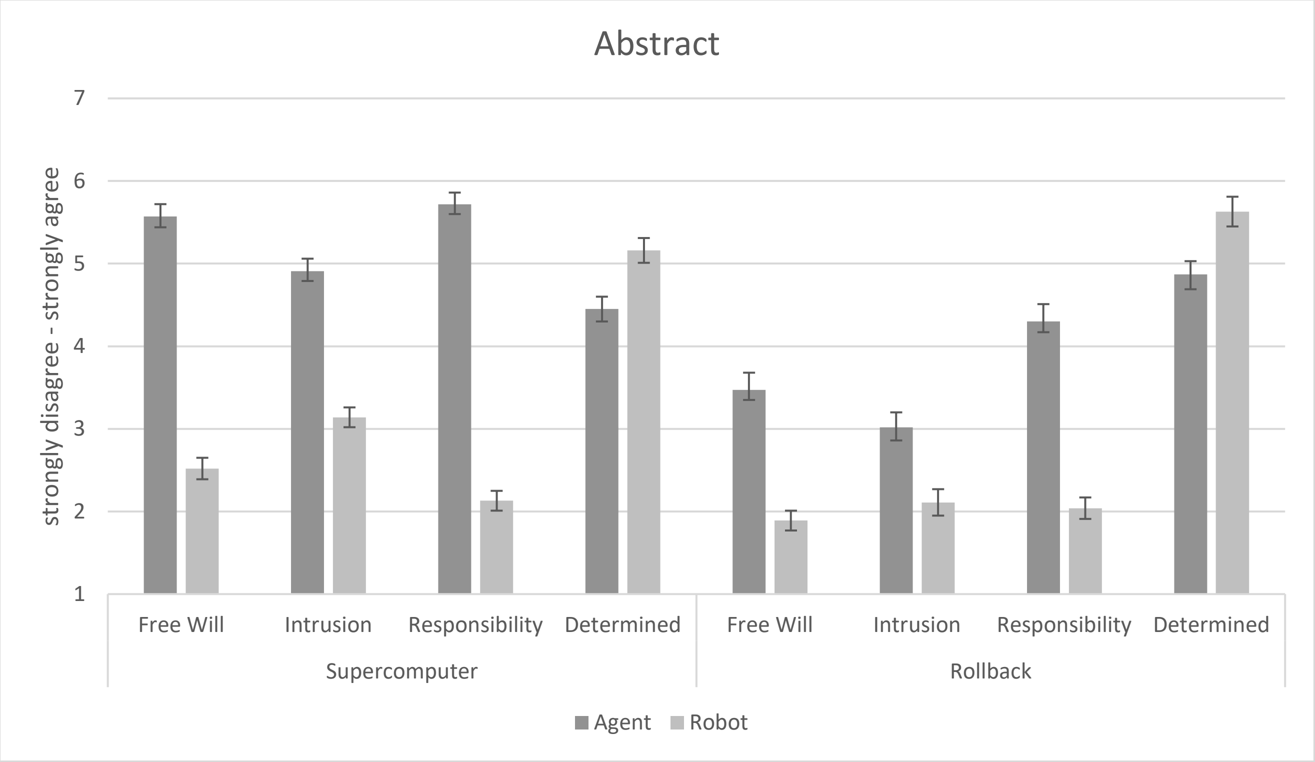

Fig. 4: Agent and Robot differences in abstract cases with 95% confidence interval. 1 = strongly disagree; 7 = strongly agree

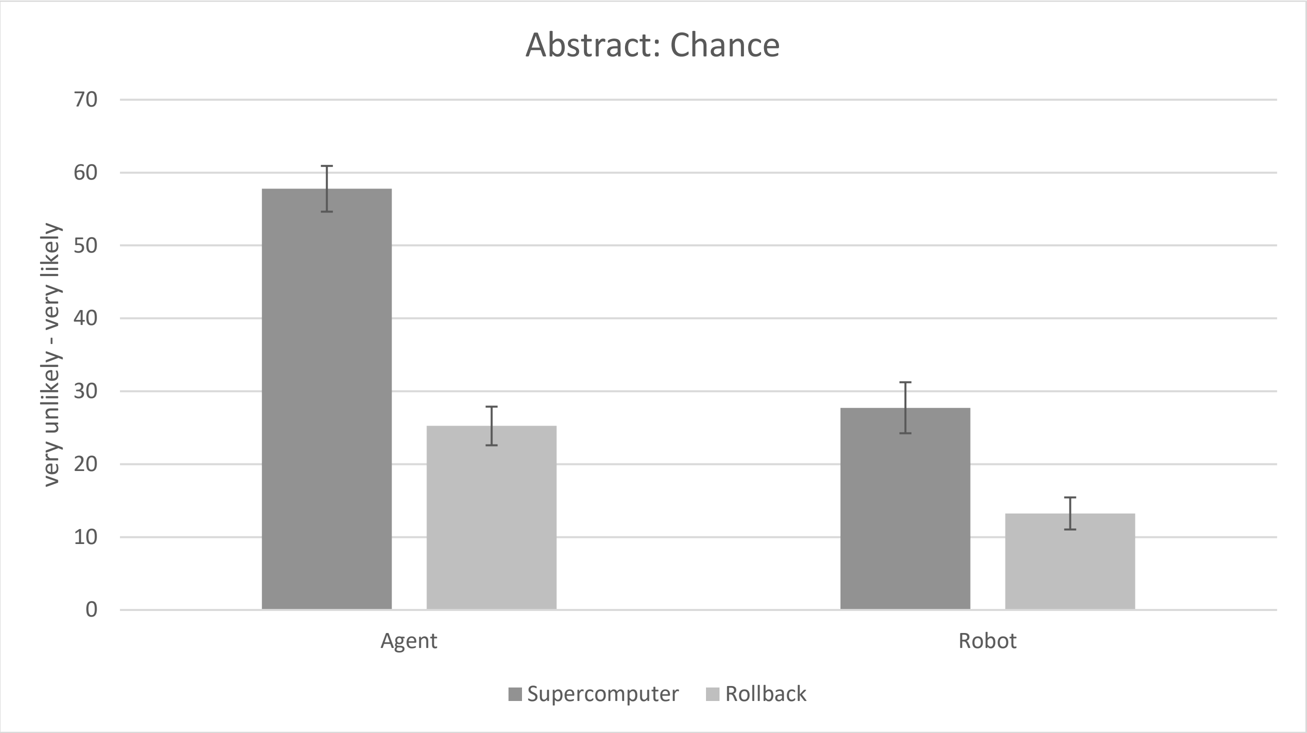

Fig. 5: Agent and Robot differences in chance judgments. 0 = very unlikely; 100 = very likely

There are two final predictions we made regarding intrusion. First, Chance asks, e.g., "What do you think the chances are that Jeremy will do something different than what the computer predicts he will do?" People made ratings on a slider scale ranging from 0 (very unlikely) to 100 (very likely). If we take ratings of 0 on Chance to indicate that people are not succumbing to intrusion and ratings greater than 0 to indicate that people are succumbing to intrusion then we can ask whether there are differences on both of our measures of free will. We predicted that there should be. In particular, those who give ratings of higher than 0 on Chance should be more inclined to attribute free will. Moreover, that effect should be more pronounced in cases featuring an agent. In other words, we predicted an interaction between Entity and Chance on free will judgments.

Beginning with the concrete versions, and collapsing across Case, 50% of participants gave a rating of greater than 0 on Chance. We found a main effect of both Chance F (1, 858) = 70.687, p<.001, ηp2= .076, and Entity F (1, 858) = 331.021, p<.001, ηp2= .278, on Free Will. Importantly, we found a two-way interaction between Chance and Entity, F (1, 858) = 9.505, p<.01, ηp2= .011.

For the abstract cases, and again collapsing across Case, 64% of participants gave a rating of greater than 0 on Chance.[16] We again found a main effect of both Chance F (1, 430) = 143.746, p<.001, ηp2= .251, and Entity F (1, 430) = 161.441, p<.001, ηp2= .273, on Free Will. Again, and

[16] We were surprised that intrusion was more prevalent in the abstract cases than in the concrete cases. One possibility is that the abstract nature of the case itself leads to more options for filling in details with one's own (indeterministic) view. In concrete cases, more stipulation/detail might impede that somewhat—though, as we find, intrusion still occurs at an alarming rate even in the concrete cases. Of course, this is just a hunch. More work on the asymmetry of intrusion effects in concrete and abstract cases would need to be done to know how best to explain these results.

importantly, we found a two-way interaction between Chance and Entity, F (1, 430) = 30.432, p<.001, ηp2= .066. The results are shown in Fig. 6.

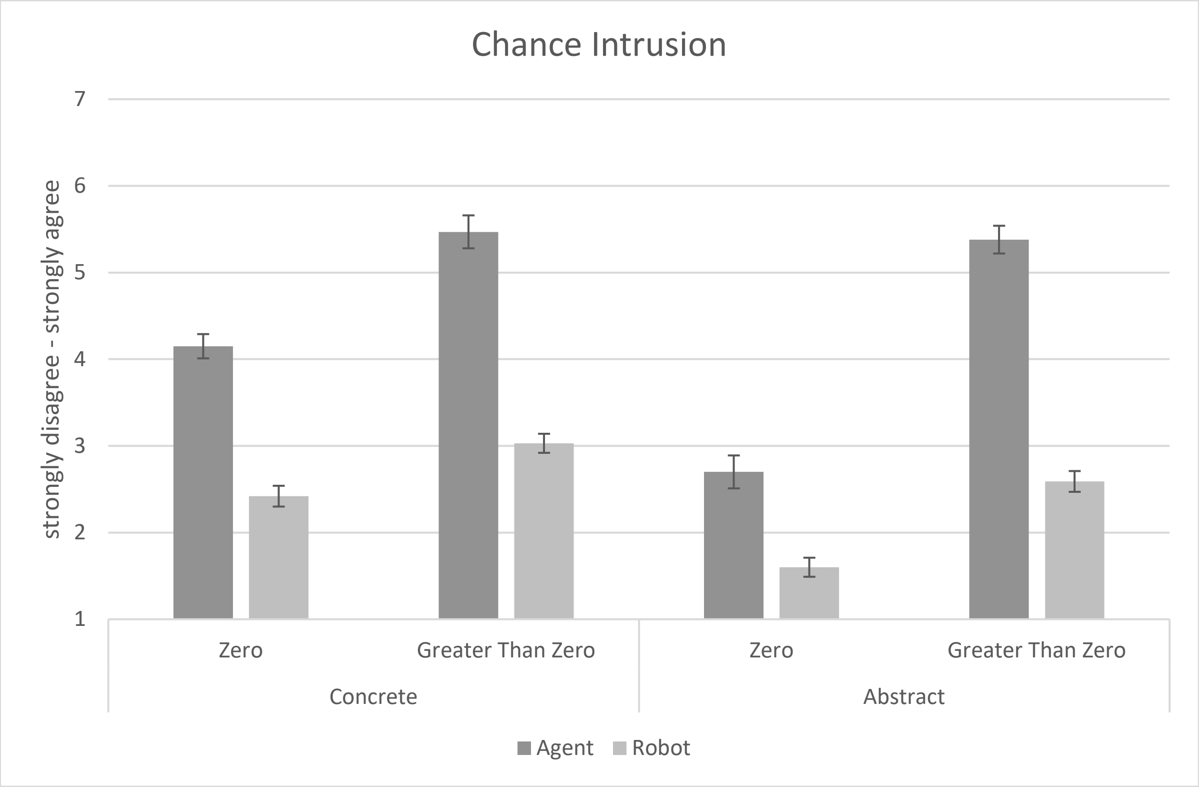

Fig. 6: Chance Intrusion Effects on Free Will for Concrete and Abstract cases with 95% Confidence Intervals. 1 = strongly disagree; 7 = strongly agree

To take all of this one step further, we make one final prediction. Rose et al. (2017) used perfect neuro-prediction and manipulation cases to manipulate intrusion. The expectation was that people's intuitive views about agency would intrude into the representation of neuroscientific scenarios when making free will judgments. And thus they hypothesized—and presented evidence—that “intuitive free will judgments cause people to misrepresent instances of perfect neuro-prediction” (2017, p. 487). In our cases here—Supercomputer and Rollback—our Entity manipulation was aimed at tapping into people’s intuitive views about agency. And so we predict that intuitive views about agency—and indeterministic metaphysics—intrude into people's representations of Supercomputer and Rollback when making free will judgments and cause them to misrepresent the scenarios. To determine whether this is the case, our focus will be on free will judgments and what is perhaps our best measure of intrusion: Chance.

We conducted four separate mediation analyses, one for each of our cases, Supercomputer and Rollback, in the concrete and abstract conditions. First, we analyze the concrete cases. For Supercomputer, a regression model with Entity as a predictor of Chance was significant, t(402) = −2.835, β = −.140, p < .01; a regression model with Entity as a predictor of Free Will was significant, t(402) = −14.821, β = −.595, p < .001; a regression model with Free Will as a predictor of Chance was significant, t(402) = 4.854, β = .236, p < .001; but a multiple regression model with both Entity and Free Will as predictors of Chance indicated that the effect of Entity was no longer significant, t(402) = -.001, β = .000, p = .999. We also tested the alternative mediation model (see e.g., Rose and Nichols 2013; Rose et al. 2012 Rose 2017; Rose and Nichols 2019 a; Rose and Nichols 2019 b; Rose and Nichols 2020; Rose et al. forthcoming; Turri, Buckwalter, & Rose, 2016). A multiple regression model with both Entity and Chance as predictors of Free Will showed that Entity significantly predicted Free Will, t(402) = −14.385, β = −.573, p < .001, but that Chance did not mediate the effect of Entity on Free Will.

For Rollback, a regression model with Entity as a predictor of Chance was significant, t(458) = −3.822, β = −.176, p < .001; a regression model with Entity as a predictor of Free Will was significant, t(458) = −11.838, β = −.484, p < .001; a regression model with Free Will as a predictor of Chance was significant, t(458) = 5.760, β = .260, p < .001; but a multiple regression model with both Entity and Free Will as predictors of Chance indicated that the effect of Entity was no longer significant, t(458) = -1.263, β = -.065, p = .207. The alternative mediation model with both Entity and Chance as predictors of Free Will showed that Entity significantly predicted Free Will, t(458) = −11.108, β = −.453, p < .001, but that Chance did not mediate the effect of Entity on Free Will.

So in the concrete versions of both Supercomputer and Rollback, we find evidence that Free Will mediates the effect of Entity on Chance. These models can be seen in Fig. 7.

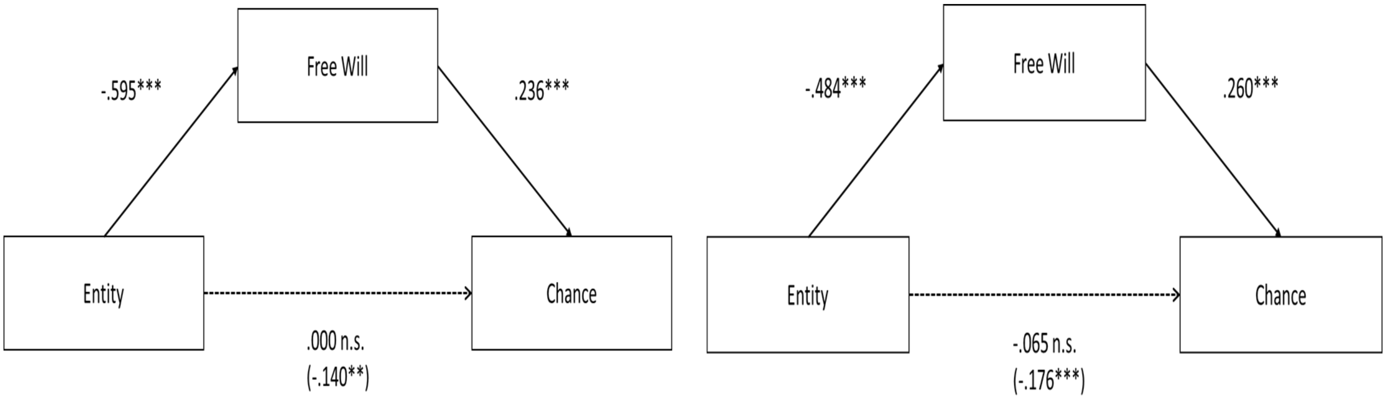

Fig. 7: Concrete Mediation Models for Supercomputer (Left) and Rollback (Right)

We consider next whether there is mediation in the abstract versions of these cases. For Supercomputer, a regression model with Entity as a predictor of Chance was significant, t(215) = −7.095, β = −.436, p < .001; a regression model with Entity as a predictor of Free Will was significant, t(215) = −15.295, β = −.723, p < .001; a regression model with Free Will as a predictor of Chance was significant, t(215) = 11.065, β = .603, p < .001; but a multiple regression model with both Entity and Free Will as predictors of Chance indicated that the effect of Entity was no longer significant, t(215) = -.012, β = -.001, p = .991. The alternative model with both Entity and Chance as predictors of Free Will showed that Entity significantly predicted Free Will, t(215) = −12.163, β = −.567, p < .001, but that Chance did not mediate the effect of Entity on Free Will.

For Rollback, a regression model with Entity as a predictor of Chance was significant, t(217) = −2.930, β = −.196, p < .01; a regression model with Entity as a predictor of Free Will was significant, t(217) = −6.639, β = −.412, p < .001; a regression model with Free Will as a predictor of Chance was significant, t(217) = 10.187, β = .570, p < .001; but a multiple regression model with both Entity and Free Will as predictors of Chance indicated that the effect of Entity was no longer significant, t(217) = .765, β = .047, p = .445. The alternative mediation model with both Entity and Chance as predictors of Free Will showed that Entity significantly predicted Free Will, t(217) = −5.887, β = −.312, p < .001, but that Chance did not mediate the effect of Entity on Free Will. These models are shown in Fig. 8.

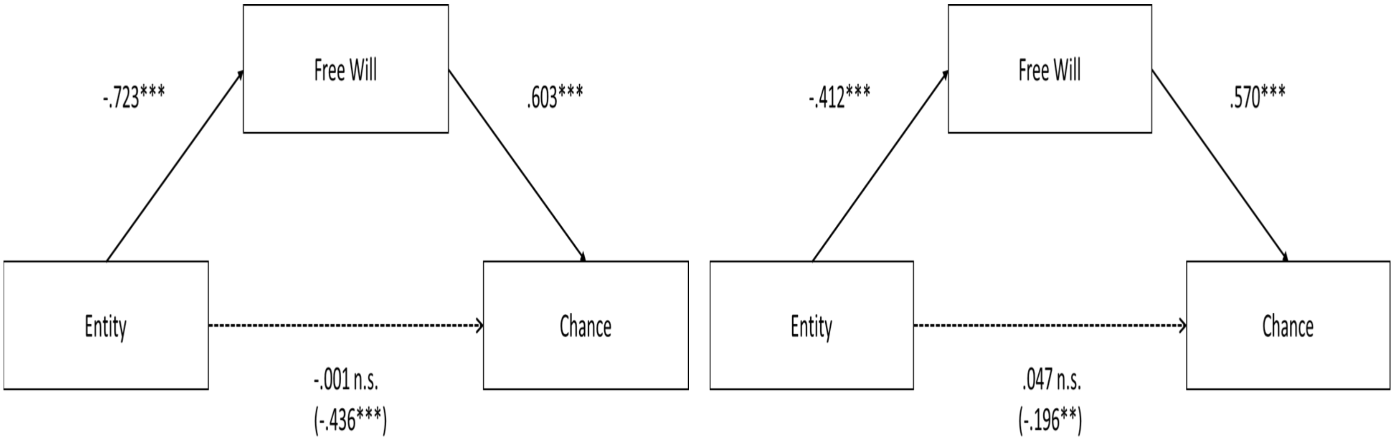

Fig. 8: Abstract Mediation Models for Supercomputer (Left) and Rollback (Right)

In every single case, we found that Free Will mediates the effect of Entity on Chance. This coheres with the findings of Rose et al. (2017) and extends those findings to the leading, classic cases supporting natural compatibilism. We also note that just like with Chance, Slight Chance (one of the items included in our composite measure of intrusion) gives a very similar picture.[17]

[17] We will explain why we looked at Slight Chance on its own in §5.

For the concrete cases we find a main effect of Case, F(1, 854)=115.707, p<.001, ηp2=.119 and Entity, F(1, 854)=19.214, p<.001, ηp2=.022 but no effect of Action, F(1, 854)=.387, p=.534, ηp2=.000. Moreover, just as with Chance, there were no significant two- or three-way interactions (all p's >.05). Similarly, in the abstract cases, there was a main effect of Case, F(1, 430)=69.074, p<.001, ηp2=.138 and Entity, F(1, 430)=12.302, p<.01, ηp2=.028, but no two-way interaction between Case and Entity, F(1, 430)=.749, p=.387, ηp2=.002. Importantly, if we instead run all the same mediation analyses using Slight Chance, in virtually every single case Free Will mediates the effect of Entity on Slight Chance and Slight Chance doesn't mediate the effect of Entity on Free Will in any of these models.[18]

In short, whether we focus on the composite measure of intrusion or individual measures like Slight Chance or Chance, we find evidence that intuitive views about the indeterministic nature of human agency influence how people understand deterministic cases like Supercomputer and Rollback when they are making free will judgments. Owing to this indeterministic intrusion, most people have a difficult time adequately tracking the determinism built into two of the central scenarios that have been used to support natural compatibilism. While this is our key finding, there are a number of other issues raised by our results that also

[18] For the concrete Supercomputer cases, a multiple regression model with both Entity and Free Will as predictors of Slight Chance indicated that the effect of Entity was not significant, t(402) = .404, β = .024, p = .686; for the concrete Rollback cases, a multiple regression model with both Entity and Free Will as predictors of Slight Chance indicated that the effect of Entity was not significant, t(458) = -.914, β = -.047, p = .361; for the abstract Rollback cases, a multiple regression model with both Entity and Free Will as predictors of Slight Chance indicated that the effect of Entity was not significant, t(217) = 1.615, β = .102, p = .108; and for the abstract Supercomputer cases, a multiple regression model with both Entity and Free Will as predictors of Slight Chance indicated that the effect of Entity was significant, t(215) = 2.257, β = .202, p = .025 and reduced so as to suggest partial mediation.

merit further consideration. So, we are now going to unpack some of the things we found along the way and discuss the downstream implications.

5. Discussion

The first thing that merits discussion is that 66% of participants didn't think the scenarios were possible. This concords with the earlier findings by Nahmias et al. (2005; 2006) which is why we borrowed their "suspension of belief" paradigm for our present study. One likely explanation for this finding is that people are indeterminists about human agency. For instance, Bloom (2012) suggests that "common sense tells us we exist outside the material world," and Knobe (2014) similarly claims that according to folk psychology, "human actions are not caused by prior events" (79). However, the commonsense view need not be transcendent or supernatural to be indeterministic. As Turri (2017c) recently pointed out, "It could be that, on the ordinary understanding, human agency fits broadly within the causal order while still being exceptional in some respects. For example, people might think that human actions are caused by psychological, neurological, and social events, even though human agents can resist causal forces in ways that inanimate objects cannot" (2-3). Importantly, according to Turri's findings, people think that human agents—unlike robots—can resist causal forces described as inevitable, guaranteed, or causally determined even though human agents are at the same time viewed as part of the natural world.

Figuring out the contours of folk views about indeterminism and human exceptionalism is a task for another day. We mention these issues here given their likely relationship with the intrusion effects that are our focus. We believe that the main reason that intrusion effects are prevalent is that indeterminism about human agency is the default folk view. That so many participants find deterministic scenarios to be impossible highlights the intuitive appeal of an indeterministic metaphysics. It also presumably partly explains why some people appear to have such a hard time tracking determinism. At a minimum, the fact that so many people found the scenarios to be impossible should weaken our confidence in their subsequent intuitions about what is possible in these scenarios (e.g., free will, moral responsibility, the ability to do otherwise, etc.). Given that these scenarios are widely deemed to be impossible—which is telling in itself—we think it is safe to assume that they likely invite confusion and misunderstanding from the outset.[19]

Second, as indicated by the main effect of Case, judgments about free will and moral responsibility tended to be higher in Supercomputer than in Rollback. This suggests that Rollback may be more compellingly deterministic than Supercomputer. An additional reason to think that Rollback might be more compellingly deterministic is that participants were more likely to find the former possible than the later. In the concrete scenarios, 77% thought the scenario was not possible in Supercomputer and 56% thought it was not possible in Rollback. This difference was significant $X^2(1)=44.739$, $p<.001$, Cramer's V=.228. In the abstract scenarios, 79% thought it was not possible in Supercomputer and 56% thought it was not possible in Rollback. This difference was significant $X^2(1)=26.431$, $p<.001$, Cramer's V=.247. These differences could be because

[19] It is worth noting that while most participants appear to find the scenarios to be counter-intuitive, the overwhelming majority correctly responded to the comprehension check we designed to make sure they understand in a broad sense that the scenarios were deterministic. So, this suggests that they don't find determinism inconceivable or incomprehensible. For instance, it's not that participants don't understand the notion of perfect prediction, they just likely think it's impossible that a computer will ever be developed that has the ability to perfectly predict everything that happens in the universe.

participants found the determinism built into Supercomputer—based as it is on perfect prediction by a computer—to be less plausible than the determinism built into Rollback.

In short, perfect prediction may not be the most intuitive way of capturing the salient features of determinism.[20] After all, predictions may be naturally viewed indeterministically given our common experiences with otherwise reliable predictions gone awry. Given that determinism is explained in terms of perfect prediction in Supercomputer, this may lead to less reliably deterministic intuitions in those cases. On the other hand, it is more difficult to give Rollback a similar indeterministic gloss since it explicitly states that given the same initial conditions and the same laws, the universe *must* turn out the same way every time it is recreated—that is, it's not treated as a prediction but rather as an entailment. So, while we do find intrusion effects even in the more convincingly deterministic rollback scenarios, these effects are less prevalent than in the supercomputer cases. This difference between the plausibility of various descriptions of determinism is something researchers should bear in mind in the future. It's not enough to simply describe determinism—it must be described in a way that people find as intuitive as possible.

Third, as expected, and as indicated by the main effect of Entity on judgments about free will and responsibility, agents were deemed to be more free and responsible than robots. This is precisely why we included some cases involving agents and some involving robots. We predicted that these latter cases would be less likely to elicit indeterministic intrusion effects, which were the primary focus of our study. The robot cases gave us a way of trying to minimize intrusion effects and served as an illustrative contrast class to the agent cases. As one can see in Fig. 3, participants' intuitions about Chance—arguably our strongest measure of intrusion (see below

[20] If this is right, it likely applies to the neuroprediction used in Nahmias et al. (2014).

for details)—were much weaker in the robot cases than in the agent cases.[21] This comports with the findings by Turri (2017c) whereby people tend to view the behavior of robots as less avoidable (i.e., more deterministic) and view the behavior as agents as more avoidable (i.e., more indeterministic). According to his findings, people tend to believe that whereas humans have the ability do otherwise (which is an important element of free will), robots largely lack this ability. Given that attributions of free will were also weaker in our robot cases, this fits with our overall expectations concerning the relationship between free will beliefs and the intrusion of an indeterministic metaphysics in otherwise deterministic scenarios.

Part of what people seem to be considering when making judgments about free will is whether the agent could have avoided doing what he or she did given the circumstances at the time. That is presumably why indeterministic intrusion is so prevalent when it comes to judgments about free will involving agents. Agents, unlike robots, are paradigmatically indeterministic beings who have the ability to do otherwise under ordinary circumstances. As such, when people are presented with agents in deterministic scenarios—which they find counter-intuitive, as we've seen—the most obvious way to preserve the agents' freedom and responsibility is to unwittingly view them through the lens of indeterminism, even in experimental contexts where researchers don't want them to do so. This kind of indeterminism may be baked into our ordinary conception of agency. When people are presented instead with robots in deterministic scenarios, there isn't the same temptation to view them through the lens

[21] It is still noteworthy that we nevertheless did find intrusion effects even in the robot cases—which speaks to the prevalence of indeterministic intrusion.

of indeterminism—which explains why both intrusion and attribution of free will are lower in these cases than in cases involving agents.

In this way, our intrusion measure Chance may tap into a very common way people think about their own agency and responsibility. After all, it often feels as if there is a real chance people can avoid doing whatever has been forecasted for them to do. Even the best predictions fall apart when it comes to human behavior. So, it only makes sense that the indeterministic notion of chance might be enmeshed with our ordinary understanding of free will. In some sense, chance provides the possibility space which we then navigate using our capacity for free and responsible agency. To say that there is a chance that an agent's future unfolds in more than one way is to view the agent as if his life is a kind of garden of forking paths, which is a common metaphor in the free will literature.

If this is right—that is, if chance plays an important role in how people ordinarily think about free will—it would help explain why 50% of people in the concrete cases and 68% of our participants in the abstract cases, seemed to be tacitly importing misplaced indeterministic assumptions into our deterministic scenarios.[22] Indeed, if we treated Chance as a genuine comprehension check, then people who answered Chance incorrectly would be excluded from our analysis of the data. Recoding Chance so that those who answered it correctly, that is, answered zero are grouped together, and those who answered it incorrectly—that is, indicated that there was a greater than zero chance—are grouped together, the findings concerning free

[22] That is, if we recode Chance, assigning 0 to those who said there was zero chance and 1 to those who said there is greater than zero chance, we find that 50%-68% of people said there was greater than zero chance that the agent could have acted differently.

will judgments change significantly, Concrete: $X^2(1)=53.432$, p<.001, Cramer's V=.257; Abstract: $X^2(1)=43.393$, p<.001, Cramer's V=.324. This can be seen in Table 2 below.

	Free Will (Yes)[23]	
	Zero Chance	**Greater Than Zero**
Concrete	27%	52%
Abstract	13%	45%

Table 2: Percent Affirming Free Will by whether the participant correctly answers Chance (Zero) or incorrectly answers Chance (Greater than Zero) for both Concrete and Abstract cases.

These findings on the relationship between Chance and judgments about free will problematize the conclusions drawn by Nahmias and colleagues (2005; 2006). As we saw earlier, 65-85% of their participants who read Supercomputer or Rollback judged that Jeremy was free and responsible. They took these results to support natural compatibilism. However, using those same scenarios coupled with a way of measuring intrusion, we found instead that most of the participants who appeared to be compatibilists are importing indeterminism. Once we include only people who said there was zero chance, and focus only on the cases featuring an agent, while collapsing across all other variables, only 39% judged that the agent was free, which is significantly below chance, binomial test, p<.001. Not only do we find much less support for natural compatibilism as a general matter than Nahmias and colleagues, but our findings suggest that people who are prone to indeterministic intrusion are far more likely to judge that an agent

[23] Ratings from 5-7 were taken to indicate free will was attributed.

in a deterministic scenario has free will. In the abstract cases—where people who are prone to intrusion are roughly four times more likely to attribute free will than people who are not prone to intrusion—this difference is especially prominent. It appears that the belief in indeterministic chance is a driving force behind people's judgments concerning free and responsible agency.

In light of these findings, we believe that we have provided evidence that (a) intrusion effects are prevalent in response to the kinds of scenarios used by natural compatibilists, and (b) intrusion effects are closely related to people's intuitions about free will even in the face of determinism. As such, we believe that our findings problematize the evidence that has been used to support natural compatibilism. It's not that no participants give properly compatibilist answers, it's just that that these answers are far less common than natural compatibilists have assumed. So, how might they respond?

The most obvious rejoinder for natural compatibilists is to deny that our intrusion items are properly construed as measures of creeping indeterminism. On this view, our items beg the question against compatibilism and our findings can be given a compatibilist-friendly interpretation. Here, the natural compatibilist is likely to appeal to the difference between the unconditional and the conditional ability to do otherwise. In an indeterministic universe, agents can have the unconditional ability to do otherwise—that is, they could have done otherwise even if everything leading up to their decision remained *exactly the same*. In a deterministic universe, on the other hand, agents merely have the conditional ability to do otherwise—that is, agents could have acted differently only insofar as something (either the past or the laws) had been different than it actually was. Compatibilists suggest that this conditional ability to do otherwise (along with other cognitive and volitional capacities) can ground free will and moral responsibility even in a deterministic universe. Incompatibilists disagree, insisting instead that free will requires indeterminism and the unconditional ability to do otherwise.

We can set the metaphysical debate aside for now. The relevant question is whether our intrusion items are more plausibility interpreted via the unconditional or the conditional lens. We think the unconditional reading is the most natural one. Consider, for instance, Chance, our primary measure of intrusion. Given the details of the scenarios, there was *no chance* (metaphysically speaking) that Jeremy would or could actually do something other than what was predicted at the time. So, the compatibilist has to insist instead that chance in this context refers to the mere *possibility* that things could have been different had the antecedent conditions been different. If we construe chance this way, though, then we must conclude that chance is always operating in the universe since, except under fatalism, there is *always* a chance that things could have been different in the compatibilist sense. This makes the concept of chance entirely vacuous (since it always applies whether the universe is deterministic or indeterministic). And while it's certainly *possible* that this is what participants had in mind when thinking about Chance, we don't think it's *plausible* to think that this is what most of them *actually* had in mind. That's precisely why we focused on Chance more than the other items for measuring intrusion. It seems to us to be the least susceptible to the compatibilist's attempted conditional gloss. When people think about an overtly indeterministic item like Chance, we think the extant research on people's preferences for the unconditional reading of both free will and the ability to do otherwise suggests that most of our participants likely viewed Chance through an unconditional lens (Deery, Bedke, and Nichols, 2013; Nadelhoffer et al., 2014; Nadelhoffer et al., 2019a; 2019b; Nichols, 2006; Wisniewski, Deutschländer, & Haynes, 2019; cf. Nahmias et al. 2004).

It’s also worth noting at this juncture that Slight Chance is very similar to Chance in this regard. The conditional reading is prima facie implausible here as well. That is precisely why we analyzed Slight Chance separately in the previous section—namely, we wanted to see whether it performed much like Chance (which it did). This is important because Slight Chance is one of the items we used for the composite measure—which had excellent reliability (α = .906). Given that Slight Chance is most naturally understood unconditionally and given that the group of items including Slight Chance has excellent reliability, this suggests that it is likely that the other items in the group are also being interpreted unconditionally. If Slight Chance were being interpreted *unconditionally* while the other intrusion items were being interpreted *conditionally*, it is unlikely that they would have fit together reliably as we found. As such, we think that our findings concerning Slight Chance bolster our unconditional reading of the other intrusion items used in the composite measure.

The compatibilist’s conditional reading of our intrusion items runs into similar problems when it comes to explaining our finding that intrusion is more prevalent in agents than in robots. Given our unconditional reading of our intrusion items, we have a readymade explanation of this asymmetry—namely, human agents are more likely to be viewed through the lens of indeterminism while robots are more likely to be viewed through the lens of determinism. Consequently, we predicted intrusion would be more prevalent in the agent cases than in the robot cases. How is the compatibilist supposed to explain this difference? The conditional gloss won’t help. After all, if we adopt the conditional reading of our intrusion items, then it turns out that the robot could have done otherwise, could have avoided the outcome, had other options than robbing the bank, etc. For on the conditional reading, this amounts to nothing more than the fact that had the past or the laws been different, the robot could have and would have done something different. Given that Turri (2017c) found that people don't think robots have the ability do otherwise precisely because they are taken to be more deterministic, this suggests that the notion of ability people are applying in these sorts of cases is the unconditional ability. Otherwise, robots, too, would be judged to have the ability to do otherwise since they clearly have the kind of conditional abilities highlighted by compatibilists.

In short, we believe that our findings suggest that not all ascriptions of free will in the face of determinism reflect an underlying compatibilist metaphysics as natural compatibilists have assumed. Some of these ascriptions reflect instead the intrusion of indeterminism. If this is correct, the natural compatibilist must devise a strategy for separating the compatibilist wheat from the indeterministic chaff if she wants to provide compelling evidence for her view. So, while it is likely true that some participants adopt a conditional reading, to problematize natural compatibilism we need only show that indeterministic intrusion is commonplace. And for that, we need only show that it is more likely than not that a number of participants adopt the unconditional reading that has been our focus in this section.

We've tried to provide both a priori reasons and empirical evidence for thinking that the unconditional reading is the most natural interpretation of most participants' judgments. But most compatibilists will likely still take issue with the wording of our items and our interpretation of the findings. This is a limitation of any study that explores folk judgments about free will because the parties to the debate often claim that certain stimuli and measures beg the metaphysical question in various ways. Because we think Chance is the most difficult intrusion item for compatibilists to explain away, we made it our focus. But if compatibilists have a better way of ensuring that participants are tracking determinism, we are clearly open to suggestions. Given that compatibilists are the ones who claim that most people think that free will and moral responsibility are compatible with determinism, we think the onus is on them to ensure that participants aren't importing indeterminism into the scenarios (just as they believe that incompatibilists have to control for bypassing intuitions). We believe our findings raise the possibility that compatibilists haven't done their due diligence on this front. In this respect, what we've done is throw down an empirical gauntlet. If compatibilists don't like our intrusion items—which is fine as far as it goes—then they will need to come up with items of their own that will succeed where ours purportedly fail. What they can no longer do in the face of our findings is ignore the issue altogether and simply assume that people comprehend the implications of determinism.

A second limitation of our study is that it was limited to participants who are in the United States. While our online sample was more diverse and representative than usual convenience samples drawn from universities, we nevertheless have to avoid making hasty generalizations. After all, our online participants are drawn from a country that is Western, Educated, Industrialized, Rich, and Democratic (WEIRD: Henrich, Heine, & Norenzayan, 2010). While there are some cross-cultural studies that support natural incompatibilism (Sarkissian et al., 2010; Wisniewski, Deutschländer, & Haynes, 2019; cf. Hainnikainen et al., 2019), much work on this front remains to be done. In the meantime, it is worth noting that the bulk of the work on folk intuitions about free will has been done using WEIRD participants. Given that we were trying to respond to and build upon this research, it made sense for us to limit our attention to participants in the United States—which reflects the original sample pools that were used by Nahmias et al.

(2005; 2006). But we are quick to acknowledge that more cross-cultural work is required before we will know whether our findings are robust and stable.

Despite these two limitations, we believe that we have made a valuable contribution to the literature by problematizing two key pieces of evidence that have been used to bolster natural compatibilism. When considered against the backdrop of the extant research that suggests either that incompatibilism is the default view or that compatibilist intuitions are driven by epistemically problematic psychological processes like performance errors, motivated cognition, affective bias, intrusion effects, and the like (Clark, Winegard, & Baumeister, 2019; Feltz et al., 2009; Feltz & Millan 2013; Nadelhoffer et al., 2014; Nadelhoffer et al., 2019a; 2019b; Nichols & Knobe 2007; Rose, 2019; Rose, forthcoming; Rose et al., 2016; Roskies & Nichols, 2008; Sarkissian et al., 2010; Wisniewski et al., 2019), our findings further complicate the project of compatibilists who want to align their view with common sense. This is not to say that the case is closed when it comes to natural compatibilism (see for example Turri, 2017a; 2017b). But if natural compatibilists want to defend their view moving forward, they will need to devise a new strategy for guaranteeing that when participants read deterministic scenarios their intuitions are not inappropriately influenced by an intrusive indeterministic metaphysics.[24]

References

Ayer, A. J. (1954). Freedom and necessity. In A. J. Ayer, *Philosophical essays* (pp. 271–84). London: Macmillan.

[24] We would like to thank Richard Cooper (Executive Editor of Cognitive Science) and three anonymous referees for very constructive feedback. Addressing their concerns greatly improved our paper. We would also like to thank Eddy Nahmias for helpful suggestions on an earlier draft of our paper.

Bear, A., & Knobe, J. (2016). What do people find incompatible with causal determinism? *Cognitive Science*, *40*, 2025–2049.

Bloom, P. (2012). Free will does not exist. So what? *The Chronicle of Higher Education*. Retrieved May 22nd, 2019, from http://chronicle.com/article/Paul-Bloom/131170/

Bourgeois-Gironde, S., Cova, F., Bertoux, M., & Dobois, B. (2012). Judgments about moral responsibility and determinism in patients with behavioral variant of frontotemporal dementia: Still compatibilists. *Consciousness and Cognition*, *21*, 851–864.

Chan, H-Y., Deutsch, M., & Nichols, S. (2016). Free will and experimental philosophy. In J. Sytsma & W. Buckwalter (Eds.), *A companion to experimental philosophy* (pp. 158–172). Hoboken, NJ: John Wiley & Sons, Ltd.

Clark, C. J., Winegard, B. M., & Baumeister, R. F. (2019). Forget the folk: Moral responsibility preservation motives and other conditions for compatibilism. *Frontiers in Psychology*, https://doi.org/10.3389/fpsyg.2019.00215

Cova, F., & Kitano, Y. (2014). Experimental philosophy and the compatibility of free will and determinism: A survey. *Annals of the Japan Association for Philosophy of Science*, *22*, 17–37.

Cover, J. A., & O'Leary-Hawthorne, J. (1996). Free agency and materialism. In J. Jordan & D. Howard-Snyder (Eds.), *Faith, freedom and rationality* (pp. 47–71). Lanham, MD: Roman and Littlefield.

Deery, O., Davis, T., & Carey, J. (2014). The free-will intuitions scale and the question of natural compatibilism. *Philosophical Psychology*, 1–26.

Deery, O., Bedke, M., & Nichols, S. (2013). Phenomenal abilities: Incompatibilism and the experience of agency. In D. Shoemaker, *Oxford studies in agency and responsibility* (pp. 126–150). New York: Oxford University Press.

Ekstrom, L. (2000). *Free will*. Boulder, CO: Westview Press.

Feltz, A., Cokely, E., & Nadelhoffer, T. (2009). Natural compatibilism versus natural incompatibilism: Back to the drawing board. *Mind & Language*, *24*(1), 1–23.

Feltz, A., & Millan, M. (2013). An error theory for compatibilist intuitions. *Philosophical Psychology*, *28*(4), 529–555.

Figdor, C., & Phelan, M. (2015). Is free will necessary for moral responsibility?: A case for rethinking their relationship and the design of experimental studies in moral psychology. *Mind & Language*, *30*(5), 603–627.

Fischer, J., & Ravizza, M. (1998). *Responsibility and control*. New York: Cambridge University Press.

Hainnikainen, I. R., Machery, E., Rose, D., Stich, S., Olivola, C. Y., Sousa, P., et al. (2019). For whom does determinism undermine moral responsibility? Surveying the conditions for free will across cultures. *Frontiers in Psychology*, *10*, 2028. https://doi.org/10.3389/fpsyg.2019.02428

Henrich, J., Heine, S.J., & Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences, *33*(2-3), 61–83.

Knobe, J. (2014). Free will and the scientific vision. In E. Machery & E. O'neill (Eds.), *Current controversies in experimental philosophy* (pp. 69–85). New York: Routledge.

Lycan, W. (2003). Free will and the burden of proof. In A. O'Hear, *Minds and persons: Royal Institute of Philosophy Supplement* (pp. 107–122). Cambridge: Cambridge University Press.

Mandelbaum, E., & Ripley, D. (2012). Explaining the abstract/concrete paradoxes in moral psychology: The NBAR hypothesis. *Review of Philosophy and Psychology*, *3*, 351–368.

May, J. (2014). On the very concept of free will. *Synthese*, *191*(12), 2849–2866.

Monroe, A., 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 & Cognition*, *27*, 100–108.

Murray, D., & Nahmias, E. (2014). Explaining away incompatibilist intuitions. *Philosophy and Phenomenological Research*, *88*(2), 434–467.

Nadelhoffer, T., Shepard, J., Nahmias, E., Sripada, C., & Ross, L. (2014). The Free Will Inventory: Measuring beliefs about agency and responsibility. *Consciousness and Cognition*, *25*, 27–41.

Nadelhoffer, T., Murray, S., & Murray, E. (2019a). Folk intuitions, free will, and the failure to comprehend determinism. [in progress]

Nadelhoffer, T., Yin, S., & Graves, R. (2019b, September 16). Folk intuitions and the conditional ability to do otherwise. https://doi.org/10.31219/osf.io/mx8fy

Nahmias, E. (2006). Folk fears about freedom and responsibility: Determinism vs. reductionism. *Journal of Cognition and Culture*, *6*, 215–237.

Nahmias, E. (2011). Intuitions about free will, determinism, and bypassing. In R. Kane (Ed.), *The Oxford handbook on free will*: *Second* Edition (pp. 555–576). New York: Oxford University Press.

Nahmias, E. (2014). Is free will an illusion? Confronting challenges from the modern mind sciences. In W. Sinnott-Armstrong (Ed.), *Moral psychology, Vol. 4: Freedom and responsibility* (pp. 1–25). Cambridge, MA: MIT Press.

Nahmias, E., Allen, C., & Loveall, B. (Forthcoming). When do robots have free will? Exploring the relationships between (attributions of) consciousness and free will. In B. Feltz, M. Missal, & Sims, A. *Free will, causality, and neuroscience*. Boston: Brill Publishers.

Nahmias, E., Coates, J., & Kvaran, T. (2007). Free will, moral responsibility, and mechanism: Experiments on folk intuitions. *Midwest Studies in Philosophy, 31*, 214–242.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2004). The phenomenology of free will. *The Journal of Consciousness Studies*, *11*, 162–179.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying free will: Folk intuitions about free will and moral responsibility. *Philosophical* Psychology, *18*(5), 561–584.

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

Nahmias, E., & Murray, D. (2011). Experimental philosophy on free will: An error theory for incompatibilist intuitions. In J. Aguilar, A. Buckareff, & K. Frankish (Eds.), *New waves in philosophy of action* (pp. 189–216). New York: Palgrave-Macmillan.

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. (2006a). Folk intuitions on free will. *Journal of Cognition and Culture*, *6*, 57–86.

Nichols, S. (2006b) Free will and the folk: response to commentators. *Journal of Cognition and Culture*, *6*, 305–320.

Nichols, S. (2012). The indeterminist intuition. *The Monist*, *95*(2), 290–307.

Nichols, S. (2015). *Bound: Essays on free will and responsibility*. New York: Oxford University Press.

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

Nowell-Smith, P.H. (1949). Free will and moral responsibility. *Mind*, *57*, 45–65.

O'Connor, T. (2000). *Persons and causes: The metaphysics of free will*. New York: Oxford University Press.

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

Rose, D. (2017). Folk intuitions of actual causation: a two-pronged debunking explanation. *Philosophical Studies*, *174*(5), 1323–1361.

Rose, D. (2019). Cognitive science for the revisionary metaphysician. In Goldman, A., & Mclaughlin, B. (eds), Metaphysics and Cognitive Science. Oxford University Press. DOI:10.1093/oso/9780190639679.003.0015

Rose, D. (forthcoming). Mentalizing objects. In Lombrozo, T., Knobe, J., & Nichols, S. (eds.), Oxford Studies in Experimental Philosophy Volume 4. Oxford University Press.

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

Rose, D., & Nichols, S. (2019a). From punishment to universalism. *Mind & Language*, *34*, 59–72.

Rose, D., & Nichols, S. (2019a). Teleological essentialism. *Cognitive Science*, *43*, e12725. **https://doi.org/10.1111/cogs.12725**

Rose, D., & Nichols, S. (2020). Teleological essentialism: Generalized. Cognitive Science, 44(3) , e12818. **https://doi.org/10.1111/cogs.12818**

Rose, D., Buckwalter, W., & Nichols. (2017). Neuroscientific prediction and the intrusion of intuitive metaphysics. *Cognitive Science*, *41*(2), 482–502.

Rose, D., Schaffer, J., & Tobia, K. (forthcoming). Folk teleology drives persistence judgments. *Synthese*.

Rose, D., Livengood, J., Sytsma, J., & Machery, E. (2012). Deep trouble for the deep self. *Philosophical Psychology*, *25*(5), 629–646.

Roskies, A., & Nichols, S. (2008). Bringing moral responsibility down to Earth. *The Journal of Philosophy*, *105*(7), 371–388.

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.

Searle, J. (1984). *Minds, brains, and science*. Cambridge, MA: Harvard University Press.

Shepherd, J. (2015) Consciousness, free will, and moral responsibility: taking the folk seriously. *Philosophical Psychology, 28*, 929–946.

Sinnott-Armstrong, W. (2008). Abstract + concrete = paradox. In J. Knobe & S. Nichols (Eds.), *Experimental Philosophy* (pp. 209–230). New York: Oxford University Press.

Smilansky, S. (2003). Compatibilism: The argument from shallowness. *Philosophical Studies*, *115*(3), 257–282.

Stace, W. T. (1952). *Religion and the modern mind*. New York: Lippincott.

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.

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

Turri, J. (2017a). Compatibilism can be natural. *Consciousness and Cognition*, *51*, 68–81.

Turri, J. (2017b). Compatibilism and incompatibilism in social cognition. *Cognitive Science, 41*(S3), 403–424.

Turri, J. (2017c). Exceptionalist naturalism: Human agency and the causal order. *The Quarterly Journal of Experimental Psychology*, 1–16. https://doi.org/10.1080/17470218.2016.1251472

Turri, J., Buckwalter, W., & Rose, D. (2016). Actionability judgments cause knowledge judgments. *Thought*, *5*(3), 212–222.

Wisniewski, D., Deutschländer, R., & Haynes, J-D. (2019). Free will beliefs are better predicted by dualism than determinism beliefs across different cultures. *PLoS ONE*, *14*(9), e0221617. https://doi.org/10.1371/journal.pone.0221617

Wolf, S. (1990). *Freedom within reason*. New York: Oxford University Press.

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*, *100*, 283–301.