Do People Understand Determinism? The Tracking Problem for Measuring Free Will Beliefs

Samuel Murray[a], Elise Dykhuis[b], and Thomas Nadelhoffer[c]

[a]Department of Philosophy, Providence College, Providence, RI, USA
[b]Department of Mathematical Sciences, United States Military Academy, West Point, NY, USA
[c]Department of Philosophy, College of Charleston, Charleston, SC, USA

Forthcoming in *Oxford Studies in Experimental Philosophy*, vol. 5

Author Note

Correspondence may be directed to Samuel Murray, Department of Philosophy, Providence College, 1 Cunningham Sq., Providence, RI 02908. Email: sfm18@duke.edu

Abstract

Experimental work on free will typically relies on deterministic stimuli to elicit judgments of free will. We call this the Vignette-Judgment model. We outline a problem with research based on this model. It seems that people either fail to respond to the *deterministic* aspects of vignettes when making judgments or that their understanding of determinism differs from researcher expectations. We provide some empirical evidence for this claim. In the end, we argue that people seem to lack facility with the concept of determinism, which calls into question the validity of experimental work operating under the Vignette-Judgment model. We also argue that alternative experimental paradigms are unlikely to elicit judgments that are philosophically relevant to questions about the metaphysics of free will.

Keywords: free will, determinism, error, intuitions, experimental philosophy

1. Thought Experiments and the Folk Psychology of Free Will

Philosophical reflection on free will aims, broadly, to answer two questions:

> *The Compatibility Question*: Is free will compatible with determinism?
> *The Traditional Question*: Does anything have free will?[1]

Answers to these questions define the conceptual landscape of the free will debate. Those who answer ‘yes’ to The Compatibility Question are *compatibilists*, while those who answer ‘no’ are *incompatibilists*. Incompatibilists who answer ‘yes’ to The Traditional Question are *libertarians*, while those who answer ‘no’ are *hard incompatibilists*.[2]

Many arguments for or against positions on The Compatibility Question utilize thought experiments. These thought experiments typically require imagining activity in some deterministic situation (or series of situations) and judging whether individuals in that situation are free. Thought experiments are pervasive. *Manipulation* arguments attempt to show that some individual can satisfy compatibilist conditions on free agency within a deterministic scenario and still act freely (Kane, 1996; Mele, 2006; Pereboom, 2001). *Ability* arguments attempt to show that determinism precludes powers of practical reasoning necessary for having free will (Taylor, 1966; van Inwagen, 1978). *Intervener* arguments attempt to show that agents can act freely even when they lack the ability to do otherwise (Frankfurt, 1971; Sartorio, 2020). Thought experiments, then, can be a valuable tool for understanding free will, especially with respect to The Compatibility Question.[3]

[1] van Inwagen (1983, p. 2).

[2] Cmpatibilists typically affirm the existence of free will. This schema leaves out those who argue that having free will is impossible (e.g., Strawson, 1986). This differs from hard incompatibilists, who hold that people *could* have free will, but the conditions for having it are not satisfied in our universe (Pereboom, 2001).

[3] A notable exception might be van Inwagen’s Consequence Argument, the conclusion of which is that determinism is incompatible with free will. Roughly, the Consequence Argument has the following form:

1) $N(P \& L)$
2) $\Box((P\&L) \rightarrow A)$
3) $N(A)$

Informally, the argument reads: Nobody has a choice about whether the distant past (before one’s birth) and the laws of nature are the way they are. Given determinism, the conjunction of a description of the past and a description of the laws at some time entails a description of the universe at every future time. Thus, if determinism is true, and nobody

When thought experiments play a pivotal role in philosophical argumentation, it can be useful to test the robustness of reactions to such experiments. Experimental philosophers have recently examined the variability of reactions to thought experiments across demographic factors (Nichols, 2011; Machery, 2017; Knobe, 2019). This experimental research can be useful in two ways. First, it can diagnose the referents of theoretically significant terms, such as 'cause' or 'free' (Vargas, 2017). Second, it can diagnose idiosyncratic or parochial reactions to thought experiments. Sometimes, these run together. Philosophers might employ a technical sense of 'free' that generates idiosyncratic reactions to thought experiments. Unchecked, we risk spinning out theories of phenomena that make no contact with people's ordinary experiences and categories (Dennett, 2006). This applies to theorizing about free will, where theories incorporate terms such as 'choice' or 'ability' that have some purchase in everyday discourse (Vargas, 2013).

Experimental work on free will beliefs has been guided by the *Vignette-Judgment Model*. On this model, individual beliefs about free will are measured by eliciting judgments in response to vignettes that encode some deterministic element. The content of free will beliefs is inferred by measuring how judgments change based on changes to the target content. The Vignette-Judgment

has a choice about what the past or laws are like, then nobody has a choice about whether the future is the way that it is. Van Inwagen interprets having a choice about whether *p* in terms of being able to render the proposition that *p* false, which interpretation he claims captures some of our intuitions about decision-making (1983, pp. 66-68). But what is this intuitive view of choice-making? *An Essay on Free Will* does not say, but elsewhere van Inwagen uses thought experiments to illustrate that the ability to choose to bring it about that *p* requires being able to bring it about either that *p* or that *~p*, and that one cannot bring it about either that *p* or that *~p* if it is inevitable that *p* (1978, pp. 215-16).

Moreover, van Inwagen uses thought experiments to counter some objections to the Consequence Argument. Lewis (1981), for example, claims that we can understand the phrase 'J was able to render *p* false' in three different ways:

A) J was able to do something such that had he done it, L [the description of the laws of nature] would be false.
B) J was able to do something such that his doing it would cause something that falsifies L.
C) J was able to do something such that his doing it (holding fixed the past) falsifies L.

If we understand the *being able to render false* relation in terms of either (A) or (B), then—Lewis argues—the Consequence Argument is unsound. (C) makes the argument sound at the expense of making it question-begging, as (C) interprets ability in a way that presupposes incompatibilism. Van Inwagen defends (C) as the correct sense of *being able to render false* and rejects the charge of question-begging by appealing to a thought experiment (van Inwagen 2004, p. 349; see also Murray & Nahmias: 2014, p. 457n33).

Model provides a procedure for sorting people within the conceptual landscape related to the Compatibility Question. Natural compatibilists tend to attribute free will or responsibility to individuals depicted in deterministic scenarios, while natural incompatibilists tend to withhold attributing free will or responsibility under those conditions. Presumably, this is because the former, but not the latter, think that having free will is compatible with the truth of determinism.

In practice, judgments elicited by different vignettes paint a complex picture about attitudes toward the Compatibility Question (May, 2014). Some have found that people tend to attribute free will to individuals depicted in deterministic vignettes (Nahmias et al., 2005; Feltz et al., 2009), while others have found opposed reactions in response to different vignettes (Nichols and Knobe, 2007; Nadelhoffer, Rose et al., 2020). Some have interpreted these results to mean that people utilize distinct conditions for attributing free will, some of which reflect compatibilist commitments and some of which do not (Knobe and Doris, 2010). Others have argued that people are fundamentally motivated to blame individuals for wrongdoing and adopt whatever conditions for free will satisfy the desire to punish (Clark et al., 2019). Thus, some see the variability in participant judgments as a feature of our thinking about free will rather than a bug.

Others have proposed error theories to explain away judgments that support countervailing conceptions of free will. The strategy is to show that these judgments are based either on peripheral content of the vignette or a misunderstanding of the target content. For example, judgments that seem to reflect incompatibilist commitments might reflect a misunderstanding of determinism as entailing epiphenomenalism or fatalism (Nahmias and Murray, 2011; Murray and Nahmias, 2014). On the other hand, judgments that seem to reflect compatibilist commitments might stem from participants importing indeterminism into the vignettes despite the stimuli being deterministic (Rose et al., 2017; Nadelhoffer, Rose et al., 2020).

This highlights an important feature of the Vignette-Judgment model. The theoretical value of some judgment depends on participant interpretations aligning with researcher expectations. Researchers design materials to exemplify scenarios that are philosophically relevant. Participants are assumed to interpret this content as the researcher intends. Folk judgments elicited by some target content can indicate underlying philosophical commitments only when such judgments are prompted by an accurate apprehension of that content (where accuracy, again, is a function of congruence with researcher intentions).

From this, error theories of free will judgments share a common purpose: identifying ways in which participant interpretations of vignettes might depart from researcher intentions. This raises a question that bears on the prospects of experimental work on free will beliefs—and experimental work on folk conceptual commitments more generally—namely, how we can tell whether participants are responding to target content in a way that aligns with researcher expectations.

The problem has been noted before, most recently by Nadelhoffer, Rose et al. (2020): "From the standpoint of experimental design, putting compatibilism to the test requires that researchers ensure that participants' intuitions are sufficiently responsive to the deterministic nature of the scenarios" (p. 3). This is *The Tracking Problem*. While they restrict the tracking problem to measuring compatibilist commitments, we believe the problem can be stated generally for any experimental work under the Vignette-Judgment Model.

2. The Tracking Problem

Suppose that participants are tracking the target content of vignettes in making judgments about free will, as the Vignette-Judgment Model presumes. A contradiction can be derived from this assumption with some additional premises.

TP-1. Participant judgments elicited by vignettes used in experimental philosophical research track the target content of the vignettes.
TP-2. The target content of vignettes used in experimental research on free will represents the philosophical notion of determinism (hereafter, determinism).[4]
TP-3. Participant judgments of free will elicited by deterministic vignettes track determinism. [1, 2]
TP-4. If participant judgments track determinism, then participants grasp the concept of determinism.
TP-5. If participants grasp the concept of determinism, then participants can draw central inferences about determinism.
TP-6. Participants cannot draw central inferences about determinism.
TP-7. Therefore, participant judgments of free will elicited by deterministic vignettes do not track determinism.[3, 4, 5, 6]

This is the Tracking Problem stated generally. This formulation of the problem defines the space of possible responses. Rejecting (TP-1) is the *Nihilist response*, as this would entail the theoretical irrelevance of experimental research under the Vignette-Judgment Model. Rejecting (TP-2) is the *Miscommunication response*, where researchers fail to properly encode determinism in the materials used to elicit judgments. Rejecting (TP- 4) is the *Nonconceptualist response* to the Tracking Problem, while rejecting (TP-5) is the *Noninferentialist response*. Finally, rejecting (TP-6) is the *Competence response*.

Below, we mount an empirical case against the *Competence response*. With the evidence for (TP-6) in place, we discuss whether the nonconceptualist or noninferentialist responses are plausible. We argue that, for determinism, neither response seems plausible. This leaves either the nihilist or miscommunication responses.

3. Grasping determinism

We think a strong empirical case can be made for denying that people can draw central inferences about determinism associated with their grasping the concept. The argument is straightforward:

[4] Put differently, the term ‘determinism’ as it occurs in The Compatibility Question and the target content of the vignette denote the same proposition.

1) People agree with statements that are not true in deterministic universes.
2) Agreeing with these statements constitutes making an error about the nature and implications of determinism
3) Therefore, people make several errors about the nature and implications of determinism.
4) If people could draw central inferences about the concept of determinism, they would not make these errors.
5) Therefore, people cannot draw central inferences about determinism.

To assess the first premise, we ran a study that utilized standard materials from previous research on judgments of free will in experimental philosophy. We also developed novel measures to assess whether participants make correct inferences about these materials based on their deterministic nature (Nadelhoffer et al., In press). We tested whether participants could draw three correct inferences about our scenarios. Only 3% of our sample (16/556) made all three inferences correctly. 34% of our sample (190/556) did not make any correct inferences.

Materials, data, preregistrations, power analyses, and analysis scripts are available in the Supplementary Materials and on the OSF project page: <https://osf.io/gzqk2/>. This study was approved by the Institutional Review Board at the College of Charleston and run via Amazon's Mechanical Turk.

3.1 Methods

We initially ran three separate experiments to examine inferences about determinism. However, because of similar design, we report methods and results for these experiments as a single study. Details about power analyses for determining sample size can be found in Supplementary Materials §1.

3.1.1 Participants

659 participants were recruited to participate in the study. Because of simultaneous enrollment, 663 participants submitted responses. Per our pre-registered exclusion criteria, 46 participants were excluded for failing attention checks, 43 were excluded for failing comprehension checks, 17 were excluded for inconsistent responding, and 1 participant did not finish the experiment. Data from 556 participants was analyzed (M_{age} = 38.7, SD_{age} = 11.1, 41% female, 81% Caucasian).

3.1.2 Materials and procedures

Participants were randomly assigned to read one of seven vignettes (described below):

> *Supercomputer* (from Nahmias et al., 2005):
> 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 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.
>
> *Universe A/B* (from Nichols & Knobe, 2007):
> Imagine a universe (Universe A) in which everything that happens is completely caused by whatever happened before it. This is true from the very beginning of the universe, so what happened in the beginning of the universe caused what happened next, and so on right up until the present. For example one day John decided to have French Fries at lunch. Like everything else, this decision was completely caused by what happened before it. So, if everything in this universe was exactly the same up until John made his decision, then it had to happen that John would decide to have French Fries.
>
> Now imagine a universe (Universe B) in which almost everything that happens is completely caused by whatever happened before it. The one exception is human decision making. For example, one day Mary decided to have French Fries at lunch. Since a person's decision in this universe is not completely caused by what happened before it, even if everything in the universe was exactly the same up until Mary made her decision, it did not have to happen that Mary would decide to have French Fries. She could have decided to have something different.

The key difference, then, is that in Universe A every decision is completely caused by what happened before the decision— given the past, each decision has to happen the way that it does. By contrast, in Universe B, decisions are not completely caused by the past, and each human decision does not have to happen the way that it does.

Rollback Concrete Bad (from Nadelhoffer, Rose et al., 2020):
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 Jim decides to rob a bank at a particular time, and every time the universe is re-created, Jim decides to rob a bank at that time.

Rollback Abstract (from Nahmias et al., 2006):
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.

Determinism [Actual World] (from Roskies and Nichols, 2008):
Many eminent scientists have become convinced that every decision a person makes is completely caused by what happened before the decision given the past, each decision has to happen the way that it does. These scientists think that a person's decision is always an inevitable result of their genetic makeup combined with environmental influences. So if a person decides to commit a crime, this can always be explained as a result of past influences. Any individual who had the same genetic makeup and the same environmental influences would have decided exactly the same thing. This is because a person's decision is always completely caused by what happened in the past.

Determinism [Alternate World] (from Roskies and Nichols, 2008):
Imagine an alternate universe, Universe A, that is much like earth. But in Universe A, many eminent scientists have become convinced that in their universe, every decision a person makes is completely caused by what happened before the decision - given the past, each decision has to happen the way that it does. These scientists think that a person's decision is always an inevitable result of their genetic makeup combined with environmental influences. So if a person decides to commit a crime, this can always be explained as a result of past influences. Any individual who had the same genetic makeup and the same environmental influences would have decided exactly the same thing. This is because a person's decision is always completely caused by what happened in the past.

Determinism [Conditional Ability] (from Nadelhoffer, Yin, and Graves, 2020):
Imagine Jim lives in a causally closed universe. In this universe, given the physical state of the universe, the laws of the universe, and the fixity of the past, at any given moment the universe is closed, like a train moving down the tracks. Whenever Jim makes a decision

> to act in a particular way, it’s always the case that he could have acted differently only if something leading up to his decision had been different. In short, at any given moment, there is one and only one choice and action genuinely open to Jim. Moreover, if you knew absolutely everything about both the history of the universe and about Jim, you could always know in advance what Jim is going to decide to do. He is not the only deciding factor when it comes to what he does. Given the way the world was long before Jim was born, everything in his life is in the cards, so to speak. Jim can make choices, but these choices are the only choices open to him. Now, for illustrative purposes, imagine that Jim decides to rob a bank.

Participants completed items about the similarity between the deterministic scenario and the actual world, the possibility of the deterministic scenario, how vividly they imagined the scenario, and a comprehension check:

> *Similarity*: How similar do you think this universe is to our own universe? (6-pt. scale, 1 = very dissimilar, 2 = dissimilar, 3 = somewhat dissimilar, 4 = somewhat similar, 5 = similar, 6 = very similar)[5]
>
> *Possibility*: Do you think this scenario is possible? [Yes/No]
>
> *Vividness*: How vividly could you imagine the previously described scenario? (5-pt. scale, 1 = very slightly or not at all, 2 = a little, 3 = moderately, 4 = quite a bit, 5 = extremely).
>
> *Comprehension*: According to the scenario, when the universe is re-created over and over again, the same initial conditions and the same laws of nature produce the exact same outcomes every time. [True/False][6]

These items were presented in random order. Afterward, participants were instructed to imagine the scenario was real regardless of how they answered the previous questions. Participants then registered judgments of free will and moral responsibility using a 7-pt. scale (1 = strongly disagree, 4 = neither agree nor disagree, 7 = strongly agree):

> *Free will*: People in this scenario can act of their own free will.

[5] Nichols and Knobe (2007) used a dichotomous similarity probe. To match this procedure, the similarity probe in this condition was dichotomous rather than continuous.

[6] This is the comprehension question for the Rollback scenarios. Each comprehension question was drawn from the original studies that used the vignette. For a full list of comprehension questions, see the section on Stimuli & Measures in the Supplementary Materials (§2).

Moral responsibility: People in this scenario can be morally responsible for their actions.

Participants then completed items taken that measure different errors about the nature or implications of determinism using a 7-pt. scale (1 = strongly disagree, 4 = neither agree nor disagree, 7 = strongly agree). The wording of the items varied according to the content of the vignette:

Bypassing items

1. It doesn't make any sense to say that Jeremy made his own choice to rob the bank.
2. It's not up to Jeremy whether or not to rob the bank.[7]

Fatalism items

1. Jeremy would have ended up robbing the bank no matter what he tried to do.
2. Jeremy will rob the bank no matter what.

Intrusion items

1. There was at least a slight chance that Jeremy could have chosen not to rob the bank even if everything (including the laws of nature) had been exactly the same prior to his decision.
2. It was open for Jeremy to choose not to rob the bank at the exact moment he decided to rob it.

The error questions were taken from Nadelhoffer et al. (In press), who used 12 items (4 per error category). We selected these items based on factor analyses (for details on item selection, see Supplementary Materials §3). The order of items was randomized across participants.

3.2 Results

3.2.1 Cross-study comparisons

[7] These items might seem to measure whether participants believe that agents in deterministic universes can be the ultimate sources of their decisions, rather than whether participants believe that mental states can be causally efficacious in a deterministic universe. However, in Nadelhoffer et al. (In press), both items were highly correlated with a different bypassing item: "What Jeremy wants and believes has no effect on what he does." Additional analyses (reported in Supplementary Materials §3) suggest that these three items measure the same construct. This provides evidence that participants interpret the notion of 'making one's own choice' and 'being up to you' in terms of causality rather than ultimacy.

Table 1 summarizes comparisons to previous findings (full discussion of comparisons are in Supplementary Materials §4).

Table 1. Summary of cross-study comparisons

Vignette	Source	Original finding[8]	Current finding[9]
Universe A/B _N_ = 75	Nichols & Knobe (2007) _N_ = 19	_Most similar_ = 90% indeterminism _Possibly responsible_ = 14% agree	*_Most similar_ = 60% indeterminism *_Possibly responsible_ = 51% agree
Supercomputer _N_ = 73	Nahmias et al. (2005) _N_ = 21	_Responsible_ = 83% agree _Free will_ = 76% agree	**_Responsible_ = 86% agree **_Free will_ = 70% agree
Rollback concrete _N_ = 71	Nahmias et al. (2006) _N_ = 86[10]	_Responsible_ = 77% agree _Free will_ = 66% agree	*_Responsible_ = 52% agree **_Free will_ = 54% agree
Rollback abstract _N_ = 84	Nadelhoffer, Rose et al. (2020) _N_ = 116	_Free will_ = 3.67 (_SD_ = 2.4) _Responsible_ = 4.40 (_SD_ = 2.2)	*_Free will_ = 4.46 (_1.9_) *_Responsible_ = 5.10 (_1.6_)
Determinism conditional _N_ = 80	Nadelhoffer, Yin et al. (2020) _N_ = 78	_Free will_ = 3.02 (_1.3_) _Responsible_= 3.84 (_1.5_)	*_Free will_ = 4.72 (_1.7_) *_Responsible_ = 4.95 (_1.5_)
Determinism Alternate _N_ = 88	Roskies & Nichols (2008) _N_ = 38	_Impossible responsibility_[11] = 5.06[12] _Appropriate blame_ = 3.67 _Impossible free will_ = 5.30	*_Responsible_ = 4.24 (_1.9_) *_Free will_ = 3.77 (_2.0_)
Determinism Actual _N_ = 85	Roskies & Nichols (2008) _N_ = 38	_Impossible responsibility_ = 3.58 _Appropriate blame_ = 5.35 _Impossible free will_ = 4.30	*_Responsible_ = 4.67 (_1.8_) **_Free will_ = 3.99 (_1.8_)

[8] Studies that reported percentages used dichotomous probes to measure judgments of free will and responsibility. In these cases, for purposes of comparison, we counted any participant who responded with a 5 or higher on the free will or moral responsibility items as agreeing either that some individual is free (or responsible) or that free will (or responsibility) is possible in a deterministic universe. Subsequent analyses do not utilize these dichotomized measurements.

[9] Current findings that _fail_ to align with some aspects of the original research are marked with *. Findings that _align_ with original findings are marked with **. These differences are discussed in Supplementary Materials §4. Because our primary aim was neither direct nor conceptual replication, we altered some measures and, at times, used different procedures. Hence, we do not claim that our results replicate or fail to replicate past results.

[10] Data on sample size was provided by Eddy Nahmias in personal correspondence.

[11] Roskies & Nichols (2008) asked whether full moral responsibility is possible in the deterministic universe they described. This differs from our item, which asked whether people can be morally responsible for their actions.

[12] Standard deviations could not be calculated based on data provided in Roskies & Nichols (2008). Roskies and Nichols measured belief in the _impossibility_ of free will and responsibility. Hence, higher scores indicated stronger belief in impossibility. Our scales were arranged so that higher magnitudes index greater attribution.

The purpose of this study was neither direct nor conceptual replication. As such, we altered some of our materials. Cross-study comparisons show that we failed to find evidence for results that align with several findings from past studies (see Supplementary Materials §4).

3.2.2 Summary statistics

Means and standard deviations for judgments of free will, moral responsibility, and errors are summarized in **Table 2**. Proportion of participants who agreed with error measures are summarized for each category, as well as how many participants agreed with all three error measures. If participant responses averaged *greater than 4*, we counted them as making an error. If participant responses averaged *lower than 4*, they were counted as not making an error.[13] Total Fail reflects the number of participants who made an error on all three categories.

[13] Participants who averaged 4 were not counted because the midpoint represented uncertainty or indifference.

Table 2. Means, standard deviations, and error rates

Vignette	Count	Free will	Moral Responsibility	Bypassing	Bypassing Miss	Fatalism	Fatalism Miss	Intrusion	Intrusion Miss	Total Fail
Universe A/B	75	3.43 (*1.9*)	4.16 (*1.8*)	5.69 (*1.1*)	64 (85%)	6.03 (*0.8*)	73 (97%)	3.57 (*2.0*)	42 (56%)	29 (39%)
Supercomputer	73	5.25 (*1.7*)	5.70 (*1.3*)	4.30 (*1.7*)	42 (58%)	5.18 (*1.4*)	55 (75%)	4.83 (*1.5*)	49 (67%)	28 (38%)
Rollback abstract	71	4.23 (*2.0*)	4.99 (*1.8*)	5.15 (*1.4*)	53 (75%)	5.47 (*1.2*)	61 (86%)	3.95 (*1.9*)	38 (54%)	30 (42%)
Rollback concrete	84	4.46 (*1.9*)	5.10 (*1.6*)	4.56 (*1.6*)	50 (60%)	5.48 (*1.3*)	71 (85%)	4.29 (*1.7*)	50 (60%)	24 (29%)
Determinism conditional	80	4.72 (*1.7*)	4.95 (*1.5*)	4.50 (*1.5*)	43 (54%)	4.81 (*1.4*)	51 (64%)	4.42 (*1.6*)	51 (64%)	25 (31%)
Determinism Alternate	88	3.77 (*2.0*)	4.24 (*1.9*)	5.22 (*1.4*)	66 (75%)	5.37 (*1.3*)	72 (82%)	3.63 (*1.8*)	49 (56%)	23 (26%)
Determinism Actual	85	3.99 (*1.8*)	4.67 (*1.8*)	5.09 (*1.3*)	64 (75%)	5.55 (*1.1*)	75 (88%)	4.09 (*1.8*)	45 (53%)	31 (36%)

Across all conditions, participants tended to agree with Bypassing items (*M* = 4.93, *SD* = 1.5), Fatalism items (*M* = 5.41, *SD* = 1.3), and Intrusion items (*M* = 4.10, *SD* = 1.8) (see **Figure 1**). 69% of participants (382/556) failed Bypassing, 82% (458/556) failed Fatalism, and 55% (305/556) failed Intrusion. 22% of participants (125) failed only one category, 40% (225) failed two, and 34% (190) failed all three. Less than 3% (16 participants) passed all three measures. **Figure 2** summarizes average error score for each vignette.

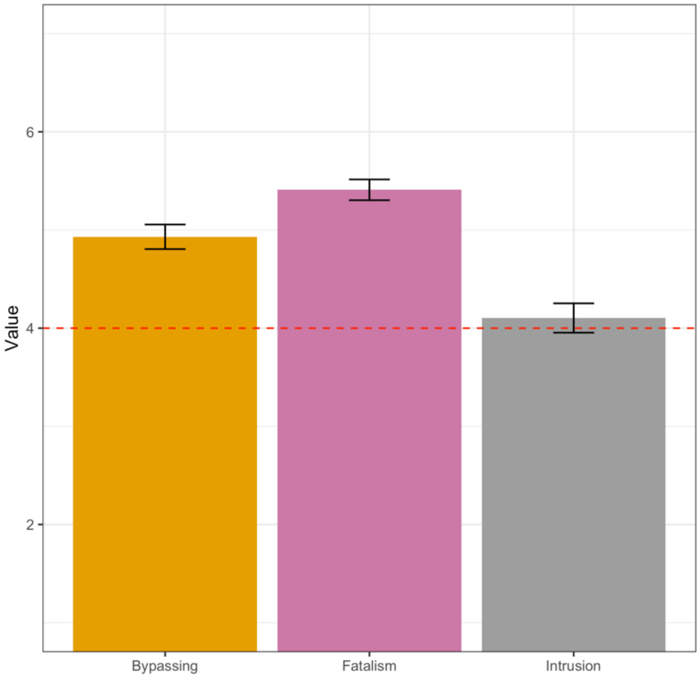

Figure 1*. Mean error score across all conditions. Error bars represent 95% confidence intervals.*

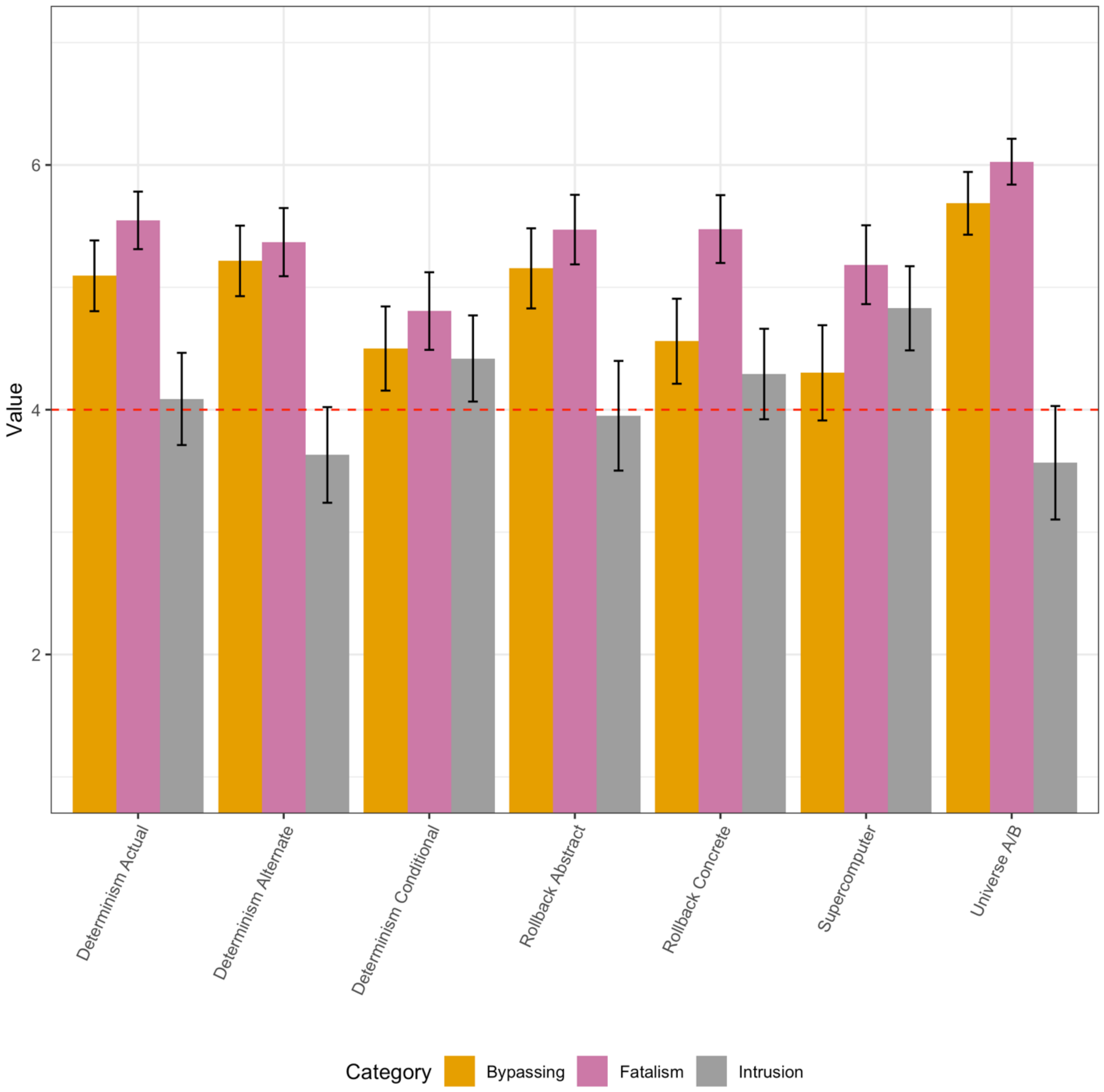

Figure 2. Mean score on error items by vignette. Error bars represent 95% confidence intervals.

3.2.3 Reliability and invariance of error measures

We computed split-half reliability to assess the internal consistency of error measures. Reliability coefficients were computed using the Spearman-Brown Prophecy formula, which generates better estimates of internal reliability for two-item scales than Cronbach's alpha (de Vet et al., 2017). Each category exhibited good internal reliability (Bypassing: *r* = .77; Fatalism: *r* = .69; Intrusion:

$r = .85$).[14] A confirmatory factor analysis showed that a three-factor model (where each category mapped to a distinct latent variable) displayed the best model fit (*CFI* = .995, *RMSEA* = .041, 90% *CI*[.00., .076], *SRMR* = 0.01; see Supplementary Materials §5.1).[15]

One-way ANOVA tests identified small effects of vignette on mean Bypassing ($F(6, 549) = 8.93$, $\eta^2 = .09$, 90% *CI*[.05, .12], $p < .001$), mean Fatalism ($F(6, 549) = 7.00$, $\eta^2 = .07$, 95% *CI*[.03, .10], $p < .001$), and mean Intrusion ($F(6, 549) = 4.97$, $\eta^2 = .05$, 95% *CI*[.02, .08], $p < .001$). To assess whether vignettes modulate responses to the error measures, we tested for measurement invariance using structural equation modeling with the *lavaan* package in R (Rosseel, 2012). To determine measurement invariance across groups, aspects of the model structure (as defined in the CFA) are specified to be equal across the vignettes while retaining good model fit, that is, whether the structure is similar across vignettes (configural invariance), whether factor loadings are similar across vignettes (metric/weak invariance), and whether the item intercepts are similar across vignettes (scalar/strong invariance).

To evaluate whether these measures and their structure are invariant, we compared X^2 statistics between the various models as well as the change in *CFI* statistics. If the X^2 comparison is $p > .05$ *or* the change in the CFI < .02, then the model exhibits invariance. We found that our model exhibits configural, weak, and strong invariance, but not strict invariance.[16] From this, we can assume that the same constructs are being measured across vignettes (measurement invariance statistics are summarized in Supplementary Materials §5.4 Table S.13).

[14] Coefficients greater than .60 indicate acceptable reliability, while those greater than .70 indicate good reliability (de Vet et al., 2017).

[15] Hu & Bentler (1999) suggest the following statistical thresholds for model fit: RMSEA < .06, SRMR < .08, and a CFI > .90. Model test statistics are summarized in Supplementary Materials Table S.4.

[16] Although a required component for full factorial invariance (Meredith, 1993), testing for residual invariance is not a prerequisite for testing mean differences because the residuals are not part of the latent factor, so invariance of the item residuals is inconsequential to interpretation of latent mean differences (Vandenberg and Lance, 2000). Thus, many researchers omit this step. We included it because residual invariance is still reported in many tests of measurement invariance (see Supplementary Materials §5.4 Table S.13).

3.2.4 Relationship between error measures and judgment

To better understand potential interactions between error and judgments of free will, we fitted two simple linear models predicting judgments of free will and moral responsibility, respectively, with bypassing, fatalism, intrusion, and vignette as predictors. The models also included all interaction terms.

An ANOVA identified a large effect of intrusion errors on free will judgments ($F(1, 505) = 404.27$, $p < .001$, $\eta^2_p = .44$, 90% $CI[.39, .49]$) and a medium effect of bypassing errors on free will judgments ($F(1, 505) = 24.58$, $p < .001$, $\eta^2_p = .05$, 90% $CI[.02, .08]$), qualified by an interaction between bypassing and intrusion errors ($F(1, 548) = 47.94$, $p < .001$, $\eta^2_p = .07$, 90% $CI[.04, .10]$) (see **Figure 3a**). There was no evidence for a main effect of fatalism errors ($p = .43$) or a two-way interaction between bypassing and fatalism errors ($p = .70$) or intrusion and fatalism errors ($p = .08$). We also found no evidence for any significant three-way interactions (all $p > .18$).[17]

[17] There was evidence for a significant four-way interaction ($F(3, 505) = 2.77$, $p = .04$, $\eta^2_p = .02$, 90% $CI[.00, .03]$). Full pairwise comparisons are reported in Supplementary Materials §5.2.

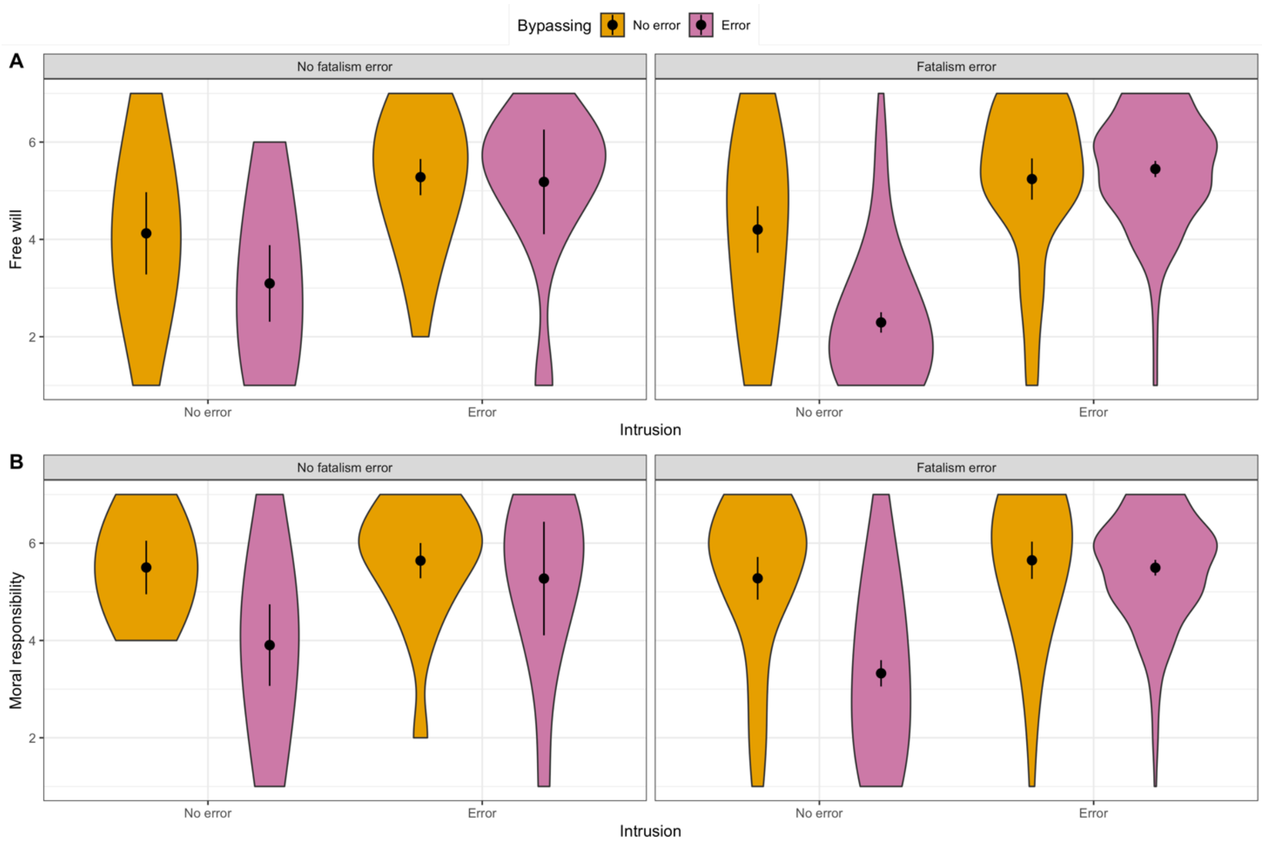

Figure 3. Average free will (Panel A) and responsibility (Panel B) ratings across error categories and error type. Error bars represent 95% confidence intervals.

To clarify the interaction between bypassing and intrusion, we computed tests of simple main effects. Participants who made no intrusion errors attributed significantly more free will when they made no bypassing errors ($M = 4.16$, 95% CI[3.78, 4.55]) compared to those who made bypassing errors ($M = 2.69$, 95% CI[2.38, 3.01]) ($t(548) = 5.81$, $p < .001$, $d = 1.07$, 95% CI[0.70, 1.43]). However, when participants made intrusion errors, there was no evidence for a difference in judgments of free will when they made no bypassing errors ($M = 5.26$, 95% CI[4.99, 5.53]) compared to when they made bypassing errors ($M = 5.31$, 95% CI[4.89, 5.73]) ($t(548) = -0.21$, $p = .83$, $d = -0.04$, 95% CI[-0.40, 0.32]).[18]

[18] This analysis papers over potentially interesting differences between different kinds of errors. For example, there might be interesting differences between individuals who make only intrusion errors compared to individuals who make all three errors. Treating each error as a separate factor does not capture these differences. We also analyzed

An ANOVA identified a large effect of intrusion errors on responsibility judgments (*F*(1, 505) = 142.20, $p < .001$, $\eta^2_p = .22$, 90% *CI*[.17, .27]) and a medium effect of bypassing errors on responsibility judgments (*F*(1, 505) = 31.57, $p < .001$, $\eta^2_p = .06$, 90% *CI*[.03, .10]), qualified by an interaction between bypassing and intrusion errors (*F*(1, 505) = 27.86, $p < .001$, $\eta^2_p = .05$, 90% *CI*[.03, .09]) (see **Figure 3b**). There was no evidence for a main effect of fatalism errors ($p = .57$) or a two-way interaction between bypassing and fatalism errors ($p = .48$) or intrusion and fatalism errors ($p = .46$). We also found no evidence for any significant three-way interactions (all $p > .19$) and no evidence for a significant four-way interaction ($p = .86$).

To clarify the interaction between bypassing and intrusion, we computed tests of simple main effects. Participants who made no intrusion errors attributed significantly more responsibility when they made no bypassing errors ($M = 5.39$, 95% *CI*[4.98, 5.79]) compared to those who made bypassing errors ($M = 3.61$, 95% *CI*[3.28, 3.95]) ($t(548) = 6.67$, $p < .001$, $d = 1.22$, 95% *CI*[0.86, 1.59]). However, when participants made intrusion errors, there was no evidence for a difference in judgments of free will when they made no bypassing errors ($M = 5.64$, 95% *CI*[5.36, 5.92]) compared to when they made bypassing errors ($M = 5.38$, 95% *CI*[4.94, 5.83]) ($t(548) = 0.98$, $p = .33$, $d = 0.18$, 95% *CI*[-0.18, 0.54]).

Bypassing errors were associated with lower judgments of free will and moral responsibility compared to people who made neither bypassing nor intrusion errors *or* just intrusion errors. However, when people make both bypassing and intrusion errors, judgments of free will and moral responsibility are statistically indistinguishable from those who make only intrusion errors.

differences between judgments of free will and moral responsibility when each of the 8 distinct error types are treated as separate levels of the same factor (see Supplementary Materials §5.3).

3.3 Discussion

3.3.1 Error and judgment

Many participants seem to conflate determinism with different constructs (bypassing or fatalism) or mistakenly interpret the implications of deterministic constraints on agents (intrusion).

Measures of item invariance suggest that participants were not responding differently to error measures across different vignettes. Hence, responses to error measures cannot be explained exclusively in terms of differences in vignettes, but rather seem to reflect participants' mistaken judgments about determinism. Further, these errors are associated with significant differences in judgments about free will in predictable ways. Participants who conflate determinism with bypassing attribute less free will to individuals in deterministic scenarios, while participants who import intrusion into deterministic scenarios attribute greater free will. This makes sense. As participants perceive mental states to be less causally efficacious, free will is diminished. However, as people perceive more indeterminism, free will is amplified.

Additionally, we found that errors of intrusion are stronger than errors of bypassing or fatalism. Because bypassing errors are associated with diminished judgments of free will and intrusion errors are associated with amplified judgments, then, if all three errors were equal in strength, we would expect a linear relationship between different errors: individuals who make bypassing errors would have the lowest average judgments, individuals who make intrusion errors would have the highest average judgments, and people who make both errors would be in the middle (as both errors would cancel out). We did not observe this relationship. Instead, participants who make intrusion errors are statistically indistinguishable from each other, no matter what other errors they make. Thus, errors of intrusion seem to trump others in the process of forming judgments of free will.

The errors people make are not incidentally related to their judgments. Instead, there are significant associations between people's inferential errors about determinism and how they attribute free will and responsibility. This evidence supports our claim that people make several errors about the nature and implications of determinism.

3.3.2 Are these errors?

We assume that the bypassing, fatalism, and intrusion items measure errors about determinism. If this assumption is correct, then agreement with these items indicates a failure to draw central inferences about the implications of determinism. This assumption can be challenged in two ways. Perhaps determinism as it is used in the context of the Compatibility Question is such that it is compatible with the propositions expressed by the various error measures, or perhaps people understand determinism in a way that differs from the conventions of philosophers interested in the Compatibility Question. We address both challenges below.

Determinism is: "…the thesis that there is at any instant exactly one physically possible future" (1983: 3). Determinism is a thesis about conditional necessities (Audi, 1993). That is, future events are necessitated *conditional on* whichever laws of nature obtain and whatever facts about the past are true. The truth of determinism thus entails a single *physically* possible future: when certain conditions are fixed (i.e., the laws of nature and the past), then only one future is possible. This conditional necessity is essential for differentiating determinism and fatalism. Consider the fatalism items again:

Fatalism 1: Jeremy would have ended up robbing the bank no matter what he tried to do.
Fatalism 2: Jeremy will rob the bank no matter what.

The thesis of fatalism entails that only one future is *logically* possible, while determinism entails that only one future is *physically* possible. Thus, if the laws of nature or the past were different, then a different future would obtain (Lewis, 1981; Vihvelin, 2004). In other words, determinism does not preclude outcomes varying across the modal landscape. It is incorrect to agree that some outcome will occur, given determinism, *no matter what* because the necessity associated with determinism is conditional on the past and laws of nature being held fixed. Even if determinism is true, it does not rule out the *possibility* of alternative sequences of events governed by different sets of natural laws.

The thesis of determinism states that the laws of nature and facts about the past jointly entail facts about the future. However, facts about the past encompass facts about individual preferences and decision-making. What we want and decide forms part of the causal chain that stretches from the past into the future. Thus, even if determinism is true, mental states make a difference as to what happens (Nahmias, 2011; Sartorio, 2005). Therefore, determinism does not entail the first bypassing statement:

Bypassing 1: It doesn't make any sense to say that Jeremy made his own choice to rob the bank.

Determinism does not preclude the possibility of making choices that cause action. It simply rules out the possibility of such choices operating independently of deterministic causal chains. Likewise, determinism does not rule out the possibility of being considered a source of one's actions. If choices stem from wants and beliefs, then it seems possible for some actions to be up to us even if determinism is true (Markosian, 1999). Thus, determinism does not preclude the second bypassing statement:

Bypassing 2: It's not up to Jeremy whether or not to rob the bank.

Still, the Bypassing items might imply that Jeremy is not the *ultimate source* of his decisions. And it might be the case that determinism precludes any agent from being the ultimate source of their decisions even if their wants and beliefs are causally efficacious. However, in a previous study, these Bypassing items were highly correlated with a statement that mental states have no effect on decisions in deterministic universes (Nadelhoffer et al., In press; see also Supplementary Materials §3). Thus, participants seem to interpret the Bypassing items in terms of whether mental states *make a difference* to decisions rather than whether agents are the *ultimate* sources of their decisions.

Finally, consider the intrusion items:

Intrusion 1: There was at least a slight chance that Jeremy could have chosen not to rob the bank even if everything (including the laws of nature) had been exactly the same prior to his decision.

Intrusion 2: It was open for Jeremy to choose not to rob the bank at the exact moment he decided to rob it.

Both items fix the laws and the past as part of the context for individual choices. Because determinism implies necessity of the consequence, fixing the antecedent necessitates the conclusion. Thus, when the laws of nature and past are fixed within deterministic scenarios, the outcome is fixed. Both intrusion items, then, reflect errors about determinism.

Both Intrusion items admit of epistemic interpretations. Given everything Jeremy knows, it is uncertain whether he will rob the bank. Participants might use this epistemic interpretation when responding to the items, in which case agreement with either item might not constitute an error about determinism. However, a model for predicting Bypassing scores found that average Intrusion scores are a significant predictor of Bypassing scores even when controlling for average Fatalism scores ($\beta = -0.11$, $p < .001$). As participants agree *more* that agents do not really make their own choices, they tend to agree *less* that it was open for the agent to make a different choice.

This suggests that people adopt a metaphysical reading of the Intrusion items, because whether some agent is seen as *really* making a choice is associated with whether alternative choices are open to that agent. To reiterate: this is an error. Participant responses to the Intrusion items suggest that people think that the future is metaphysically open. But this marks a failure to appreciate how determinism entails a metaphysically closed future.

A further problem with applying an epistemic interpretation of Intrusion items is that 36% (202/556) of participants agreed with both Bypassing and Intrusion items. If we suppose that participants adopt an epistemic reading of the Intrusion items, then we should assume they adopt an epistemic reading of the Bypassing items. Consider a pair of such items:

- In Universe A, it was open for John to choose not to have French Fries at the exact moment he decided to have them. [Intrusion]
- In Universe A, it's not up to John whether or not to eat fries. [Bypassing]

An epistemic reading of the Intrusion item seems inconsistent with an epistemic reading of the Bypassing item. From John's perspective, it would seem open, given what John knows, to choose to eat French fries or not and thus up to him whether or not he chooses to eat fries. However, if we take a metaphysical reading of the Bypassing items, then we should adopt the same reading of the Intrusion items. But agreeing with a metaphysical interpretation of the Intrusion items is clearly an error.

The error measures used in our study reflect genuine inferential errors with respect to determinism. Participant responses to our error measures might invite a different challenge. Why think that the sense of determinism at issue in the Compatibility Question fixes reference? Participants might have a different understanding of determinism, and we could prioritize the meaning of 'determinism' that reflects common usage. However, there are two problems with this response. First, the concept of determinism underlying judgments observed in our studies seems incoherent. Many participants made some combination of errors that included intrusion (255/556, 46%), meaning that almost half of our sample judged that it was both possible for people to choose to do something different and that whatever happened would have happened no matter what. The folk concept of determinism might embed this inconsistency, but we think this is unlikely. Further, if we concede that people have a distinct concept of determinism from the one at issue in the Compatibility Question, then folk intuitions do not have any evidential import for these theoretical discussions. Put differently, if people believe that some distinct notion, *determinism*,* is compatible with having free will, this does not obviously bear on whether *determinism* is compatible with free will. At best, it would open a discussion as to why *determinism* (and not *determinism**) is at issue in discussions of the Compatibility Question.

Because many participant make inconsistent judgments, we think it is plausible that people are making errors about determinism rather than operating with a distinct (potentially incoherent) concept of determinism. Hence, we also think that our error measures in fact measure *errors*.

3.3.3 What explains the errors?

Errors can arise from various sources. Just because people make errors with respect to a particular concept does not mean they cannot draw inferences centrally related to it. Recall the argument meant to establish that people cannot make central inferences about determinism:

1) People agree with statements that are not true in deterministic universes.
2) Agreeing with these statements constitutes making an error about the nature and implications of determinism
3) Therefore, people make several errors about the nature and implications of determinism.
4) If people could draw central inferences about the concept of determinism, they would not make these errors.
5) Therefore, people cannot draw central inferences about determinism.

Different sources of error provide different evidence about one's facility with a concept. Some errors might be incidental or procedural, rather than conceptual.

Perhaps participants make errors because they do not understand the materials or respond in bad faith. However, we included standard comprehension and attention checks to screen participants prior to analyzing data. If participants did not attentively read the materials, we would expect more participants to fail our comprehension checks. Moreover, the pattern of responses is somewhat predictable. Judgments of intrusion are associated with greater attributions of free will and responsibility, while judgments of bypassing and fatalism are associated with diminished attributions. If participants were responding in bad faith, we would expect less predictable results. It is impossible to completely rule out bad faith responding, but the results suggest that such responding did not occur at rates that call into question their validity.

Some have suggested that judgments of free will and responsibility are driven by biases toward blame validation (Clark et al., 2019). Accordingly, insofar as determinism might threaten the justifiability of blame, participants might refuse to accept the deterministic aspects of the scenario when making judgments about them. This is unlikely to serve as a general explanation. Some of the vignettes had highly abstract content and included abstract items about the possibility of free will and responsibility generally. These abstract scenarios are unlikely to activate biases toward blame validation, as there is no concrete instance of wrongdoing that prompts a desire to blame and punish (Nichols and Knobe, 2007). Further, some of the vignettes described non-actual universes. It is unclear what would motivate a refusal to accept that determinism could be true in *any* universe, even if there were biases toward rejecting the actual world's being deterministic.

Finally, some errors might stem from bad materials. Clearer representations of determinism might elicit fewer errors. While this is possible, note that it comes at the expense of denigrating widely used stimuli. Within the larger context of thinking about how everyday intuitions bear on theorizing about the Compatibility Question, this response puts us back to square one: intuitions don’t tell us much because the materials are not structured appropriately to draw out meaningful judgments. Also, this response raises the question of what constitutes good materials. The vignettes we used presented determinism in several different ways. It is unclear how better to describe determinism in a way that does not resort to overly technical language.

If all or most errors were procedural or incidental, then the argument for the inability to make central inferences would be invalid. However, because the errors seem mainly to be conceptual, the argument seems valid. Moreover, we think the errors people make indicate a lack of competence with respect to the concept of determinism. In terms of the Tracking Problem, we think this constitutes good evidence for (TP-6).

4. Back to the tracking problem

The Tracking Problem results from a contradiction built on a series of seemingly plausible premises and implications of these premises:

TP-1. Participant judgments elicited by vignettes used in experimental philosophical research track the target content of the vignettes.
TP-2. The target content of vignettes used in experimental research on free will represents the philosophical notion of determinism (hereafter, determinism).[19]
TP-3. Participant judgments of free will elicited by deterministic vignettes track determinism. [1, 2]
TP-4. If participant judgments track determinism, then participants grasp the concept of determinism.
TP-5. If participants grasp the concept of determinism, then participants can draw central inferences about determinism.
TP-6. Participants cannot draw central inferences about determinism.
TP-7. Therefore, participant judgments of free will elicited by deterministic vignettes do not track determinism.[3, 4, 5, 6]

[19] Put differently, the term ‘determinism’ as it occurs in The Compatibility Question and the target content of the vignette denote the same proposition.

We have already made an empirically motivated case for (TP-6). Proponents of (TP-1) and (TP-2) will likely insist that better materials can elicit theoretically meaningful judgments. However, in the absence of better materials, what else could be said about the problem?

(TP-4) and (TP-5) make substantive claims about the relationship between having a concept and the ability to draw inferences about that concept. Some well-known arguments for externalism about mental content might suggest that (TP-5) is false. For example, Burge (1979) argued that people can grasp the concept of arthritis even when mistakenly inferring that someone has arthritis in their stomach. This is because individuals possess concepts in virtue of operating within linguistic communities that have a history of successfully referring to the phenomenon picked out by the concept, even if individuals occasionally fail to successfully refer. Some inferences might also be peripheral to a concept. For example, someone could grasp the concept of water without being able to infer that water boils at 212 degrees Fahrenheit at sea level.

Content externalist responses to the Tracking Problem are inadequate for two reasons. First, people do not seem to operate within a linguistic community that has a history of successfully referring to deterministic scenarios. Worries about determinism have always been relatively academic. In the medieval and early modern period, philosophers and theologians wrestled with the possibility of human freedom in the context of God's omniscience and providential control, or theological determinism (Murray, 1995). With the advent of Newtonian mechanics and the turn to quantitative physics, the relevant notion of determinism became decidedly more scientific. Instead of questioning whether God left room for freedom, scholars wondered whether there was any wiggle room in a fundamentally physical universe governed by iron-clad laws (Ismael, 2013). Thus, determinism is not a topic around which some portion of our everyday discourse might be organized.[20] Second, concepts and inferences seem to dissociate for experiential concepts. That is, people might acquire some concept through perceptual experience while also making certain errors with respect to the concept. This explains why concept possession for arthritis and water can tolerate some degree of ignorance because these concepts are acquired through perceptual acquaintance. Notably, determinism is not a perceptual feature of any situation: deterministic and indeterministic universes are perceptually indistinguishable (Horgan, 2015). Thus, it is unclear how people would acquire the concept of determinism independently of testimony. But, in learning about it, people would also acquire information that supports making certain kinds of inferences. If those inferences cannot be made, this is good evidence that people do not grasp the concept.

Someone might object that (TP-5) and (TP-6) equivocate on the relevant inferences. That is, the inferences at issue in (TP-5) are not *central* inferences. However, if the inferences measured in our experiment are not relevant to determining whether people grasp the concept of determinism, then what are the relevant inferences? The core features of determinism are that, when the thesis of determinism is true at some world, the future is necessary *conditional on* certain facts about the past and laws of nature obtaining in that world. How else can we assess whether people grasp the concept than by asking people about the modal strength of determinism (fatalism), the causal implications of determinism (bypassing), or the possibility of indeterminism (intrusion)? If nothing else, the argument presented above puts the onus on the proponent of this skeptical response to produce a distinctive list of central inferences.

[20] This might not seem to be the case for theological determinism. After all, theological determinism raises difficult questions about God's grace and justice, both of which are central to the experiences of religious believers. However, the difficulties themselves are also academic. For example, in the early 16th and 17th century, the Dominicans and Jesuits engaged in heated debates about the relationship between God's foreknowledge and the allocation of grace. Eventually, the Pope was asked to resolve the dispute between the two orders. However, the Pope determined that too little was known to make a definitive statement, and further debate on the topic was prohibited without special dispensations. Thus, the debate was not considered practically serious enough as a matter of faith for the Catholic Church to feel compelled to register an official statement on the problem of theological determinism (Murray, 2011: Introduction).

The last option is experimental nihilism (rejecting TP-1). That is, no matter how well-designed some vignette is, participants are not able to respond to the target content because they lack the core concepts implicated in such content. We want to briefly explore some of the implications of this option and what it might mean for future work on free will beliefs. For one, we are not suggesting general experimental nihilism for research into folk-psychological judgments related to philosophical concepts. As we mentioned above, the soundness of the Tracking Problem varies depending on the research questions.

Experimental nihilism might be relative to the Vignette-Judgment model. After all, the primary issue might be that vignettes cannot encode target content about determinism in a way that elicits theoretically meaningful judgments. Some researchers have proposed moving to perceptual or feature-based paradigms to study attributions of free will and responsibility (Sosa et al., 2021; Machery et al., In press). However, this does not address the Tracking Problem outlined here. Determinism is not a perceptual feature of a situation, so it must be stipulated for it to inform judgments in a perceptual paradigm. But the stipulation must provide some content to the notion of determinism, which raises the issues of how to make this stipulation and whether people can comprehend the implications of the stipulation. For perceptual paradigms to address the Compatibility Question, stimuli must incorporate deterministic elements in the same way proponents of the Vignette-Judgment model do. The same comprehension questions arise for perceptual paradigms as vignette paradigms.

Others have advocated using natural language paradigms to assess folk psychological categories of free will and responsibility (Monroe and Malle, 2010). The idea is that researchers should examine folk-psychological constructs through everyday conceptualizations rather than eliciting judgments through abstract or bizarre vignettes. We think that this is a promising methodology for studying folk-psychological constructs related to attributions of agency. However, Monroe and Malle (2010) found that people's lay conceptualizations of free will made no reference to metaphysically loaded constructs like determinism: "People's responses [to a prompt about the conditions for free will] provide…no assumptions of substance dualism, indeterminism or original causes" (p. 216). We think that these results are consistent with the interpretation of our results presented above. That is, if people lack facility with the concept of determinism, then we would expect that such concepts would not figure in folk conceptualizations of free will. Thus, Monroe and Malle find exactly what we would expect if people lack facility with the concept of determinism. However, if this interpretation is correct, then natural language paradigms will not furnish evidence that bears on the Compatibility Question. Thus, while such paradigms might be useful for studying other aspects of psychological attribution, they come at the expense of delivering philosophically edifying results.

In conclusion, the Tracking Problem is not restricted to the experimental study of free will beliefs. It is a general problem for work operating under the Vignette-Judgment model. This is because the model relies on participants to make judgments in response to vignettes that encode some philosophically relevant target content and to understand that target content in a way that aligns with researcher expectations. The Tracking Problem exploits two different ways of failing in these operations. Participants might respond to peripheral content or depart from researcher expectations. With respect to free will beliefs, we suspect that participants are responding to peripheral content because they do not have the concepts necessary for responding to the target content in a way that aligns with researcher expectations. The situation might be different for different concepts. For example, participants likely have a concept of 'knowledge', so there is unlikely a parallel case to be made for (TP-6) applied to experimental epistemology. Either way, the conceptual landscape seems too variegated to draw any general lessons from the Tracking Problem abstracted from a particular domain of study. We suggest that experimental researchers who presume the Vignette-Judgment model consider the Tracking Problem when developing materials to ensure vignettes elicit theoretically relevant judgments.

Acknowledgements

Thanks to Al Mele, Eddy Nahmias, Shaun Nichols, and Gunnar Björnsson for feedback on the error items. Michael J. Murray, Tim O'Connor, and John Churchill provided feedback in the early stages of research. This research was presented at the 2021 Society for Philosophy and Psychology meeting and to the Moral Attitudes and Decisions Lab at Duke University. Walter Sinnott-Armstrong and Jana Schaich Borg, among others, gave helpful criticisms. Finally, thanks to Joshua Knobe and two anonymous referees for the volume for insightful comments that helped to improve our paper. This research was partly supported by an internal opportunity award from the College of Charleston.

References

Audi, R. 1993. Modalities of knowledge and freedom. In *Action, Intention, and Reason* (Ithaca, NY: Cornell University Press), 253-80.

Burge, T. 1979. Individualism and the mental. *Midwest Studies in Philosophy* 4, 73-121.

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, 10,* Article 215. https://doi.org/10.3389/fpsyg.2019.00215.

de Vet, H.C.W, Mokkink, L.B., Mosmuller, D.G., Terwee, C.B. 2017. Spearman-Brown prophecy formula and Cronbach's alpha: different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology* 85: 45-49. doi: 10.1016/j.jclinepi.2017.01.013.

Dennett, D.C. 2006. Higher-order truths about chmess. *Topoi* 25:1-2, 39-41.

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

Frankfurt, H.G. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* 68:1, 5-20.

Horgan, T. 2015. Injecting the phenomenology of agency into the free will debate. In D. Shoemaker (ed.) *Oxford Studies in Agency and Responsibility* 3. Oxford: Oxford University Press, 34-61.

Hu, L. and Bentler, P.M. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal* 6:1, 1-55.

Ismael, J.T. 2013. Causation, free will, and naturalism. In Ross, D., Ladyman, J., and Kincaid, H. (eds.) *Scientific Metaphysics* (Oxford: Oxford University Press), 208-35.

Kane, R. 1996. *The significance of free will*. Oxford: Oxford University Press.

Knobe, J. 2019. Philosophical intuitions are surprisingly robust across demographic differences. *Epistemology and Philosophy of Science* 56:2, 29-36.

Knobe, J. and Doris, J. 2010. Responsibility. In J. Doris and The Moral Psychology Research Group. *The Moral Psychology Handbook*. Oxford: Oxford University Press.

Lewis, D. 1981. Are we free to break the laws? *Theoria* 47:3, 113-21.

Machery, E. 2017. *Philosophy within its proper bounds*. Oxford: Oxford University Press.

Machery, E., Kneer, M., Willemsen, P., and Newen, A. In press. Beyond the courtroom: Agency and the perception of free will. In P. Henne and S. Murray (eds.) *Advances in Experimental Philosophy of Action*. London: Bloomsbury.

Markosian, N. 1999. A compatibilist version of the theory of agent causation. *Pacific Philosophical Quarterly* 80:3, 257-77.

May, J. 2014. On the very concept of free will. *Synthese* 191:12, 2849-2866.

Mele, A.R. 2006. *Free will and luck*. Oxford: Oxford University Press.

Meredith, W. 1993. Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58, 525-43.

Monroe, A. E., & Malle, B. F. (2010). From uncaused will to conscious choice: The need to study, not speculate about people's folk concept of free will. *Review of Philosophy and Psychology, 1*(2), 211–224. https://doi.org/10.1007/s13164-009-0010-7.

Murray, D., & Nahmias, E. (2014). Explaining away incompatibilist intuitions. *Philosophy and Phenomenological Research*, *88*, 434–467. https://doi.org/10.1111/j.1933-1592.2012.00609.x

Murray, M.J. 1995. Leibniz on divine foreknowledge of future contingents and human freedom. *Philosophy and Phenomenological Research* 55:1, 75-108.

Murray, M.J. 2011. *Dissertation on Predestination and Grace*. New Haven, CT: Yale University Press.

Nadelhoffer, T., Rose, D., Buckwalter, W., and Nichols, S. 2020 Natural compatibilism, indeterminism, and intrusive metaphysics. *Cognitive Science* 44:8, e12873.

Nadelhoffer, T., Yin, S., and Graves, R. 2020. Folk intuitions and the conditional ability to do otherwise. *Philosophical Psychology* 33:7, 968-96.

Nadelhoffer, T., Murray, S., and Dykhuis, E. In Press. Intuitions about free will and the failure to comprehend determinism. *Erkenntnis*.

Nahmias, E. (2011). Intuitions about free will, determinism, and bypassing. In R. Kane (Ed.), *The Oxford Handbook on Free Will* (2nd edition: pp. 555–576). New York: Oxford University Press.

Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. 2005. Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology* 18:5, 561-84.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2006). Is incompatibilism intuitive? *Philosophy and Phenomenological Research*, *73*, 28–53. https://doi.org/10.1111/j.1933-1592.2006.tb00603.x

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.

Nichols, S. 2011. Experimental philosophy and the problem of free will. *Science* 331:6023, 1401-1403.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The Cognitive science of folk intuition. *Noûs*, *41*, 663–685. https://doi.org/10.1111/j.1468-0068.2007.00666.x

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

R Core Team. 2008-2020. *R*: *A language and environment for statistical computing*. R Foundation for Statistical Computing. https:www.R-project.org/

Rose, D., Buckwalter, W., & Nichols. (2017). Neuroscientific prediction and the intrusion of intuitive metaphysics. *Cognitive Science*, *41*, 482–502. https://doi.org/10.1111/cogs.12310

Roskies, A., & Nichols, S. (2008). Bringing moral responsibility down to Earth. *The Journal of Philosophy*, *105*, 371–388. https://doi.org/10.5840/jphil2008105737

Rosseel, Y. 2012. lavaan: An R package for structural equation modeling. Journal of Statistical Software 48:2, 1-36.

Sartorio, C. 2005. Causes as difference makers. *Philosophical Studies* 123, 71-96.

Sartorio, C. 2020. *Causation and free will*. Oxford: Oxford University Press.

Sosa, F.A., Ullman, T., Tenenbaum, J.B., Gershman, S.J., and Gerstenberg, T. 2021. Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition* 217, 104890.

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

Taylor, R. 1966. *Action and purpose*. Englewood Cliffs, NJ: Prentice-Hall.

van Inwagen, P. 1978. Ability and responsibility. *Philosophical Review* 87:2, 201-224.

van Inwagen, P. 1983. *An Essay on Free Will*. Oxford: Oxford University Press.

van Inwagen, P. 2004. Freedom to break the laws. *Midwest Studies in Philosophy* 28:1, 334-50.

Vandenberg, R.J. and Lance, C.E. 2000. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods* 3:1, 4-70.

Vargas, M. 2013. *Building Better Beings*. Oxford: Oxford University Press.

Vargas, M. 2017. Contested terms and philosophical debates. *Philosophical Studies* 174:10, 2499-2510.

Vihvelin, K. 2004. Free will demystified: A dispositional account. *Philosophical Topics* 32, 427-450.