Folk Intuitions about Free Will and Moral Responsibility: Evaluating the Combined Effects of Misunderstandings about Determinism and Motivated Cognition

Kiichi Inarimori (Hokkaido University)
Yusuke Haruki (The University of Tokyo)
Kengo Miyazono (Hokkaido University)

Abstract

In this study, we conducted large-scale experiments with novel descriptions of determinism. Our goal was to investigate the effects of desires for punishment and comprehension errors on people's intuitions about free will and moral responsibility in deterministic scenarios. Previous research has acknowledged the influence of these factors, but their total effect has not been revealed. Using a large-scale survey of Japanese participants, we found that the failure to understand causal determination (intrusion) has limited effects relative to other factors and that the conflation of determinism and epiphenomenalism (bypassing) has a significant influence, even when controlling for other variables. This leads to the increased prevalence of incompatibilist responses. Furthermore, our results demonstrated a close association between the attribution of free will/responsibility and retributive desire. While further research is needed to establish the causal relationship between these factors, this association is consistent with Cory Clark and colleagues' (2019) study that increased desire contributes to increased compatibilist responses and their claim that a definitive intuition about free will may be elusive.

1. Introduction

Philosophy of free will scholars often discuss the compatibility or incompatibility of free will and determinism. According to compatibilism, the free will necessary for moral responsibility is compatible with determinism. According to incompatibilism, it is not. Intuition has played a central role in this debate. As experimental philosophy has developed, folk intuitions about free will have attracted much interest. Numerous experimental studies have been carried out (for a comprehensive review, see Inarimori et al., 2024). Some studies (e.g., Nahmias et al. 2005, 2006) support natural compatibilism, according to which folk intuitions are compatibilist. Other studies (e.g., Nichols & Knobe 2007) support natural incompatibilism, according to which folk intuitions are incompatibilist.

Why do different studies support different views? Or, what drives compatibilist versus incompatibilist responses? A promising factor relates to people's desire to uphold responsibility for human agents. Studies by Cory Clark and colleagues (2019) suggest that

> people do not have one intuition about whether free will is compatible with determinism. Instead, people report that free will is compatible with determinism when desiring to uphold moral responsibility (Clark et al. 2019, p1).

Clark et al. (2014) found that people's beliefs about free will are motivated by desires for punishment. The authors conducted a study examining the effect of people's desire to punish an agent on their free will beliefs. They measured participants' desires to punish the agent in the given vignettes between two conditions: immoral and morally neutral. Participants in the immoral conditions were presented with the description of a home robbery by an unemployed man, while participants in the morally neutral conditions were presented with the description of an unemployed man stealing an aluminum can from the recycling bin. They found that the presentation of an immoral action increased people's desire to punish, and that increased desire was positively correlated with attributions of moral responsibility and beliefs in free will. This study did not examine free will beliefs in deterministic scenarios. However, a later series of studies by Clark et al. (2019) suggest that desires for punishment also influence people's responses in deterministic situations. For example, they found that when participants were presented with morally relevant actions, their agreement with the compatibilist view increased. When presented with morally neutral actions, they did not. This suggests that the desire to maintain moral responsibility (i.e., motivated cognition) influences responses to compatibility questions.

Clark et al. (2019) argue that people's motivation to attribute free will might explain divergent results in previous studies. A famous study by Shaun Nichols and Joshua Knobe (2007), for example, found that people judge moral responsibility to be incompatible with determinism when asked specific abstract questions (e.g., "Is it possible in universe A [a deterministic universe] for a person to be fully morally responsible for their actions?"). However, Nichols and Knobe also found that people judge moral responsibility to be compatible with determinism when given a concrete description of wrongdoing (e.g., an actual murder). According to Clark et al. (2019, p18), the difference in intuitions between abstract and concrete cases can be explained by a difference in motivation, one that relates to attributing punishment to an agent in the two different cases. The concrete case leads to an increased motivation to punish, which results in an increase in compatibilist responses.

Another promising explanation for the divergent responses relates to a possible misunderstanding of determinism. In the experimental philosophy of free will, participants are often presented with descriptions of deterministic universes to elicit their intuitions about determinism. Here is an example 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. (Nichols & Knobe, 2007, p. 669)

In many cases, after reading the description of a deterministic universe as described above, participants report their intuitions about the compatibility of free will and moral responsibility with the universe; however, in this type of study, two types of miscomprehensions are known to influence participants' responses: *bypassing* and *intrusion*.

Bypassing

Determinism and epiphenomenalism are distinct concepts. It is important to note that determinism does not necessarily entail epiphenomenalism. In an epiphenomenalistic world, an agent's mental states, such as their desires, beliefs, and intentions, are bypassed. Her mental states have no causal effect on her actions. But determinism need not entail bypassing. Even if a person is determined to φ, her φ-ing can still be caused by her mental state in a deterministic way. Put otherwise, the direct cause of an agent's actions can be her mental state even if factors beyond her control ultimately determine her actions. Eddy Nahmias and Dylan Murray (2010) found that most participants exhibiting incompatibilist intuitions also agree with bypass statements.

Intrusion

There can also be an erroneous intrusion of deterministic assumptions (e.g., alternative possibilities) into one's understanding of determinism. Research by Thomas Nadelhoffer and colleagues (2020) suggests that most people who show compatibilist responses fail to understand that determinism excludes alternative possibilities.

A more recent study by Nadelhoffer and colleagues (2023) suggests that almost all participants in previous experiments might have been guilty of bypassing and intrusion. Nadelhoffer et al. examined the comprehension of determinism among participants who had passed a standard comprehension check about the meaning of determinism. Drawing from Nichols and Knobe (2007), Nadelhoffer et al. found that 67% of people agree with one or more intrusion statements. As many as 98% of people agree with one or more bypassing statements. They also found a negative correlation between bypassing and compatibilist responses and a positive correlation between intrusion and compatibilist responses.

It could be argued that this criterion is unduly stringent. Rather than focusing on the number of individuals classified as "complete passers," who disagree with any bypassing or intrusion statements, it would be more appropriate to focus on the number of individuals classified as "midpoint passers," whose average agreement with bypassing or intrusion statements is below the midpoint. However, Nadelhoffer et al. (2023) contend that the majority of participants fail to meet the criteria for either the bypassing or intrusion checks, even when the focus is on the number of midpoint passers and the average agreement to comprehension materials (footnote 31). Even if one focuses on average agreement with both bypass and intrusion statements, on average, many people still agree with these statements (bypassing 91% and intrusion 57%). Similarly, as many as 80% of the subjects failed to pass the comprehension checks about determinism in their other experiment using a scenario similar to the one in Nahmias et al. (2005, 2006).[1]

[1] This rate includes both (a) the conflation of determinism with bypassing or the failure to understand the exclusion of alternative possibilities and (b) the conflation of determinism with fatalism (the idea that it is modally impossible for one to act otherwise than one actually does). We did not focus on this because, as Nadelhoffer et al. (2023) show, the conflation of determinism and fatalism itself has no significant effect on people's responses.

What is important here is that, regardless of the criterion employed, the overall effect of the above-discussed factors on people's responses is uncertain. The focus of our research is to determine the total effect of these factors on folk intuitions. The significance of desires for punishment relative to comprehension errors has not been investigated. Although Clark et al. (2019) have demonstrated that motivated cognition might explain some proportion of compatibilist intuitions, they did not examine its effect size relative to the effect size of comprehension errors. Moreover, the overall effect of comprehension errors on folk intuitions themselves has not been sufficiently examined. Although Nadelhoffer et al. have conducted extensive research on determinism's comprehension issue, they provided only one type of comprehension check to each participant (based on the assigned condition [bypass, intrusion, or fatalism])[2]. As such, the full impact of comprehension errors on the folk responses or intuitions of people who have passed both bypass and intrusion materials has not been uncovered.

2. Current Research

Our goal in this study is to uncover the total effects of desires for punishment and comprehension errors on people's intuitions about free will and moral responsibility in deterministic situations.

To measure the effects of the desires and compare them with the effects of comprehension errors, we developed new statements for measuring participants' desire to punish an agent in a given scenario. We gave both of these statements and the relevant comprehension materials to participants. To uncover the full effect of comprehension errors, we also needed to collect data from both those who misunderstand determinism and those who understand determinism to some degree. Given the rather low comprehension rates found in Nadelhoffer et al. (2023), we speculated that most participants would not understand determinism. We therefore conducted a large-scale experiment with a sample size of 1,200 participants, collecting data from those who passed comprehension checks. We also gave each participant the bypass and intrusion statements (participants in Nadelhoffer et al. [2023] were given either bypass or intrusion material depending on the conditions).

Our research is also an attempt to improve comprehension rates by developing both a new scenario and videos describing that scenario[3]. We created a new version of the so-called "rollback case" scenario. This is based on Nahmias et al.'s (2010, 2014) description of a universe that is re-created over and over again (with the same things happening each time). We chose the rollback case scenario because previous studies have suggested that people who have been exposed to it are less likely to agree with both bypass (Nahmias & Murray, 2010) and intrusion (Nadelhoffer et al., 2020).[4] Based on this new version of the rollback case, we created videos describing the rollback universe with the expectation that a video explanation would be easier for people to understand than a written description and, in turn, improve comprehension rates. To demonstrate the effectiveness of the videos, we compared comprehension rates

[2] Some studies following Nadelhoffer et al. (2023), such as Cova and Martinez (2024) and Murray et al. (forthcoming), have investigated the total effects of comprehension errors. However, there is still controversy surrounding the total effects of comprehension errors and their seriousness, as well as uncertainty regarding the total effects compared to retributive desire.

[3] We created both videos describing determinism (Universe A) and indeterminism (Universe B). You can access these videos from the following link.
Universe A: https://youtu.be/ZpUMa_3er5s
Universe B: https://youtu.be/L-WoveH25WM

[4] Nadelhoffer et al. (2023), in contrast, used the supercomputer scenario (Nahmias et al., 2005, 2006) and the Nichols and Knobe (2007) scenario.

between three conditions: the Nichols & Knobe (2007) scenarios, the new rollback, and the rollback videos.

In sum, we measured participants' intuitions, understandings of determinism, and desires for punishment in three different scenarios. We then analysed the overall effect of misunderstanding and the desires on folk intuitions.

3. Experiment

3.1 Participants

We recruited 1,219 participants from Lancers.[5] They responded to a question about their intuitions in a deterministic universe, followed by questions about the comprehension of determinism and their desire to punish an agent in each scenario. This experiment was conducted on the Qualtrics platform (https://www.qualtrics.com/). All materials were originally given in Japanese and completed on participants' own digital devices. Participants were randomly assigned to one of our three conditions: (a) Nichols and Knobe's original condition (N&K), (2) the rollback condition (Rollback), or (c) the rollback video condition (Video). On average, it took participants 8.72 minutes to complete the study. They received ¥88 for participating (¥605/hour).[6]

A total of 220 participants were excluded from our statistical analyses. Three participants did not complete the survey, 46 did not pass "surface" attentional checks, 148 did not pass a "deep" attentional check (described below), and 26 claimed that Universe A is more like ours while also claiming that Universe A is impossible (an incoherent pair of judgments). Our final sample thus comprised 999 adults. The mean age was 42.09±10.34 years (548 men, 449 women, and 2 selected others). The number of participants in each condition was as follows: N&K = 332, Rollback = 335, and Video = 332. This represented the required sample size (n = 323 for each) to detect a small effect (f =.10) using one-way analysis of variance (ANOVA) under standard parameters (α =.05 and 1 – β =.80). This effect was calculated using G*Power software (version 3.1.9.7). Our study was carried out in accordance with the Declaration of Helsinki and its amendments. The Ethics Committee of Hokkaido University approved our experimental protocol.

3.2 Materials and Methods

Depending on the condition to which they were assigned, participants were presented with one of three scenarios describing a deterministic world. In the N&K condition, participants were given a Japanese-translated description of determinism from Nichols and Knobe (2007). In the other two conditions, participants were presented with our new version of rollback in Japanese text or video with Japanese narration. The English translation of our original scenarios is as follows:

> *Imagine a universe (Universe A) that is re-created over and over again, starting from the same initial conditions with the same laws of nature. In this universe, the same conditions and the same laws of nature produce the exact same outcomes. Every time the universe is re-created,*

[5] Lancers is a major Japanese online cloud-sourcing service.

[6] We set the reward at ¥88 based on two factors:

1. The average minimum wage (¥964/hour) in Japan at the time of the experiments (February 2023).
2. The expected completion time (5 minutes), which we calculated based on average completion times in a pilot study.

> *everything must happen in the same way. Therefore, every time the universe is re-created humans will make all the same decisions. For example, in this universe, a man named Taro decides to eat French fries at some specific time. Every time the universe is re-created, Taro then decides to eat French fries at that time.*
>
> *In contrast, imagine a universe (Universe B) that is re-created over and over again, starting from the same initial conditions with the same laws of nature. In this universe, the same conditions and the same laws of nature produce the exact same outcomes. The one exception is human decision-making. Whenever the universe is re-created, humans can make different decisions. For example, when this universe is re-created for the umpteenth time, a man named Taro decides to eat French fries at some specific time. But, every time the universe is re-created, Taro can decide to eat chocolate, eat pudding, drink beer, or the like instead of eating French fries.*

As we explained in Section 2, we rewrote the original version of the rollback scenario and created the new version and videos to make it more understandable. An important modification from the original version (Nahmias and Murray 2010) is that the new version contrasts two universes (deterministic and indeterministic). We introduced this contrast to emphasize determinism's implications. Another difference is that the new version contains only morally neutral actions, while Nahmias et al.'s original version describes a man named Jeremy's robbing a bank.

Participants in N&K, Rollback read Japanese texts. In Video, the scenario was presented as a narration and video with subtitles. The video was designed to enhance the understanding of a deterministic world by showing concrete examples, which were expected to reduce the comprehension error rate. After being presented with descriptions of determinism, we recorded participants' attitudes. We did so in three ways (following Nadelhoffer et al. 2023):

1. We asked which of the two universes is the most similar to our own ("Universe A" or "Universe B").
2. We asked whether participants thought that Universe A is possible ("Yes" or "No").
3. Participants completed a "surface" comprehension check about the deterministic world: "In Universe A, everything that happens is completely caused by what happened before it" ("True" or "False").

Next, we assessed each participant's intuition by asking them to evaluate a fictional murder scenario in Universe A. The scenario and relevant questions were as follows (originally presented in Japanese):

> *In Universe A, a man named Takashi becomes attracted to a woman other than his wife. He comes to believe that the only way to be with the woman is to kill his wife. Takashi knows that, if he puts poison in the tea his wife routinely drinks, then she will certainly drink it. Before leaving for work one day, he mixes cyanide into his wife's tea, thereby murdering her.*

- *Takashi killed his wife of his own free will.*
- *Takashi is morally blameworthy for killing his wife.*
- *Takashi killing his wife was a bad thing.*

Participants rated the extent of their agreement according to three criteria:

1. Free will attribution to the murderer (FW).
2. Moral responsibility attribution to the murderer (MR).
3. Wrongness of the behavior (Wrongness).

The responses were recorded using a 7-point Likert scale, ranging from 1 (do not agree at all) to 7 (strongly agree).

Immediately after the responses, participants were presented with another set of questions. Each question was presented as follows:

Bypassing

- *In Universe A, what Takashi wanted and believed had no effect on whether he killed his wife.*
- *In Universe A, Takashi would have decided to kill his wife no matter what he wanted or believed.*
- *In Universe A, it does not make any sense to say that Takashi made his own choice to kill his wife.*
- *In Universe A, it is not up to Takashi whether or not to kill his wife.*

Intrusion

- *In Universe A, Takashi can avoid doing what he does.*
- *In Universe A, Takashi has the ability to change his mind about killing his wife.*
- *In Universe A, there was at least a slight chance that Takashi would not kill his wife.*
- *In Universe A, it was possible for Takashi to do something other than kill his wife.*

Desire

- *I want Takashi to be morally blamed.*
- *I want Takashi to be punished.*

Surface attentional check

- *This question is designed to ensure that you have answered the question carefully. Be sure to answer 5 to this question.*

Deep attentional check

- *In Universe A, Takashi did not murder his wife with cyanide.*

The first eight of these questions were intended to assess comprehension errors about the deterministic worlds (the possibility of *bypassing* and *intrusion*). The bypass statements used in our experiment are based on statements originally provided by Nadelhoffer et al. (2023) in their second study. The intrusion statements used in our experiment are based on statements originally provided in the rollback cases by Nadelhoffer et al. (2020). This allowed us to simultaneously assess the dominant misunderstanding of the deterministic world (bypass and intrusion) for each participant.

We added two questions to measure each participant's *desire* to punish the murderer (i.e., motivated cognition). We also included two statements as attentional checks. One was a *surface attentional check* (as in previous studies). The other was a *deep attentional check* intended to include only reliable data in our analyses. This was due to a recent issue regarding the data quality of online experiments. For example, recent study by Florian Cova and Tristan Martinez (2024) reported that many participants

who passed surface attention checks failed deep attention checks. Importantly, we designed our deep attention check so that correct responses were on the *left* side of the 7-point scale. As Cova and Martinez also indicated, respondents who select a random answer are more likely to choose the higher end (i.e. the right side) of the scale.

Participants could refer to the description of the deterministic world to answer each question. The order of questions was randomized, and participants used a Likert scale, ranging from 1 (do not agree at all) to 7 (strongly agree).

3.3 Analysis

Our analysis commenced with the exclusion of unreliable data based on attentional checks. Initially, a surface-level check required participants to select a specific number. Those who chose different numbers were excluded. We then employed a more rigorous, deep attentional check. This involved a statement that was deliberately inconsistent with the previously presented scenarios. Participants who agreed with that statement (those indicating a response above the midpoint) were also excluded. We judged responses containing inconsistencies (such as that Universe A resembles our world but its existence is impossible) to be invalid and excluded them from the analysis.

For the initial step in our main analysis, we compared comprehension errors across conditions. We hypothesized that participants in Rollback and Video would exhibit fewer misunderstandings than those in N&K. To categorize comprehension errors, we defined participants demonstrating a relatively accurate understanding of the deterministic world concept as "passers". "Complete passers" were those who rated all items on the bypass and intrusion scale below the midpoint (i.e., lower than four). "Midpoint passers" were those whose average response across the eight items was below the midpoint, but who rated at least one item above the midpoint. We employed chi-squared tests to compare the proportion of complete passers and midpoint passers across conditions. We also conducted one-way ANOVAs with the condition as a between-subject factor. This was to independently assess potential differences in raw responses to bypass and intrusion items among the conditions.

In the next step, we examined the moral intuitions of both complete and midpoint passers to directly test whether misunderstandings about deterministic worlds would influence moral intuitions. To this end, we independently performed two-way ANOVAs on FW and MR, with the conditions (N&K, Rollback, and Video) and comprehension (midpoint passers versus non-passers) as between-subject factors. Additionally, to determine whether moral intuitions align with compatibilism even among individuals with bypass judgments, we utilized independent samples t-tests to compare FW and MR scores between participants who scored below and above the midpoint on the bypass questions.

In concluding our analysis, we employed multiple linear regression to explore how (in)compatibilist responses are shaped. In so doing, we simultaneously considered individual differences in comprehension errors and motivated cognition. We also created two independent models for FW and MR. In each model, the groups (N&K, Rollback, Video) were treated as a factor variable. We included bypass, intrusion, and desire scores as independent variables in the regression model to understand their collective impact.

3.4 Results

Before proceeding to the statistical analyses, we identified 220 out of 1219 responses as unreliable.

These were subsequently excluded from our statistical analyses. Notably, 148 of these participants failed the deep attentional check (nearly three times more than the number who failed the surface attentional check). This result underscores the importance of incorporating deep attentional checks in online surveys.

In our initial analysis, we focused on assessing the comprehension rate of the deterministic world concept among Japanese participants. As a preliminary step, we evaluated the internal consistency of the bypass and intrusion items as these were translated into Japanese and tested for the first time. The calculated Cronbach's alpha values were 0.65 for bypass and 0.84 for intrusion (similar to those found in Nadelhoffer et al.'s 2023 study). This suggests adequate internal consistency for measuring comprehension errors in the Japanese translation.

Next, we examined whether different descriptions of the deterministic world influenced the participants' agreements to both bypass and intrusion statements. The proportions of complete passers (N&K = 1.20%; Rollback = 1.49%; Video = 0.60%) and midpoint passers (N&K = 19.88%; Rollback = 22.39%; Video = 25.00%) did not show significant differences across the conditions (χ^2s < 2.50, *p*s > .29) (Figure 1A). Unfortunately, the overall proportion of passers was low (22.42% for midpoint passers and only 1.10% for complete passers when averaged across conditions). This finding aligns with Nadelhoffer et al.'s (2023) research and highlights a general difficulty in comprehending the relevant concepts.

When assessing the mean scores for comprehension errors across conditions using ANOVA, we observed a significant effect of the assigned groups on the bypass score ($F_{2, 996}$ = 7.06, η^2_p = .014, p < .001). Post hoc analysis using Bonferroni's multiple comparisons correction revealed that scores were higher in N&K (mean score = 4.69 ± 1.52) than in Rollback (4.41 ± 1.40) and Video (4.28 ± 1.40) (*t*s = 2.51, *p*s < .036) (Figure 1B). Conversely, we did not observe the group difference in Intrusion (mean score: N&K = 2.55 ± 1.52; Rollback = 2.37 ± 1.48; Video = 2.40 ± 1.50) ($F_{2, 996}$ = 1.47, η^2_p = .003, p = .23) (Figure 1C). In sum, although using the new version of the rollback scenario and video reduced bypass scores on average, the overall proportion of participants who correctly understood deterministic situations did not improve based on the presentation method. This suggests that, although certain approaches might influence specific aspects of comprehending deterministic universes, they do not necessarily enhance an overall understanding of that complex concepts.

Comparisons of compatibilist responses between participants having bypass judgments and those who do not revealed significant differences for FW (3.34 ± 2.26 vs. 4.48 ± 2.19, t_{997} = -9.93, p < .001) and MR (5.33 ± 2.03 vs. 6.16 ± 1.51, t_{997} = -6.76, p < .001), suggesting that having a bypass judgment contributes to incompatibilist responses (Figure 2). Still, compatibilist responses (average scores above the midpoint) were observed for MR even in individuals with bypass judgments, thereby providing support for the idea of a 'folk compatibilist theory' or natural compatibilism, which is the view that most laypeople have compatibilist intuitions[7].

[7] Perhaps these responses do not support compatibilist theories. Mainstream compatibilist requirements for free will involve a causal integration between an agent's mental state and their action, including responsiveness to reasons for performing or not performing an action (e.g., Fischer & Ravizza, 1998). Thus, most compatibilist theories might not approve of attribution of moral responsibility in bypass situations. Instead, the apparent compatibilist responses confirmed among these participants might be Free-Will-No-Matter-What (FWNMW) intuitions. These intuitions admit the attribution of free will and moral responsibility regardless of the causal efficacy of agents' mental states. Adam Feltz and colleagues propose the FWNMW hypothesis and argue that most compatibilist responses do not genuinely support compatibilist theories (Feltz, Cokely, and Nadelhoffer 2009; Feltz & Millan 2013). However, the empirical plausibility of this hypothesis is controversial (see Andow

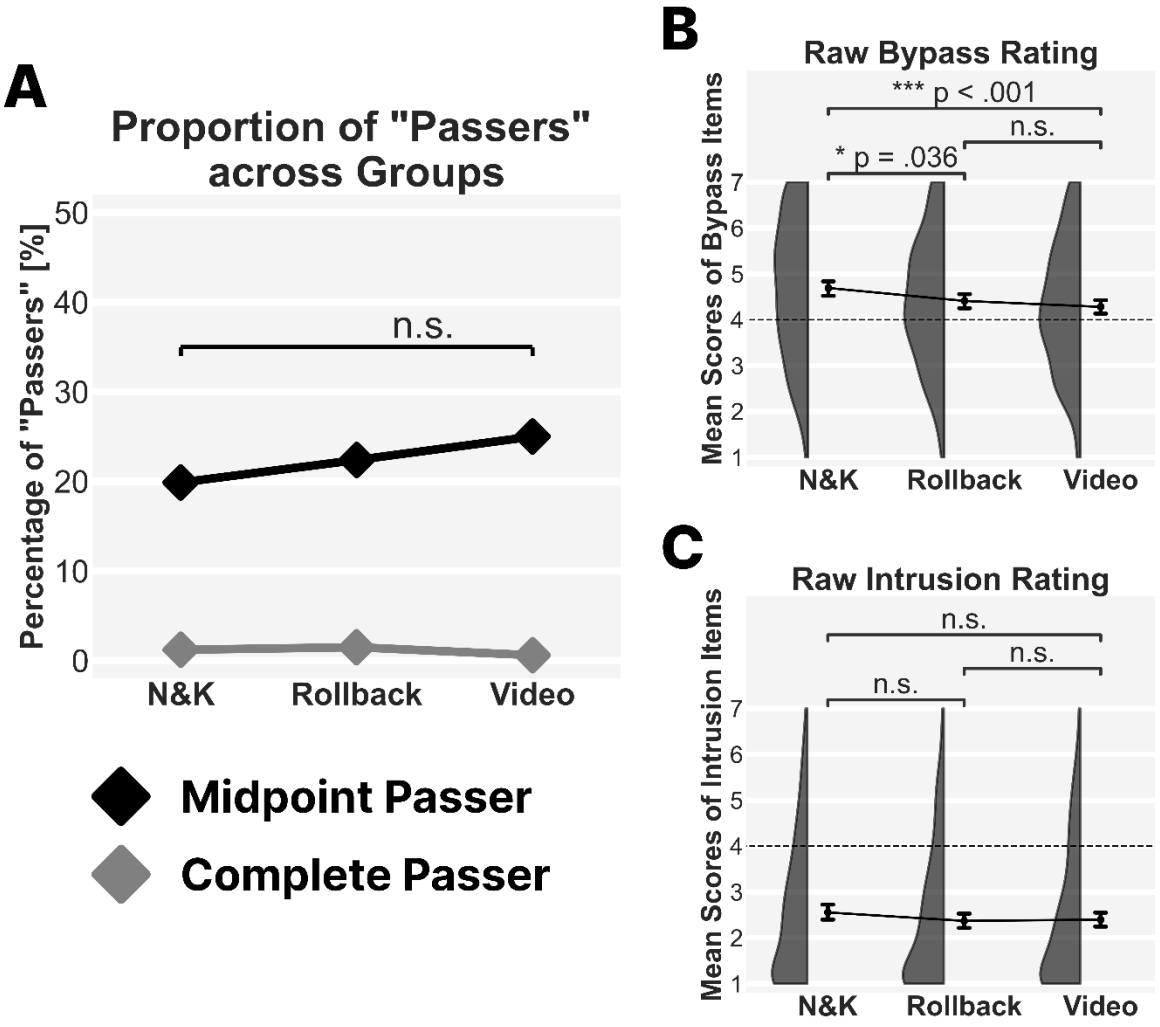

Figure 1. Comprehension errors across conditions (or assigned groups). A). The proportion of midpoint and complete passers did not differ between conditions (χ^2 = 2.50, p = .29; χ^2 = 1.26, p = .53). Participants who scored below midpoint among all questions regarding bypass and intrusion were labeled as complete passers, while midpoint passers were classified as participants whose averaged scores in bypass and intrusion questions were both below midpoint. B), C). People in N&K condition scored higher in bypass questions, compared with Rollback and Video conditions (ps < .037). Conversely, there was no significant difference in scores of intrusion questions across conditions. Group mean scores are depicted as a point estimate with 95% confidence intervals, with the data distribution shown as a half-violin plot for each condition. Dashed lines represent a midpoint of each item. N&K: Nichols & Knobe (2007)'s original condition; n.s.: no significance.

and Cova 2016). While our focus in this paper is on the total effects of comprehension mistakes, and it is beyond our scope to test the FWNMW hypothesis, we will come back to this point in Section 5.

Figure 2. Moral intuitions of people with bypass judgments. Participants were divided based on their average bypass scores (below/above midpoint). Independent samples t-tests indicated that individuals without bypass errors demonstrated more compatibilist responses regarding a murderer in a deterministic world, in terms of both free will and moral responsibility attribution (*p*s < .001). Still, even participants with bypass judgments displayed compatibilist tendencies on average, as they scored above the midpoint in moral responsibility attribution (average score = 5.33 ± 2.03). Mean scores are presented as point estimates with 95% confidence intervals, and the kernel density indicates the data distribution. The dashed line represents the midpoint.

In our subsequent analysis, we explored the moral intuitions of mid-point passers by using two-way ANOVAs to assess the effects of comprehension (midpoint passers versus non-passers), the conditions (N&K, Rollback, and Video), and the interaction of comprehension and the conditions. The results revealed distinct patterns for attribution judgment of FW and MR. For FW, the effects of the conditions ($F_{2, 993}$ = 17.58, η^2_p = .033, p < .001) and comprehension ($F_{1, 993}$ = 33.77, η^2_p = .032, p < .001) were both significant while the interaction between them was not ($F_{2, 993}$ = 0.79, η^2_p = .001, p = .45). Post hoc analysis indicated that participants in Rollback and Video condition showed more compatibilist intuitions than those in N&K condition (ps < .001). Also, midpoint passers had more compatibilist responses than non-passers (p < .001). Conversely, the ANOVA for MR revealed a significant interaction between conditions and comprehension ($F_{2, 993}$ = 3.73, η^2_p =.007, p =.02) in addition to the main effects of both factors (condition: $F_{2, 993}$ = 13.88, η^2_p = .027, p < .001; comprehension: $F_{1, 993}$ = 12.69, η^2_p = .012, p < .001). Importantly, post hoc analysis revealed that incompatibilist responses were more prevalent among N&K participants who had comprehension errors than any other combinations (Bonferroni corrected ps <.001; see Figure 3).

In sum, when evaluating free will attribution in deterministic scenarios, better comprehension of presented scenarios and the introduction of rollback scenarios were associated with more compatibilist responses. Notably, for moral responsibility attribution, the effect of better comprehension was observed only in the context of the N&K original scenario. Participants in the Rollback and Video conditions exhibited relatively compatibilist intuitions regarding moral responsibility, irrespective of their comprehension level of deterministic situations. This implies that, some important differences between N & K and Rollbacks, which were not measured in this experiment, may have influenced people's responses to MR questions. We shall return to this point in the Discussion section.

Finally, our regression analyses revealed that assigned group, individual comprehension errors, and motivated cognition independently contributed to participant's moral intuitions. Specifically, the FW model accounted for about 24% of the variance (R^2_{adjusted} =.24). The effects of all predictors, the N&K group contrasted with Rollback and Video (ps <.001), bypass ($\beta_{standardized}$ = –0.29, t = –9.16, p <.001), intrusion ($\beta_{standardized}$ = 0.06, t = 1.98, p =.048), and desire ($\beta_{standardized}$ = 0.21, t = 7.43, p <.001) were significant. Interestingly, MR was better predicted by these variables (R^2_{adjusted} = .46). The significant predictors of MR were N & K group contrasted Rollback and Video (ps < .001), bypass ($\beta_{standardized}$ = -0.13, t = -4.99, p < .001), and desire ($\beta_{standardized}$ = 0.59, t = 24.46, p < .001), whereas intrusion did not have a significant effect ($\beta_{standardized}$ = 0.03, t = 1.23, p = .22). These results suggest that 1) more bypass judgments, 2) less desire to punish, and 3) introduction of a deterministic world via N&K scenarios were consistently associated with more incompatibilist responses. However, intrusion judgments were less correlated with moral intuitions when modeled simultaneously with bypass and desire, calling into question the relationship between intrusion and increased compatibilist responses proposed by Nadelhoffer and colleagues (2020; 2023). This inconsistency would be because of (1) the negative correlation between intrusion and bypass (Spearman's ρ = –.45 in the present sample) and (2) the fact that far fewer people have misunderstanding when it comes to intrusion compared with bypass.

Importantly, our results revealed that the desires for punishment more strongly correlated with increased compatibilist responses than comprehension errors when it comes to MR, suggesting a profound influence of motivated cognition on moral intuitions. To further substantiate this finding, we conducted a drop-one analysis that assesses the impact of the bypass and desire variables. This analysis involved observing changes in the model's R-squared value when individually removing each variable. The findings indicated that removing the bypass variable led to a decrease in the variance explained: for the Free Will (FW) model, it dropped from 24% to 17%, and for the MR model, from 46% to 45%. In

contrast, eliminating the desire variable resulted in a more significant decrease in explained variance: for the FW model, it fell from 24% to 19%, and for the MR model, it plummeted from 46% to 14%. These results underscore the crucial relation between both bypass and desire and moral intuitions, with a particularly significant impact of the desire variable in the context of moral responsibility. It is important to note that this result primarily indicates correlational relationships and does not necessarily imply that retributive desire is the primary driver of responsibility attribution. We will further examine this topic in the subsequent section.

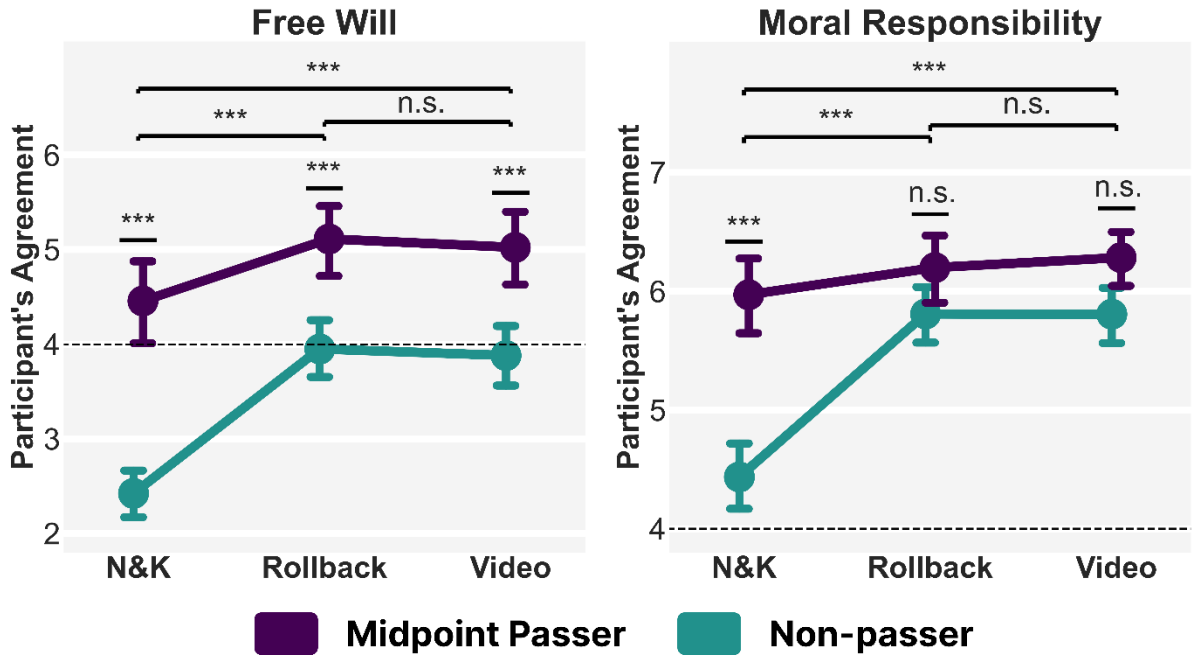

Figure 3. Moral intuitions across conditions and comprehensions. Participants whose averaged scores in bypass and intrusion questions were both below midpoint were labeled as midpoint passer (N = 224) while others were classified as non-passer (N = 775). For free will attribution, a two-way ANOVA revealed the significant main effect of condition and comprehension without their interaction. According to post hoc analysis, midpoint passers and people in Rollback and Video condition had more compatibilist responses than non-passers and those in N&K condition, for each (*p*s < .001). For moral responsibility attribution, there was a significant interaction between the condition and comprehension, in addition to their significant main effects. Post hoc analysis indicated that non-passers in N&K condition had less compatibilist intuitions compared with other combinations (*p*s < .001). The upper three annotations indicate the statistical significance of the condition while lower three represent the significance of comprehension for each plot. Dashed lines represent a midpoint of each item. N&K: Nichols & Knobe (2007)'s original condition; n.s.: no significance; ***: *p* values below .001

4. Discussion

Our research leads to at least five discussion points.

First and foremost, our findings challenge Nadelhoffer et al.'s (2020) intrusion hypothesis, according to which many compatibilist responses are explained by the intrusion of indeterministic assumptions into people's understanding of determinism. Our regression analysis showed that intrusion has only a weak effect on FW scores and no independent effect on MR scores when controlling for the effects of bypassing judgments and the desire for punishment. As stated, intrusion's limited explanatory power in this analysis is attributable to its low mean score. The mean bypass score among conditions was 4.46 ± 1.45, while for intrusion it was 2.44 ± 1.50. This indicates that the majority of participants comprehended that determinism precludes alternative possibilities and necessitates every event. In other words, most participants were free from intrusion. Alternatively, intrusion's limited explanatory power can be explained by the inverse correlation between bypass and intrusion scores. This suggests that intrusion effects might actually reflect bypassing effects; i.e., the compatibilist responses that intrusion appears to explain might be due to (increased or decreased) bypassing judgments. Overall, our findings challenge the importance of intrusion and raise doubts about the intrusion hypothesis.

Second, our regression analysis highlights that the desire for punishment is closely associated with both FW and MR scores. Notably, study participants' desire (rather than comprehension errors) is predominantly associated with MR scores. One possible interpretation of this correlation is that the presentation of wrongful actions triggers a retributive desire. This, in turn, increases compatibilist responses, which aligns with two of Clark et al.'s (2019) claims: (1) people's desire to uphold moral responsibility drives compatibilist intuitions and (2) commitments to free will or responsibility attributions are inherently motivation-dependent. However, given that our study did not control or manipulate participants' desire, we are unable to ascertain how to interpret the correlation between desires and MR scores at this stage. Further research is required to determine how to interpret the role of desires in this context.

Third, participants who comprehended determinism were more inclined to be compatibilists. Out of the 999 participants who passed attention checks, 22.42% scored below the mean on the bypass and intrusion material (midpoint passer). Figure 2 illustrates how midpoint passer responses were generally compatibilist (on average, they agreed with the attribution of free will and responsibility to determined actions across three conditions). Our study suggests that people who correctly understand determinism are likely to respond in a compatibilist way when faced with certain immoral actions. This supports natural compatibilism. Note that, however, average MR scores of non-passers (participants who did not pass either bypass or intrusion checks) were also above the midpoint in all conditions. Moreover, many participants who conflated bypassing with determinism admitted responsibility attribution, which could pose a challenge for compatibilists. We shall return to this point in the next section.

Fourth, there were significant scenario effects on intuitions. These were independent of the effects of desires for punishment and comprehension errors. N&K participants tended to show incompatibilist responses about both free will and moral responsibility compared with Rollback and Video participants. Even if we take factors such as bypass and intrusion into account, this difference is still statistically significant. Participants in the rollback conditions tended to be compatibilists about moral responsibility regardless of the degree to which they understood determinism. This suggests that the rollback account of determinism itself increases compatibilist responses. One potential explanation is that the expression included in N&K affects how an action is represented. An important feature of the original N&K is its

emphasis on causal necessity. The N&K contains the phrase “completely causally determined.” This might weaken the explanatory significance of the relevant agents’ mental states, which, in turn, leads to a weakened attribution of free will and responsibility. According to the “explanation hypothesis” (Björnsson 2014), we attribute moral responsibility to an agent only if her mental state provides a significant explanation for her actions. In light of this, we can explain the reduced attribution of responsibility in N&K as resulting from the reduced explanatory significance of mental states in the relevant scenario.

It seems equally plausible, however, that the rollback cases did not adequately describe a deterministic world. In contrast with the approach taken by N&K, the rollback cases do not explicitly assert that every event is causally necessitated by natural laws and preceding factors. It is therefore possible to understand rollback cases as describing a rollback universe which produces the same event every time, while still allowing for the theoretical possibility of producing different events. It can thus be posited that N&K and Rollback represent distinct universes: a deterministic universe and a non-deterministic rollback universe. If this is the case, the divergent responses observed between these two kinds of scenarios can be attributed to the fact that they reflect different intuitions about different universes. Based on the intrusion score in rollback cases, it appears that the majority of participants understood the rollback universe presented as a deterministic universe, but this possibility cannot be completely excluded.

In any case, the significance of scenario effect indicates the potential for certain variables to be influenced by the presentation of disparate universes, which may, in turn, affect participants' responses across different conditions. As rollback cases have been employed in other studies within the field of experimental philosophy of free will, this is not a concern that is unique to our own study. Nevertheless, further investigation is necessary to address this issue.

Fifth, our research has some limitations in terms of generality. Our experiment was conducted on Japanese participants. We must, therefore, recognize that the effect of desires for punishment might not generalize to other cultures. We need further cross-cultural research. There are also potential concerns related to the translation of materials. It is possible that our results simply reflect cultural or linguistic differences. Nonetheless, we consider this to be unlikely given that our research successfully replicated prior research. Most empirical research on determinism comprehension has been conducted in Western countries. However, we replicated both the comprehension problem and the effects of bypassing on people’s responses discussed in Nadelhoffer et al. (2023). We also replicated the compatibilist tendencies in concrete conditions from Nichols and Knobe (2007). It is, then, reasonable to assume that our research does not invoke significant problems related to cultural differences and/or material translations[8].

[8] One might posit that the limited explanatory power of the intrusion variable is attributable to cultural differences or translation issues. Indeed, a recent study by Murray and colleagues (forthcoming) also investigated the total effects of comprehension errors on Amazon's Mechanical Turk. Their findings indicated that the influence of bypassing errors is less pronounced than that of intrusion errors. In other words, their findings were contrary to our own. However, our findings regarding intrusion errors are consistent with those of other studies conducted in Western countries, which also raises questions about the significance of intrusion errors. For instance, a recent study by Cova and Martinez (2024) suggests that pervasive intrusion errors may be attributed to the quality of responses or the issue with the survey platform. Therefore, our findings regarding intrusion errors cannot be easily attributed to cultural differences.

5. Philosophical Implications

This study has important implications for the philosophical debate about natural compatibilism versus natural incompatibilism. The results indicated that individuals who demonstrated an understanding of deterministic scenarios were predominantly inclined towards compatibilist perspectives. This aligns with the folk compatibilist theory, or natural compatibilism, which posits that the majority of laypeople possess compatibilist intuitions. The fact that the majority of laypeople have compatibilist intuitions does not automatically settle the philosophical debate between compatibilism and incompatibilism. Nevertheless, the question of whether people have compatibilist or incompatibilist intuitions is relevant in determining the burden of proof (see Nahmias et al., 2006). Therefore, given our findings, it falls upon those who espouse incompatibilism to elucidate why incompatibilism is a tenable position despite its apparent counter-intuitiveness. It seems that our results provide, if not definitive, counterevidence to incompatibilist theories.

However, it is premature to conclude that folk intuitions are compatibilist or that compatibilist theories are intuitively justified at this stage. This is for at least three reasons.

First, it remains controversial whether people who understand determinism are compatibilists. We have examined intuitions only in one concrete case. Nichols and Knobe's (2007) much-discussed study suggests that (a) people tend to give compatibilist responses when their intuitions are probed with concrete descriptions of immoral actions and (b) people tend to give incompatibilist responses when their intuitions are abstractly probed without presentations of concrete actions. People with a correct understanding of determinism might show incompatibilist responses when asked about their intuitions in abstract conditions.

Second, most participants gave compatibilist responses even when making bypass judgments. This seems to be consistent with compatibilist theory. There is, though, a notable caveat here. Most compatibilist theories do not agree with attributions of free will and moral responsibility in a world where mental states have no causal efficacy. Mainstream compatibilist theories (e.g., Fischer & Ravizza 1998) require causal integration between an agent's mental state and her actions (such as responding to reasons for performing or not performing some action). Even if people intuitively grant free will/moral responsibility attributions while making bypassing judgments, it is questionable whether such intuitions are consistent with a compatibilist theory (see Feltz and Millan 2013). We found that most participants whose bypass scores were above the midpoint also gave compatibilist responses about moral responsibility. Mean FW scores were 3.39 and mean MR scores were 5.68 for the 649 participants whose mean agreement with bypass statements was above the midpoint. This finding suggests that the folk are insensitive to compatibilist standards for attributing responsibility. It also suggests that their outwardly compatibilist responses do not justify mainstream compatibilist theories of moral responsibility.

Third, our research found that people's desire for punishment is related to compatibilist responses. If compatibilist intuitions are primarily driven by desires, then their value as evidence for compatibilist theories becomes suspect. The substantial role of motivated cognition might undermine the compatibility question about free will and determinism itself, a question that is prescriptive in nature and whose answer should be independent of personal desires. We can, of course, interpret desire's effect the other way around. Compatibilist responses themselves might cause the desire to uphold moral responsibility (rather than the other way around). If so, then desire's strong effect does not imply that people's free will and responsibility attributions inherently depend on motivation. Our experimental

results do not allow us to determine which model is correct. Further research is needed to determine how exactly desire affects intuitions.

Conclusion

The primary goal of our study was to uncover the overall effects of desires for punishment and comprehension errors on people's intuitions about free will and moral responsibility in deterministic situations. These factors are expected to influence people's responses to compatibility questions. The total effect of each factor has, though, not been fully investigated. We set out to achieve this goal via a large-scale experiment with a new version of the rollback scenario and videos describing it.

By employing a large-scale survey with new descriptions of determinism and comprehensive measurements, we found that intrusion's effects are limited compared with other factors. We also found that bypass significantly affects people's responses, even when we control for other factors. This leads to an increase in incompatibilist responses. It is also noteworthy that there is a strong correlation between FW/MR scores and desire variables. Moreover, in the context of responsibility attribution, the effect of comprehension errors is less pronounced than that of desire. These findings lend support to the view put forth by Clark et al. (2019) that motivated cognition plays a pivotal role in the attribution of free will and moral responsibility, and that there is no firm intuitive basis for the concept of free will.

However, given the settings of our study, there are multiple possibilities regarding the causal direction of retributive desire. It is unclear whether it is increased desire that drives attribution of responsibility and whether motivated cognition increases compatibilist responses. Furthermore, in light of the significant scenario effects we found on responsibility attribution, which are unreducible to other variables, our study suggests the existence of other potential factors, which are not tracked by either bypassing or intrusion sentences nor by desire sentences.

Acknowledgements

This research is funded by JSPS KAKENHI (21H00464: 22KJ0110: 22KJ0109) and TOYOTA Foundation (D22-ST-0028). Preliminary findings of this study were presented at the 5th research meeting on Philosophy and Ethics of blame held in 2023, Kanazawa, and the 4th European Experimental Philosophy Conference held in 2023, Zurich. We are grateful for the valuable feedback received during the presentation, which helped improve this manuscript. We are also grateful to Gunnar Björnsson for his helpful feedback on earlier versions of the manuscript.

References

Andow, J., & Cova, F. (2016). Why compatibilist intuitions are not mistaken: A reply to Feltz and Millan. *Philosophical Psychology*, 29(4), 550–566. https://doi.org/10.1080/09515089.2015.1082542

Björnsson, G. (2014). Incompatibilism & 'Bypassed' Agency. In A. R. Mele (ed.), *Surrounding Free Will*, Oxford University Press, 95–112.

Clark, C. J., Luguri, J. B., Ditto, P. H., Knobe, J., Shariff, A. F., & Baumeister, R. F. (2014). Free to punish: A motivated account of free will belief. *Journal of Personality and Social Psychology*, 106(4), 501–513. https://doi.org/10.1037/a0035880

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. https://doi.org/10.3389/fpsyg.2019.00215.

Cova, F., & Martinez, T. (2024). Failure to comprehend determinism or failure to measure comprehension? Methodological issues in experimental philosophy of free will. https://doi.org/10.31234/osf.io/3j6v8

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

Feltz, A. & Millan, M. (2013). An Error Theory for Compatibilist Intuitions. *Philosophical Psychology*, 28 (4): 529–555. https://doi.org/10.1080/09515089.2013.865513

Fischer, J. M. & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

Inarimori K, Honma S and Miyazono K (2024) Do we have (in)compatibilist intuitions? Surveying experimental research. *Frontiers in Psychology*. 15:1369399. doi: 10.3389/fpsyg.2024.1369399

Murray, D. & Nahmias, E. (2014). Explaining Away Incompatibilist Intuitions. *Philosophy and Phenomenological Research*, 88 (2): 434–467. https://doi.org/10.1111/j.1933-1592.2012.00609.x

Murray, S., Dykhuis, E., & Nadelhoffer, T. (Forthcoming). Do People Understand Determinism? The Tracking Problem for Measuring Free Will Beliefs. *Oxford Studies in Experimental Philosophy*, 5.

Nadelhoffer, T., Rose, D., Buckwalter, W., & Nichols, S. (2020). Natural Compatibilism, Indeterminism, and Intrusive Metaphysics. *Cognitive Science*, 44 (8): e12873. https://doi.org/10.1111/cogs.12873.

Nadelhoffer, T., Murray, S., & Murry, E. (2023). Intuitions About Free Will and the Failure to Comprehend Determinism. *Erkenntnis*, 88 (6): 2515–36. https://doi.org/10.1007/s10670-021-00465-y.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying Freedom: Folk Intuitions About Free Will and Moral Responsibility. *Philosophical Psychology*, 18 (5): 561–584. https://doi.org/10.1080/09515080500264180.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2006). Is Incompatibilism Intuitive? *Philosophy and Phenomenological Research*, 73 (1): 28–53. https://doi.org/10.1111/j.1933-1592.2006.tb00603.x

Nahmias, E. & Murray, D. (2010). 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*. Palgrave-Macmillan, 189–215.

Nichols, S. & Knobe, J. (2007). Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions. *Noûs*, 41 (4): 663–685. https://doi.org/10.1111/j.1468-0068.2007.00666.x.