Philosophical thought experiments elicit conflicting intuitions

Anonymous CogSci submission

Abstract

In many cases, when encountering a philosophical thought experiment, different people give different answers. Some people say that the correct answer is A while others say it is B. Despite this fact, they may nonetheless share the same intuitions. That is, many of the participants who ultimately select option B may have an intuition drawing them toward option A, and *vice versa*. Two studies explored the possibility that people have such conflicting intuitions. Study 1 used a self report methodology; Study 2 used mouse-tracking. Both studies found evidence for conflicting intuitions, but both also found evidence of systematic variation across different answers to different thought experiments. In particular, the results suggest that when people ultimately judge that answer B is correct, they are especially inclined to have an intuition drawing them toward A to the extent that a large proportion of participants judged that A is the correct answer.

Keywords: thought experiments; cognitive conflict; intuition

Existing research on intuitions about philosophical thought experiments typically finds that different participants give different answers. If one simply presents a philosophical thought experiment and asks for an answer, one typically finds that some participants say that the answer is A while others say that the answer is B. Research on intuitions about these thought experiments is often concerned with questions about what is drawing people to each of the possible answers (Almeida et al., 2025; Cushman, 2013; Nichols, 2014).

Recent work has distinguished two different ways of explaining the split observed in people's responses (Fischer et al., 2023; Knobe, 2026). One approach would be to say the split is primarily a matter of *different people having different intuitions*. Some people have intuition A, while others have intuition B. A second approach would be to say that the split is primarily a matter of individual people having *conflicting intuitions*. That is, perhaps it is primarily a matter of individual people having both intuition A and intuition B.

To illustrate, consider research on how people respond to thought experiments about free will (Deery et al., 2015; Hannikainen et al., 2019). This research finds that some participants say that the person in the thought experiment has free will, while others say that the person does not have free will. What is happening within the minds of the participants who say that the person does not have free will?

One view would be these participants simply don't have an intuition drawing them in the direction of attributing free will. The other view would be that even though these participants conclude in the end that the person does not have free will, they do feel an intuition drawing them to the view that the person does have free will.

Within existing work, there are a number of studies that have explored this question as it applies to one or another individual thought experiment. These studies provide important insights regarding individual thought experiments, but since the different studies use such different methods, they do not facilitate comparisons across different thought experiments. Within work on the Ship of Theseus thought experiment and on individuation of persons, researchers have asked participants directly whether both answers to the thought experiment are correct (Dranseika, 2024; Dranseika et al., 2024). Within work on lying and on personal identity, researchers have asked whether one answer is right in a certain sense while the other is right in a deeper sense (Knobe, 2022; Skoczeń, 2026). Within work on free will, researchers have given participants contradictory statements embedded within a long list of other statements, so that participants could potentially endorse both without noticing the contradiction in their intuitions (Deery et al., 2015). Within work on legal interpretation, researchers have used within-subjects designs to show that few participants give the same answer consistently, while most participants waver between opposing answers from one case to the next (Engelmann et al., 2024). Within work on dualism, researchers have explicitly asked participants about the degree to which they are experiencing conflict ("How divided did you feel when giving your response?"; Cruz & Mata, 2026).

In the present research, we explore this question more systematically by applying the same method across a range of different thought experiments. Specifically, we look at people's intuition from 15 different thought experiments taken from various areas of philosophy (brain in a vat, trolley problem, Twin Earth, etc.). The aim is to get a better understanding of the overall prevalence of conflicting intuitions across thought experiments and also to determine whether there are differences such that people show more conflicting intuitions in certain types of cases than in others.

In addition to asking participants to introspect about whether they had conflicting intuitions about the philosophical thought experiments, we employed a behavioral method, *mouse-tracking* (Freeman, 2018; Kieslich & Hilbig, 2014), to evaluate whether they showed signs of conflict while they decided which answer to provide. In this paradigm, partici-

pants are asked to select one of two responses, A or B, which are lateralized on participants' computer screens. Numerous studies have demonstrated that participants' hand movements as they guide the cursor toward a response capture important aspects of their decision-making process—such as the degree to which they are confident about one answer, or torn between the two. This method has been successfully applied to document, for example, continuous updating during spoken-word recognition (Spivey et al., 2005), the impact of gender stereotypes on face perception (Barnett et al., 2021) and response competition during interference tasks, such as the well-known Stroop and Flanker tests (Ye & Damian, 2023).

We report the results of two experiments, obtaining both self-reported (Study 1) and behavioral (Study 2) evidence of intuitive conflict. These two studies made use of the same 15 thought experiments. Our research questions, sampling and analysis plans were pre-registered on *AsPredicted.com*. Anonymized data, materials, analysis scripts, and a full report of the pre-registered analyses, have been made available on the project's *Open Science Framework* (OSF) repository.

Study 1

In this first study, we used a self-report measure to explore the degree to which people have conflicting intuitions about philosophical thought experiments. Participants were presented with a philosophical thought experiment and then asked to provide an answer to a question about it. However, after providing their own answer, participants were asked whether they also had an intuition drawing them toward the opposite answer—the one that they did not choose. Our aim was to assess the overall degree to which participants report having an intuition drawing them toward the opposite answer and also to determine whether the tendency to give this response differs systematically across different answers to different philosophical thought experiments.

Method

Participants A U.S. nationally representative sample of 1799 participants were invited to take part in Study 1, of whom 89 failed the attention check at the beginning of the study. We excluded an additional 93 participants for failing the comprehension question about the thought experiment they received. Thus, we analyzed data from 1617 participants.

Procedure In a between-subjects design, participants were assigned to a philosophical issue from the battery of 15 thought experiments. Each thought experiment was taken word-for-word from a study in the existing literature. Materials for all the thought experiments are available on the OSF repository for this paper.

After reading the thought experiment, participants were asked to make a judgment about that thought experiment. Thus, for the Truetemp question, they were asked whether the person in the story knows that it is 22°; for the Twin Earth question, they were asked whether the substance on Twin Earth is water; and so forth. All judgments were framed as a choice between two options.

Regardless of which option participants selected, they were then asked whether they also felt an intuition drawing them toward the opposite option. This second question was phrased as follows: "On the previous page, you answered that [*participant's answer*]. Even though you concluded in the end that this was the right answer, we want to know whether you initially felt pulled in both directions. When answering did you also have the feeling that [*opposite of participant's*

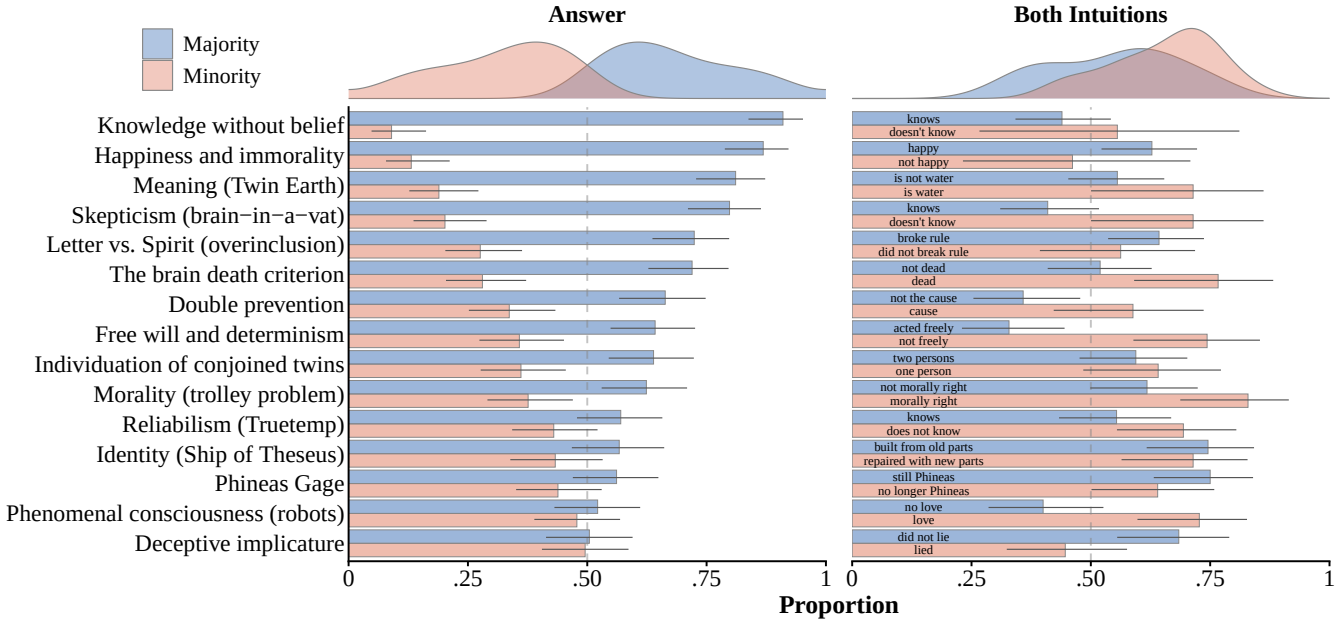

Figure 1: Proportion of (left) each answer by issue and (right) proportion who had both intuitions by issue and answer.

answer], or did you not have that feeling?" Participants then selected either "I did" or "I did not."

Results

Figure 1 shows the proportion of participants who gave each answer on each philosophical issue and the proportion giving each answer who reported having both intuitions.

For each separate issue, we asked whether there was a significant difference between answers such that people are more inclined to reporting having both intuitions when they give one answer rather than the other, using $\chi 2$ tests of independence (see Figure 1). A significant difference (i.e., $p < .05$) arose for seven of the fifteen thought experiments: the trolley problem, the brain death criterion, skepticism (brain-in-a-vat), phenomenal consciousness (robots), deceptive implicature, free will and determinism, and double prevention. On six of these seven issues, participants who gave the more common answer (i.e., the majority) were less likely to report having both intuitions. The only exception was the deceptive implicature scenario, where the majority response ("did not lie": 50.4%) was only slightly more popular than the minority response ("lied": 49.6%).

To assess whether there is variation in the probability of having both intuitions across issues, we compared a model with a random intercept of issue to an intercept-only model in a likelihood ratio test. The random intercept model (*AIC* = 2185) provided better fit to the data than the intercept-only model (*AIC* = 2200), $\chi^2_{(df = 1)} = 16.97$, $p < .001$-indicating that there was significant variation in the tendency to report conflicting intuitions across thought experiments.

We obtained the overall estimated probability of having both intuitions in the population of philosophical issues from the grand intercept of this model, which was equal to .58, 95% *CI* [.54, .63]. The confidence interval excluded 50%, suggesting that—in the population of philosophical questions from which these fifteen thought experiments are drawn—participants are more likely than chance to feel drawn to both possible answers.

We then looked at two different possible predictors of the pattern of conflicting intuition across the different answers on the different issues. First, for each issue, we calculated the entropy of the distribution of answers on that issue. Thus, for each issue, the entropy would be high if the issue is highly divisive, with many participants choosing each possible answer, and the entropy would be low if the vast majority of participants choose one answer and only very few choose the opposite answer.

Second, for each individual answer on each issue, we calculated the *popularity opposite*, i.e., the proportion of participants who gave the opposite answer. In short, one predictor works at the level of the philosophical issues themselves, with the idea being that participants might have more conflicting intuitions when the issue is more divisive (e.g., lying, where both answers were equally popular) and less conflicting intuition where there is a clear majority in favor of one answer

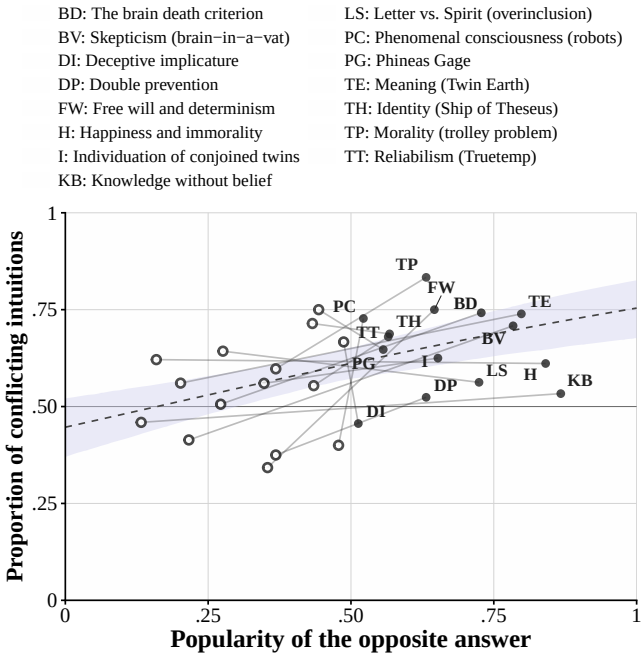

Figure 2: Proportion who report having both intuitions by the popularity of the opposite answer.

(e.g., knowledge without belief, where the majority of participants attributed knowledge). By contrast, the other predictor works at the level of the individual answers, with the idea being that participants might have more conflicting intuitions when they choose an answer such that many participants chose the opposite answer (e.g., those specific participants who were in the minority that did not attribute knowledge on the knowledge without belief issue).

For entropy, we ran a mixed-effects logistic regression model with entropy as a fixed effect and a random intercept of issue (*AIC* = 2185). Entropy was unrelated to the probability of reporting both intuitions, *OR* = 2.76, 95% *CI* [0.79, 9.71], $z = 1.59$, $p = .11$. Thus, highly divisive issues were no more likely to elicit conflicting intuitions than were issues on which a clear majority of participants chose the same answer. For popularity opposite, we ran a mixed-effects logistic regression model with popularity opposite as a fixed effect and a random intercept of issue (AIC = 2169). Popularity opposite positively predicted whether participants reported having both intuitions, *OR* = 3.80, 95% *CI* [2.03, 7.10], $z = 4.19$, $p < .001$ (see Figure 2).

Discussion

In this first study, we used a self-report measure to ask whether participants who say that a particular answer to a philosophical question is the correct one also say that they had an intuition drawing them to the opposite answer. Using this measure, we find that participants respond in approximately half of all cases that they had an intuition drawing them toward the opposite answer of the one they gave.

However, this response was not found equally for all answers to all philosophical questions. One factor that predicts this response is the popularity of the opposite answer. When a majority of participants give one answer and a minority give the other, those participants who give the majority answer often say that they had no intuition drawing them toward the minority answer, whereas those participants who give the minority answer specifically show a tendency to say that they had an intuition drawing them toward the majority answer.

Study 2

In this second study, we moved away from self-report methods and shifted to a method in which cognitive conflict was measured using a mouse-tracking paradigm. The key question was whether this measure would show the same effects observed in Study 1. Thus, we wanted to know whether participants would show evidence of conflicting intuitions in our 15 philosophical thought experiments and whether they would be especially likely to show evidence of conflict in cases where many of the other participants selected the opposite answer from the one that they ultimately chose.

For this second study, we developed a control version of each of our thought experiments. This control version was developed by altering the part of the thought experiment that might make people feel drawn to the minority answer. For example, in the Twin Earth thought experiment, this part is the information indicating that although the substance on Twin Earth is different from water in its deeper microphysical properties, it is exactly the same in its superficial properties. The control version is therefore one in which the substance is different not only in its deeper microphysical properties but also in its superficial properties. This control version would be predicted to not generate cognitive conflict.

In the original versions of the thought experiments, one hypothesis would be that even those participants who ultimately choose one of the two answers also have an intuition drawing them toward the opposite answer. We test this hypothesis by asking whether participants who receive the original thought experiments show a greater tendency to deviate in their mouse movements in the direction of the answer that they ultimately choose against, relative to those who receive the control versions. Then, for each answer, we also calculate the popularity opposite, i.e., the proportion of participants who chose the answer that was the opposite of the answer that a given participant chose. We ask whether popularity opposite predicts the deviations observed in mouse-tracking and whether it explains any differences we might observe between the original versions and the control versions.

Method

Participants We recruited a U.S. nationally representative sample of 598 participants (median age = 45 years old; 48% women, 50% men, 2% non-binary) on Prolific.com, of whom 15 dropped out and 583 completed the study. We excluded (n = 4) participants whose distribution of response times had an interquartile range below 20 ms, and those (n = 8) who provided eight or fewer correct answers out of the 15 comprehension questions. Additionally, we excluded (i = 728) responses when the corresponding comprehension check was answered incorrectly. These exclusion criteria resulted in a final dataset of 7816 responses provided by 571 participants (∼13.7 responses per participant).

Procedure After providing informed consent, participants were familiarized with the study display by completing a practice question. After the brief practice block, participants were assigned to one of two versions, either the original version or a control version, of each thought experiment from the battery of 15 thought experiments employed in Study 1. Thus, each participant provided fifteen responses—one per vignette pair in either the original or control condition.

Each thought experiment was composed of three screens. On the first screen, we presented the thought experiment but omitted the key question. Participants then had to press

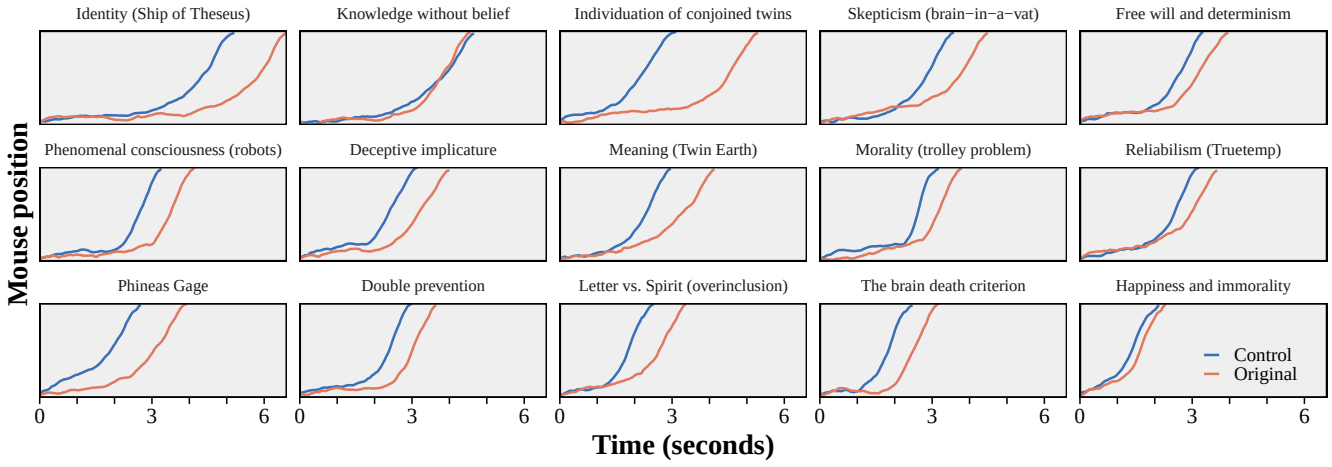

Figure 3: Mean horizontal position over time for each thought experiment and condition.

a "Next" button centered toward the bottom of their screen. Then, the second screen revealed the target question and two lateralized response options on the top half of the screen. To accommodate different screen sizes, the horizontal spacing between response buttons (Mdn = 850 px, IQR = 218 px) was adjusted according to the browser's viewport width. Mouse positions were continuously sampled (at a median rate of ∼ 64 Hz) from the display of the response page to the moment participants clicked on either response button. The third screen presented a comprehension question with the correct and incorrect answers lateralized as in the response page. The left-right placement of response options—both for the key question and the comprehension question—was randomly assigned on each trial.

To statistically compare mouse trajectories of varying durations, each trajectory was time-normalized to 101 equally-spaced time points. Additionally, trajectories toward left-side responses (i.e., with negative x endpoints) were folded along the vertical midline so that all trajectories terminated at positive-valued endpoints.

Results

Figure 3 shows mean (horizontal) mouse positions against corresponding mean timestamps for each philosophical issue and condition.

We treated the horizontal mouse position as our dependent measure–where 0 represents the position at the origin, positive values represent distance toward the option that the participant eventually chose and negative values represent distance in the opposite direction. All trajectories began with a point intermediate between the two options and ended with the option that the participant eventually chose. However, at some point along the trajectory, participants might deviate in the direction of the option that was not chosen. The aim of each of these analyses was to determine what predicts deviations of this type.

First, we ask whether trajectories show greater curvature in the original condition than in the control condition. For this analysis, we regressed the mouse position on condition, time step and their two-way interaction. If the degree of curvature depends on condition, it would show up as an interaction between condition and time step, such that a particular value of the variables predicts a more negative horizontal position for specific time steps.

Condition moderated the effect of time step on mouse position, $\chi^2_{(df=100)} = 833.0, p < .001$. Figure 4 shows the difference between the original condition and the control condition at each time step. Values below 0 on the y-axis indicate an effect such that participants show greater curvature in the original condition than in the control condition. Analyses of the simple effect of condition revealed significantly greater curvature in the original condition during most of the second half of the trajectory, i.e., between the 52nd and 95th time steps.

Second, just as in Study 1, we calculate *popularity opposite*, i.e., the proportion of participants who chose the answer that

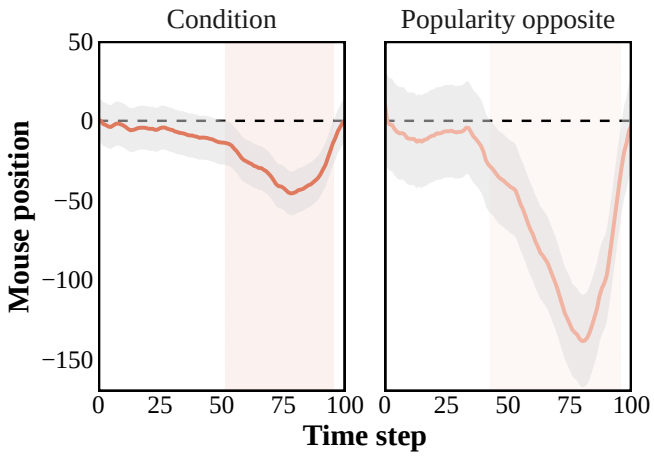

Figure 4: Effects of condition and popularity of the opposite view on mouse position by time step. Y values represent the predicted deviation (left) of the original (vs. the control) condition and (right) from a unit increase in popularity opposite.

was the opposite of the answer that a given participant chose. We then ask whether deviations are predicted by popularity opposite. We regressed the horizontal position on popularity opposite, time step and their two-way interaction. Here, a significant interaction would indicate that when a participant ultimately chooses an option that is endorsed only by a minority, there are some time steps during which that participant shows different curvature from a participant who chooses an option that is endorsed by the majority.

There was a significant interaction between popularity opposite and time step, $\chi^2_{(df=100)} = 1867, p < .001$. Analyses of the simple effect of popularity opposite revealed a negative effect on mouse position during roughly the second half of the trial, i.e., between the 43rd and 95th time steps (see Figure 4).

Having found significant effects for both condition and popularity opposite, we construct a single model in which these two variables compete to predict curvature—i.e., we entered both condition and popularity opposite as moderators of the effect of time step. In this model, popularity opposite moderated the effect of time step on mouse position, $\chi^2_{(df=100)} = 1084, p < .001$, whereas condition did not, $\chi^2_{(df=100)} = 26.24, p = 1$. The simple effect of popularity opposite revealed greater curvature in the original versions than in the control versions between the 51st and the 94th time steps.

Exploratory analyses In Experiment 1, popularity of the opposite view also explained greater intuitive conflict when providing the minority (vs. the majority) response to each thought experiment. We conducted two exploratory analyses, reported in full in the online repository, to understand whether trajectories during minority and majority responses differed, and why.

First, mouse trajectories showed greater curvature during minority (vs. majority) responses—during the latter part of the mouse trajectory (from the 66th to the 91st time step).

Second, this effect was explained by popularity opposite: In a second model controlling for popularity opposite, popularity opposite predicted greater curvature during the second half of the trajectory (49^{th} to 93^{rd} time step) whereas response (majority vs. minority) did not.

Discussion

In Study 2, participants displayed greater curvature in the original thought experiments than in control versions designed to elicit a single intuitive reaction. This difference in curvature was greatest during the second half of participants' decision processes and was explained by the popularity of the opposite view—which was greater on average in the original condition than in the control condition. In other words, participants deviated the most from the response option they ultimately selected when the alternative response was very popular among participants in this study as a whole.

General Discussion

Two studies looked for the presence of conflicting intuitions in philosophical thought experiments, applying the same methodologies to a battery of fifteen different thought experiments from different parts of philosophy. Study 1 used a self-report methodology; Study 2 used mouse-tracking. Study 1 found that participants report conflicting intuitions in slightly more than half of all cases. Both studies found that the proportion of participants showing conflicting intuitions varied substantially across different answers to different experiments. Both studies found that participants were more likely to have an intuition drawing them toward the answer they did not ultimately select if that answer was itself selected by a large proportion of participants.

The present results suggest that people often have conflicting intuitions about a single thought experiment. One way to explore this issue is to consider the total proportion of participants who are drawn to a particular answer to be the proportion who select that answer in the end plus the proportion who select the other answer and to say that they felt drawn toward the answer they did not select. When we do this with the results from Study 1, we find that (1) almost all answers on all philosophical experiments have the property that the majority of participants were drawn to them, and (2) one of the answers to each thought experiment was such that the vast majority of participants were drawn to that answer. Thus the proportion of participants who have each intuition greatly exceeds the proportion who select that answer when ultimately making a choice.

A key question for further research will be how people choose between conflicting intuitions. When people have intuition A and also have intuition B, how do they arrive at a final decision? One possible answer to this question would be that the psychological process people used to choose between intuitions is closely related to the processes that generated those two intuitions in the first place. Another possible answer would be that the psychological process people used to choose between these intuitions is something distinct—a separate process from the one people used to generate the intuitions in the first place.

One obvious way to spell out the first type of answer would be to say that people choose between intuitions using a process of repeated sampling (much like in other domains of cognition; Ratcliff et al., 2016). In each iteration of this process, people have an intuition drawing them in one direction, and they continue the process until this process of evidence accumulation allows one intuition to reach a certain threshold. Thus, people might first have intuition A, then intuition B. Then, to decide between the two, they simply continue this process, with each new intuition being generated from the same probability distribution that was used to generate the first two. One virtue of this answer is that it straightforwardly predicts the relationship we observed between the popularity of each answer and how strongly participants were intuitively drawn to it. The higher the probability of having a given intuition in the first place, the higher the probability of selecting that intuition at the end of the process of choice.

A natural way to spell out the second option would be that people first generate certain intuitions, then go through a distinct process of reasoning that enables them to choose between those intuitions. On this view, people first generate intuition A and intuition B; then they go through a distinct process of reasoning in which they choose between A and B. A virtue of this second answer is that it makes sense of a clear pattern in our mouse-tracking data. Curvature toward the opposite answer tended to occur abruptly and at the later stages of the trajectory. This pattern is very different from what we see in cases that are well modeled in terms of a process of gradual evidence accumulation like the one predicted by the first option (Spivey et al., 2005). It suggests that the resolution of conflicting intuitions is a discrete event later in the decision-making process.

Regardless, the present results suggest something about the nature of the phenomenon that requires explanation. Consider a thought experiment in which the majority of people choose answer A but a substantial minority choose answer B. The present results suggest that we do not usually need a theory that explains why a substantial minority don't have an intuition drawing them toward answer A. Rather, in most cases, what we need is an explanation of why certain people do have an intuition drawing them to answer A but ultimately decide in favor of an intuition drawing them in the opposite direction.

References

Almeida, G. D. F. C. F., Flanagan, B., & Hannikainen, I. R. (2025). Trait empathy predicts a preference for the spirit of the law: Nationally representative survey evidence. *Journal of Research in Personality*, *116*. https://doi.org/10.1234/eg.article

Cruz, F., & Mata, A. (2026). Love is in the soul, math is in the brain: Dualist intuitions and belief in psychological

science. *Journal of Experimental Social Psychology*, *122*, 104845.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*, 273–292.

Deery, O., Davis, T., & Carey, J. (2015). The Free-Will Intuitions Scale and the question of natural compatibilism. *Philosophical Psychology*, *28*(6), 776–801.

Dranseika, V. (2024). Two ships of Theseus. *Synthese*, *203*(6), 201.

Dranseika, V., Nichols, S., & Shoemaker, D. (2024). The identity of what? Pluralism, practical interests, and individuation. *Philosophy and Phenomenological Research*, *109*(3), 757–773.

Engelmann, N., Hannikainen, I. R., González-Garcı́a, C., & Ruz, M. (2024). Understanding rule enforcement using drift diffusion models. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*, 100–105. https://escholarship.org/uc/item/k0e1y2

Fischer, E., Allen, K., & Engelhardt, P. E. (2023). Fragmented and conflicted: folk beliefs about vision. *Synthese*, *201*(3), 84.

Freeman, J. B. (2018). Doing psychological science by hand. *Current Directions in Psychological Science*, *27*(5), 315–323.

Hannikainen, I. R., Machery, E., Rose, D., Stich, S., Olivola, C. Y., Sousa, P., & Zhu, J. (2019). For whom does determinism undermine moral responsibility? Surveying the conditions for free will across cultures. *Frontiers in Psychology*, *10*, 2428.

Kieslich, P. J., & Hilbig, B. E. (2014). Cognitive conflict in social dilemmas: An analysis of response dynamics. *Judgment and Decision Making*, *9*(6), 510–522.

Knobe, J. (2022). Personal identity and dual character concepts. In *Experimental Philosophy of Identity and the Self: Experimental Philosophy of Identity and the Self* (p. 49).

Knobe, J. (2026). Conflicting intuitions. *Ergo*.

Nichols, S. (2014). The episodic sense of self. In J. D'Arms & D. Jacobson (Eds.), *Moral Psychology and Human Agency: Moral Psychology and Human Agency*. Oxford University Press.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Skoczeń, I. (2026). Are Lying and Perjury Dual Character Concepts?. *Law and Philosophy*.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, *102*(29), 10393–10398.

Ye, W., & Damian, M. F. (2023). Effects of conflict in cognitive control: Evidence from mouse tracking. *Quarterly Journal of Experimental Psychology*, *76*(1), 54–69.