In press at *Ergo*

Conflicting Intuitions[1]

Joshua Knobe
Yale University

Research on intuitions about philosophical thought experiments shows a striking pattern. Often, there are powerful intuitions on one side and also powerful intuitions on the exact opposite side. A question now arises about how to understand this pattern. One possible view would be that it is primarily a matter of *different people having different intuitions*. I present evidence for the view that this is not the correct understanding. Instead, I suggest, it is primarily a matter of individual people having *conflicting intuitions*. That is, it is primarily a matter of individual people having an intuition on one side and also having an intuition on the opposite side.

One striking characteristic of the traditional questions of philosophy is the degree to which they tend to be *confusing*. Just try introducing the problem of free will in a typical philosophy course. You will not find that all of the students immediately converge on a particular answer and take that answer to be obviously correct. Instead, even the very first time you explain the question, you will find students expressing wildly different views that appear to provide support for completely different theories. Similar points could be made about numerous other philosophical questions, including questions about skepticism, dualism, consequentialism, or personal identity.

Over the past few decades, work in experimental philosophy has looked in detail at people's ordinary way of thinking about each of these questions and many others besides, and there are now sophisticated research programs exploring people's intuitions in each of these separate areas. In this paper, I argue that these separate research programs seem gradually to be uncovering a more general fact about how to understand the distinctive way in which traditional philosophical questions are confusing.

At least at first, it might seem that the obvious interpretation of what is happening in these areas is that *different people have different intuitions*. For example, it might be that some people have compatibilist intuitions, while others have incompatibilist intuitions, and similarly for each of the other traditional questions of philosophy. Then the puzzlement we feel regarding these questions might be explained in terms of the conflict between the intuitions of different people.

I will argue that experimental research has been providing evidence for a very different hypothesis. On this hypothesis, the difficulty is in large part a matter of *individual people having conflicting intuitions*. Thus, when a person encounters a thought experiment designed to get at one of

[1] For comments on an earlier draft, I am grateful to Guilherme Almeida, Taylor Davis, Vilius Dranseika, Eugen Fischer, Ivar Hannikainen, Shaun Nichols, John Protzko and Kevin Tobia.

the traditional questions of philosophy, she will often have an intuition pulling her in one direction but also an intuition pulling her in the exact opposite direction.

This might seem like an obvious hypothesis, and perhaps it is. Still, I think that its implications haven't been fully appreciated. If true, it would have important implications both empirically and philosophically.

Empirically, the conflicting intuitions hypothesis has important implications when it comes to what exactly our theories about people's intuition are supposed to explain. For example, suppose we are studying people's intuition about causation. We think of an interesting case and give it to a sample of participants. We then find that approximately half of the participants say that this is a case of causation, while the other half say that this is not a case of causation. Now suppose we want to develop a theory about the underlying cognitive processes that generate people's causal intuitions. What exactly is that theory supposed to explain?

If we assume that the difference in responses is due to different people having different intuitions, we would naturally assume that the theory should explain why different people have different intuitions regarding this case. So we might think that our theory of the underlying cognitive processes should explain why half have one intuition and the other half have the opposite intuition.

However, if we conclude that the difference in responses is due to each individual person having conflicting intuitions, there is an important sense in which there is more that needs explaining. An adequate theory would not merely have to explain why half of the participants have the intuition that it is a case of causation; it would have to explain why those people are conflicted. That is, it would have to explain why they have one intuition and also why they have the exact opposite intuition.

At the same time, there is also an important respect in which the conflicting intuitions hypothesis says that a theory of intuitions needs to explain less. Suppose you run a study in which you present participants with a thought experiment and tell them that they have to choose one answer or the other. In such a case, it might be that most participants have an intuition drawing them toward one answer and also an intuition drawing them toward the exact opposite answer. Still, they will have to choose. In some way or other, each participant will have to select one of the two answers, and we will probably find that some participants choose one while others choose the other. The key point now is that a theory about the cognitive processes that generate people's intuitions might not have anything to say about how the different participants choose different answers. As we will see, it can easily turn out that there is one cognitive process that generates people's intuitions and then a completely separate process that enables them to select an answer when their intuitions conflict.

Philosophically, this claim will have important implications if you assume that there is a distinction between the philosophical importance of *intuitions* about a question vs. *beliefs* about which answer to that question is actually correct (see, e.g., Bengson, 2013). To illustrate, suppose that you are working on a question in the philosophy of language. You think about a particular English sentence and immediately have the intuition that this sentence is true. Then you might reflect philosophically about whether that intuition is correct. Ultimately, you might conclude that the

sentence is not true. In such a case, you have an intuition that the sentence is true but a belief that the sentence is not true. These seem to be importantly different states.

If you do see things in this way, you face two different questions about the philosophical importance of empirical facts about how people ordinarily think. One question is about the importance of ordinary *intuitions*. If you learn that most native speakers do not share your intuition about the sentence, how would this impact your philosophical inquiry? The other is about the importance of ordinary *beliefs*. If you learn that most ordinary folks who reason philosophically about this question end up arriving at the opposite belief, how would this impact your philosophical inquiry? At least potentially, you might think that these two questions have two very different answers. For example, if you learn that most ordinary native speakers don't share your intuition about the sentence, you might think that this information would fundamentally call into question one of the starting points of your inquiry. By contrast, suppose you find that most people share your intuitions, but when they reason from those intuitions, they end up concluding that your theory is not correct. In such a case, there is no sense at all in which the starting points of your inquiry has been undermined. Basically, what is happening is just that ordinary folks are engaging in the same kind of reasoning that philosophers most characteristically engage in, and they are then arriving at a different answer.

To the extent that you see things this way, you should think that the conflicting intuitions hypothesis has important philosophical implications. Suppose we give a person a question with two options, and she chooses the first option. If the conflicting intuitions hypothesis holds in this case, then the fact that she chooses the first option and not the second does *not* mean that her intuitions are drawing her toward the first option but not the second. On the contrary, the idea would be that she actually has conflicting intuitions but then arrived at the conclusion that only one of her two intuitions was correct. Thus, if we want to bring out the philosophical significance of her intuitions, we should not be exploring the significance of the option she ended up choosing; we should instead be exploring the significance of the conflicting intuitions themselves.

The remainder of this paper will be exploring the evidence that supports the conflicting intuitions hypothesis, but before we begin, a brief note to lower expectations. I will not be arguing that existing empirical findings provide slam-dunk evidence in favor of this hypothesis. Instead, as we will see, existing evidence is somewhat indirect and inconclusive, and it might well turn out in the end that the conflicting intuitions hypothesis is not quite right. Nonetheless, evidence from a number of different sources does seem to be pointing toward it, and at this point, I think it is the hypothesis supported by the preponderance of existing research.

1. Themes from existing research

In this section, I review some themes from existing research in experimental philosophy. Later on, I will be arguing that these themes have some surprising further implications, in particular, that they provide support for the conflicting intuitions hypothesis. But that will have to wait. For the moment, I will be confining myself to articulating themes that are already present within the existing literature.

1.1. Scope of the present account

This paper will focus entirely on questions that occupy a certain distinctive status in the history of philosophy. Specifically, I will be focusing on questions that have given rise to *long-standing debate*. In questions that have this special status, some philosophers argue that the correct answer is A, while others argue that the correct answer is B. Then, many years later, one still finds a debate with recognizably similar positions. In other words, even after many aspects of the larger philosophical discussion have changed, one can still find some philosophers defending a position that is recognizably A and others defending a position that is recognizably B.

Long-standing debates play an important role in the history of philosophy. A few examples include: the debate between compatibilism and incompatibilism, the debate between dualism and physicalism, the debate between deontology and consequentialism, the debate between views according to which there are objective moral truths and views according to which there are not.

I hasten to add that the claims I will be arguing for here only apply to these questions and do not also apply to other sorts of questions. Suppose we instead turn instead to questions about which the vast majority of philosophers agree (e.g., the question in normative ethics as to whether it is morally wrong to torture innocent children). If we look at intuitions about those sorts of questions, we might not find that those intuitions have any of the features I will be discussing here.

In thinking further about these issues, it will prove helpful to have a running example involving a specific philosophical question. I will be focusing here on the debate in the philosophy of law between *textualism* and *purposivism*. The best way to understand this debate is to begin with a simple example. Imagine a town in which there are a lot of loud parties that keep people up at night. The town makes a rule that says it is illegal to make music in your home after midnight if you live within 50 feet of another person. One day, a person is coming home to her apartment late at night and, almost inaudibly, she starts singing a little song to herself. Now, here is the question: Did this person violate the rule?

In a case like this one, it seems that the person violates the text of the rule, but does not violate the purpose of the rule. Thus, the characteristically textualist response would be to say that the person violated the rule, while the characteristically purposivist response would be to say that the person did not violate the role.

Experimental research over the past few years has given us extremely detailed information about the pattern of people's intuitions in thought experiments like this one (Almeida et al. in press; Almeida in press; Hannikainen et al. 2022; Struchiner et al. 2020), and I will be using this thought experiment as a case study throughout the remainder of the paper. However, nothing that I say is intended to be specific to this one thought experiment, and whenever I illustrate a claim using it, I will also provide evidence for the same claim from other areas of experimental philosophy.

Before moving onward, it might be helpful for me just to report my own subjective reaction on encountering this thought experiment. As soon as I heard it, I felt the intuition that the agent violated the rule, but I also felt the intuition that the agent did not violate the rule. I subsequently learned a bit about the philosophical arguments that have been offered in this debate and began to arrive at a belief about which side was correct. However, even if I would now respond that one side

is correct and the other is incorrect, it would be completely mistaken to deny that I have intuitions drawing me in both directions.

The core hypothesis I will be defending regarding this case is that ordinary folks – people with no prior training in philosophy – tend to react to it in that same way. When they encounter this case, they tend to have conflicting intuitions.

1.2. Split responses

When experimental philosophers first began running studies on people's intuitions about the traditional problems of philosophy, one might reasonably have expected the result to be that people's responses would sometimes line up pretty clearly with one or another of the traditional answers to those questions. For example, one might have expected that people's responses to questions about the mind-body problem would overwhelmingly favor dualism, or their answers to questions about metaethics would overwhelmingly favor realism, or that their responses to questions about free will would overwhelmingly favor incompatibilism.

One striking finding from the last twenty years of experimental philosophy research is that none of these things happened. In no case did we find that people's responses overwhelmingly favored one or another specific view. Instead, what we found in each case was what I will call *split responses*. When you look at studies on people's intuitions regarding any of these questions, what you find is that there are some responses that seem to favor one view and other responses that seem to favor the opposite view.

As one illustration of this broader phenomenon, consider again our case study involving the textualism/purposivism debate. In an important study, Flanagan and colleagues (2023) presented participants with a number of different vignettes in which textualism and purposivism come apart. In all vignettes, the agent performs a behavior that violates the text of the rule but does not violate the purpose of the rule. Each vignette then involved a different specific rule (a rule against cars in the park, a rule against dogs in a restaurant, a rule against phones in a classroom, etc.).

Flanagan and colleagues made their data available, and I used those data to create a figure that shows the number of participants giving textualist or purposivist responses for each vignette (Figure 1).[2]

[2] All code for all figures that appear in this paper can be found at https://osf.io/5skw6. In almost all cases, the figures that appear here are very different from the figures included in the papers where the experimental results were originally reported, but this should not be seen as an implicit criticism of those original papers. The reason why figures in this paper are so different is just that this paper is focused on a very different question.

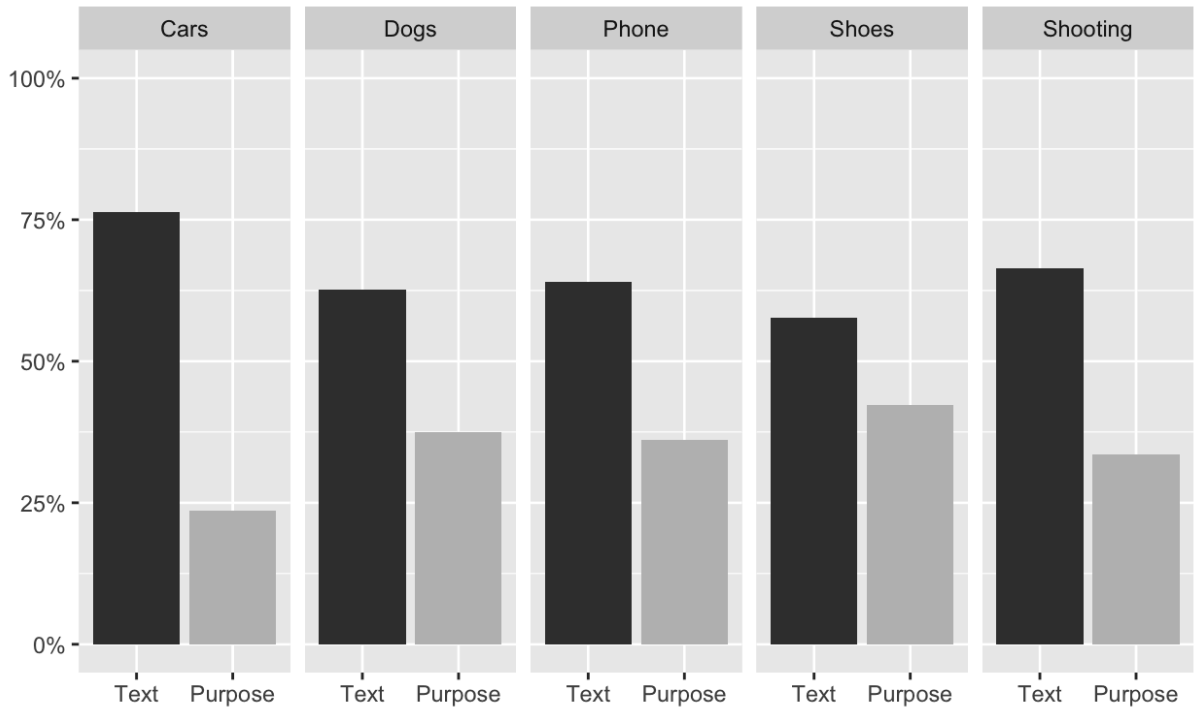

Figure 1. Percent of participants giving textualist vs. purposivist response by vignette. Based on data from Flanagan et al. (2023).

Looking at these results, one finding that immediately stands out is that participants are not equally divided between textualist and purposivist responses. Instead, the proportion of textualist responses is substantially higher. This is an important finding, and one might plausibly think that it provides evidence against the conflicting intuitions hypothesis. If one assumes that each individual participant has conflicting intuitions, one would need to have some explanation of why participants show this disproportionate tendency to arrive in the end at textualist responses. This is an important issue, and I will have a whole lot to say about it in section 2.1.

But in this section, I want to focus on a much more straightforward finding. The results show that participants in this study give split responses. A substantial portion give textualist responses, but a substantial portion give purposivist responses. Thus, research in this area has focused on trying to explain both responses. Researchers have developed theories about why so many people give textualist responses and why so many give purposivist responses.

This basic result has been obtained again and again in studies on all sorts of issues. For another example, consider intuitions about consciousness. Within this literature, there has been a great deal of work about whether people have broadly functionalist intuitions or whether their intuitions depend on facts about physical realizers. The big question driving this research is whether there is a difference between the criteria determining people's intuition about different mental states and, if so, how to understand that difference. But suppose we ignore that whole issue and simply look at the distribution of responses for a single mental state.

For one nice example, consider a recent study by Sytsma and Snater (2023). Participants were asked to imagine an entity that is functionally identical to the human mind but that is physically realized on a computer.

> Imagine that in the future scientists are able to exactly scan a person's body and brain at the molecular level. Using the information from the body scan they can create an android body that is externally indistinguishable from the original person. And using the information from the brain scan they can create a perfect computer simulation of the working brain. They can then embed that computer in the android body to create a duplicate of the person.
>
> Imagine that scientists scan your body and your brain and use that information to create an exact android duplicate of you. What, if anything, do you think this duplicate would be capable of?

Then they were asked whether they agreed or disagreed with the statement: "The android would feel pain when she [he] is injured." Participants rated this statement on a scale from 1 ("Disagree Strongly") to 7 ("Agree Strongly").

Figure 2 shows the results. Clearly, these results do not point to a single response that almost all participants share. Instead, what we see are split responses. There is a sizable proportion of participants saying that they agree strongly, but there is also a sizable proportion saying that they disagree strongly.

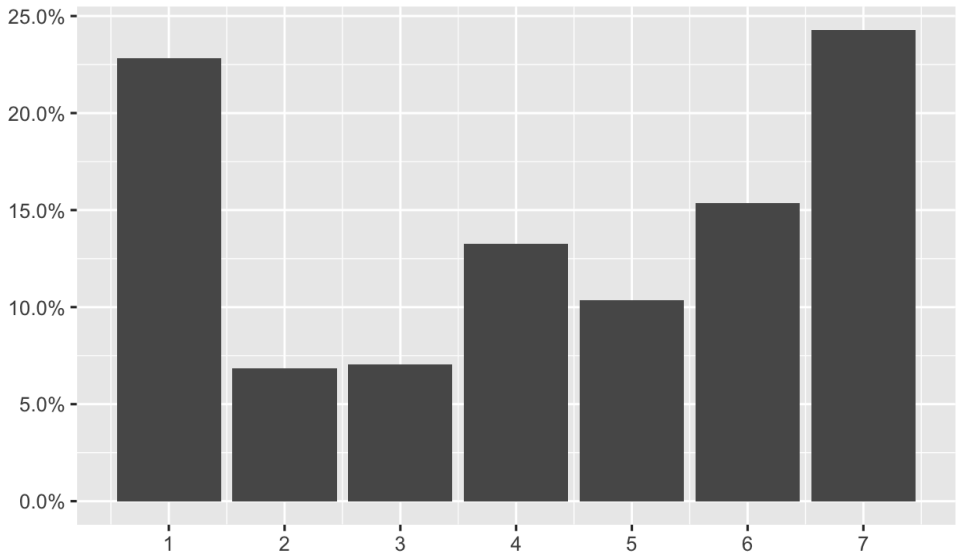

Figure 2. Histogram showing the distribution of responses to a question about whether an android can feel pain. Based on data from Sytsma and Snater (2023).

For yet another example, consider intuitions about personal identity. In a well-known study, Tobia (2015) gave participants vignettes about a person who lost some of his most important traits.

Participants were then randomly assigned either to a condition in which the traits that were lost were morally good or to a condition in which the traits that were lost were morally bad.

In the first condition, participants were told to imagine that Phineas is a kind person who always helps others. One day, he gets into an accident in which a spike goes through his head. The person who exists after the accident is very different in character. This person is cruel and irresponsible. Participants are then asked whether they agree that the person who exists after the accident isn't even really Phineas at all. In the other condition, participants got a story that went in the opposite direction. Phineas starts out as a cruel and irresponsible person. Then he gets into an accident. The person who exists after the accident is a very kind person who always helps others. The question is again whether they agree that the person who exists after the accident isn't even really Phineas at all.

Figure 3 shows the results. The most salient finding is that participants are more inclined to say that the person after the accident isn't really Phineas when morally good traits are lost, and agreement with the statement in the condition is at approximately the midpoint ($M = 3.3$). However, it is also important to note the nature of the distribution. Even though the mean response in the condition where good traits are lost is higher and is at approximately the midpoint, a substantial proportion of participants in that condition strongly disagree with the statement. So an adequate explanation of the responses would have to explain not only why agreement is higher in that condition but also why a substantial proportion of participants in that condition strongly disagree.

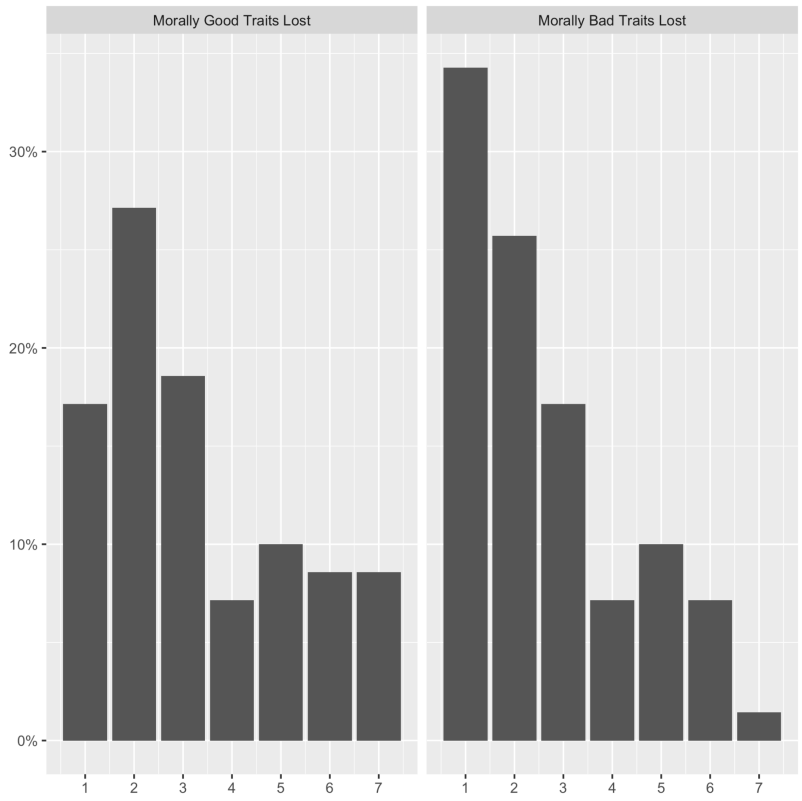

Figure 3. Histogram showing the distribution of responses to a question about whether a person who exists after moral traits have been lost is still the same person as the one who existed previously. Based on data from Tobia (2015).

Finally, consider intuitions about the problem of free will. There has been an enormous and very complex debate within the experimental philosophy literature about whether people should be understood, ultimately, as having compatibilist or incompatibilist intuitions. But putting that debate to one side for a moment, it is clear that people show split responses. If you simply describe a deterministic universe and ask participants whether agents in that universe have free will, a substantial proportion will answer "yes" and a substantial proportion will answer "no."

Figure 4 displays the results from a large-scale (N = 5,268) study using this approach (Hannikainen et al. 2019). We will be exploring some further results from this study below, but for the moment, we can just note a very simple fact. Approximately half of participants say that the agent in a deterministic universe has free will, while the other half say the exact opposite.

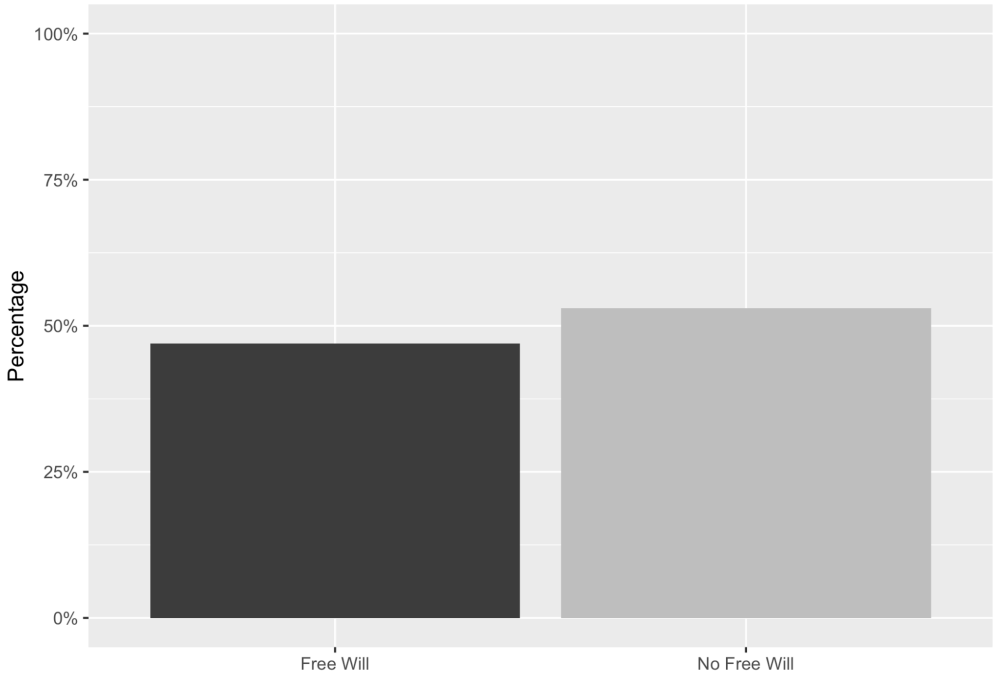

Figure 4. Responses to a question about whether agents in a deterministic universe can have free will. Based on data from Hannikainen et al. (2019).

Across all of these different cases, we find the same basic result, which is that *the disagreement among philosophers is mirrored in the responses of ordinary folks*. If you put together a sample of philosophers and asked them about these cases, you would find that different philosophers would give different answers, and that same pattern appears to arise in the responses of ordinary folks. Different people are giving different responses.

1.3. *Opposing processes*

How can we explain these patterns of folk responses? Clearly, it will not be enough just to say something like "The folk view is A" or "The folk view is B." We need an account that explains why many people give one of the responses but also explains why so many people give the opposite response.

Within the existing experimental philosophy literature, the usual approach to this problem is to develop a theory that posits what I will call *opposing processes*. On this sort of theory, there is a cognitive process within people's minds that tends to generate A intuitions, and then a distinct cognitive process that tends to generate B intuitions. Taken together, the whole theory then explains why the question tends to yield split responses.

Within this literature, different opposing process theories tend to posit completely different cognitive processes. For example, theories designed to explain people's intuition in the philosophy of law explain those intuitions using completely different processes from the ones that appear in

theories designed to explain intuition about personal identity. Yet, despite this obvious difference, these different theories are quite similar in their more abstract structure.

For a first example, let's return to our case of the textualism/purposivism debate. Almeida and colleagues have argued that the split in people's responses in this case can be explained using opposing processes (Almeida, in press; Almeida et al. in press). The explanation relies on a more general theory from cognitive science about how people think about certain categories. This theory says that when people are thinking about certain categories, they can make use of two different processes: one that focuses on concrete features and another that focuses on more abstract values (Knobe et al. 2013; Reuter 2019). As an example, consider the way people ordinarily think about the category of scientists. The theory says that there are actually two very different processes in people's minds that can be used to think about a category like this one. One process focuses on the concrete features associated with being a scientist (running experiments, developing theories, using statistics, etc.). The other focuses on more abstract values associated with being a scientist (e.g., the quest for empirical truth).

Importantly, there will be cases in which these two different processes yield different intuitions. For example, consider a person who has a job as a biology professor and who spends her days running experiments, conducting statistical analyses, and so forth. Then suppose that this person actually isn't at all engaged in an effort to find the truth. She is simply propping up a preconceived dogma as a way of becoming more famous and successful within her profession. Is this person a scientist? The theory says that the two different processes should lead to two opposing intuitions. On one hand, she has all of the right concrete features, so the first process should yield the intuition that she is a scientist. On the other, she doesn't embody the relevant abstract values, so the second process should yield the intuition that there is a deeper sense in which she is not truly a scientist at all.

Almeida and colleagues argue that this more general theory can explain why we find a split between textualist and purposivist responses. The core idea is that one of these processes tends to generate textualist intuitions, while the other tends to generate purposivist intuitions. Consider again the rule against making music at night. In the case we described above, one process would look at the concrete features, note that the agent satisfies all of those features, and then determine that the agent violated the rule. Then the other process would understand the rule in terms of a more abstract value, note that the agent is not going against that abstract value, and therefore determine that there was a deeper sense in which the agent was not violating the rule.

For another example, consider intuitions about sacrificial dilemmas such as the trolley problem. Influential early work by Greene and colleagues (2001) argued that intuitions on these dilemmas should be understood in terms of opposing processes, and this approach has been worked out in a number of different ways within subsequent research. As an illustration, let's consider Cushman's (2013) hypothesis that these judgments should be understood in terms of *model-based* vs. *model-free* representations of value.

This hypothesis draws on a much larger framework developed in cognitive science and artificial intelligence for thinking about how people represent the value of possible actions (e.g., Daw & Shohamy, 2008; Sutton & Barto, 2018). On one hand, people have a causal model that enables

them to make predictions about the consequences of those actions. They can then develop a model-based estimate of the value of an action by thinking about those consequences. On the other, people have the capacity to form representations of the value of the actions themselves. Using this capacity, they can simply assign a model-free value to performing a given action in a given situation. Existing theories suggest that people's ability to perform skillful actions involves a complex interplay of these two representations. When you are trying to decide whether to perform a particular action, you may have a model-based representation and a model-free representation that assign different values to that same action, and behaving skillfully involves using these two representations in a way that enables you to choose the best action.

The claim then is that these two distinct processes tend to lead to two different intuitions in sacrificial dilemmas. Model-based cognition tends to lead to consequentialist intuitions, while model-free cognition tends to lead to deontological intuitions (Cushman 2013). For example, in the footbridge version of the trolley problem, model-based cognition might tend to yield the intuition that pushing the person off the footbridge would yield the best consequences and is therefore morally permissible, whereas model-free cognition might tend to yield the intuition that pushing a person off a footbridge is a bad action in itself and is therefore morally wrong.

Finally, let's consider intuitions about personal identity. In studies on this topic, participants are often introduced to a person who exists at a particular time and told about a property that this person has (a psychological property, a biological property). Then they are told that this property has been lost. Now comes the key question: Is the person who exists after the loss of this property still the very same person who existed at the beginning of the story? Or should we say instead that the original person doesn't even exist anymore, and the person who exists at the end of the story is a fundamentally different person? As we have seen, results from these studies show split responses. Some participants think that the person described at the end of the vignette is the very same person as the one described in the beginning, while others think that if the person at the end of the vignette does not have this property, she isn't even really the same person at all (e.g., Tobia 2015).

How are these split responses to be explained? Within the existing literature, there are a number of different proposals (e.g., Knobe 2022; Tierney 2020), but I will be focusing here on the one developed by Nichols (2014). On this explanation, people have two fundamentally different ways of thinking about the self. The "thick" conception is closely tied to having certain properties, while the "thin" conception is not tied to any properties. Thus, on the thin conception of personal identity, it would make sense to say that, e.g., you could have been a mosquito in a previous life. The mosquito might not have many psychological or biological properties in common with you, but on this conception of this self, being the same self is not ultimately a matter of sharing any such properties.

Nichols argues that the distinction between these two conceptions arises from the distinction between two different cognitive capacities, the capacity for semantic memory and the capacity for episodic memory. When you represent an event using semantic memory, you represent the event as containing different agents, and you could represent one of those agents as having the property of being *you*. But when you represent an event using episodic memory, you are doing something fundamentally different. You are not representing yourself as an agent within the event, who

happens to have certain properties; rather, you are representing the event as a whole from a particular agential perspective. Nichols argues that these two different cognitive capacities ultimately give rise to two different conceptions of the self (thick and thin), which then generate two very different patterns of intuitions regarding philosophical thought experiments.

In this section, we have been introducing the basic idea of opposing process theories, and we have been illustrating that idea with a sequence of different examples. As the examples show, different opposing process theories posit completely different underlying cognitive processes. Yet, despite this difference in content, there seems clearly to be a more abstract structure that all of these theories share.

In what follows, our focus will be on an aspect of such theories that might not initially seem especially salient. This is the fact that opposing process theories are fundamentally *intrapersonal*. That is, these theories do not suggest that some people have a process in their minds that tends to generate A intuitions while others have a process in their minds that tends to generate B intuitions. Instead, they say that individual people generally have within their minds both a process that tends to generate A intuitions and a process that tends to generate B intuitions.

2. Conflicting intuitions

Thus far, we have been discussing two key themes from existing research. One is that people tend to show split responses; the other is that these split responses are explained by opposing processes. I now want to argue that these ideas have an important consequence that has not been sufficiently explored.

Specifically, if one suggests that the split responses are due to opposing processes, it becomes natural to begin thinking of the split responses themselves in a very different way. Suppose we observe that half of the participants give response A, while the other half give response B. The most natural initial assumption would be that half of the participants have A intuitions, while the other half have B intuitions. Then we might begin looking for a theory that explains why different people have different intuitions. At least initially, it might be thought that theories of opposing processes provide just such an explanation.

I will be arguing that this is not the right way to understand opposing processes. If people's intuitions are the product of opposing processes, we should not expect to find that some people have A intuitions and others have B intuitions. Instead, we should expect to find that people generally have both A and B intuitions. When a person has both of these intuitions, I will say that the person has *conflicting intuitions*.

At its core, the argument for this claim is very simple. Suppose people have a process within their minds that generates A intuitions and also a process that generates B intuitions. Then, if both of these processes run to completion, people will have both A intuitions and B intuitions. Thus, the person will have conflicting intuitions.

Of course, I don't mean to suggest that every conceivable opposing process theory would have to predict conflicting intuitions. Clearly, we could create an opposing process theory that did not make that prediction. For example, we could have a theory that posits an additional mechanism such that whenever one of the processes runs to completion, the other process immediately stops

operating. Then there would be opposing processes, but there would be no prediction of conflicting intuitions. In proposing a theory like this one, it certainly feels like we would be introducing an ad hoc auxiliary hypothesis just to avoid predicting conflicting intuitions, but at least in principle, it does seem that such a theory would be coherent.

However, when one looks at the actual opposing process theories that have been put forward in the existing experimental philosophy literature, it seems that this sort of auxiliary hypothesis would not be plausible for any of them. Just to give one illustration, the claim that intuitions about sacrificial dilemmas are explained by a mix of model-based and model-free cognition is embedded in a much larger theory about how these two forms of cognition work together in human decision-making more broadly. Within that larger theory, it would make no sense to suggest that whenever people have a model-free value representation, they stop looking for a model-based value representation. (If the mind worked in that way, it would not be able to achieve the distinctive benefits of having these two forms of cognition; see, e.g., Morris et al., 2021.) Thus, if this theory of intuitions about sacrificial dilemmas is even roughly on the right track, there would be little hope for auxiliary hypotheses that prevent conflicting intuitions. The theory would pretty much have to predict that an individual person can have both a consequentialist intuition and a deontological intuition about the very same case.

To sum up, what I've been offering so far is an argument about how to understand the implications of existing theories. That argument says that existing theories that aim to explain split responses tend to have a particular form. Then it says that if a theory of that form is correct, we should not expect to find that the split responses are mostly due to different people having different intuitions. Instead, we should expect to find that individual people have conflicting intuitions.

The key question now is whether that prediction is in fact correct. In what follows, I review all empirical evidence I have been able to identify that bears directly on this question. Different evidence bears on the question in quite different ways, and for obvious reasons, none of the studies reviewed here are explicitly described as attempts to test the conflicting intuitions hypothesis. As a result, all conclusions derived from this review will necessarily be provisional. Not only is it possible that future research will provide further empirical evidence, it is also very possible that some existing studies have provided evidence that bears on this question in ways I simply failed to see.

3.1. *Case study*

To begin with, let's turn to our usual case study. Why do people show split responses when they are forced to choose between textualist and purposivist answers? One possible view is that it is primarily a matter of different people having different intuitions; another view would be that it is primarily a matter of each individual person having conflicting intuitions. Existing research does not conclusively settle the question as to which of these views is correct, but I will argue that the results of recent studies do provide evidence that favors the second view over the first.

We can begin by exploring a simple question regarding the pattern of responses across participants. Do we find that some participants give exclusively textualist responses while others give exclusively purposivist responses? Or do we find that the majority of participants give some textualist responses and some purposivist responses?

A recent study by Flanagan and colleagues (2023) is very nicely designed to address this problem. Each participant received four different cases. Two were the sorts of cases we've been discussing so far, in which a person violates the text of a rule but does not violate the purpose (these cases are known in the literature as "overinclusion cases"). The other two were cases with the opposite structure, where the person does not violate the text but does violate the purpose ("underinclusion cases").

Thus, the purely textualist pattern of responses would be to say that the rule was violated in both of the overinclusion cases and neither of the underinclusion cases. Conversely, the purely purposivist pattern would be to say that the rule was violated in both of the overinclusion cases and neither of the underinclusion cases. We can now ask what proportion of participants give those pure patterns of responses and what proportion give a mix of textualist and purposivist responses.

Figure 5 shows the distribution of responses. In this figure, the third bar from the left shows the percentage of participants who gave purely textualist responses, while the third bar from the right shows the percentage who gave purely purposivist responses. Clearly, the overwhelming majority of participants are not showing either of these patterns. Most participants are showing a mix of textualist responses and purposivist responses.

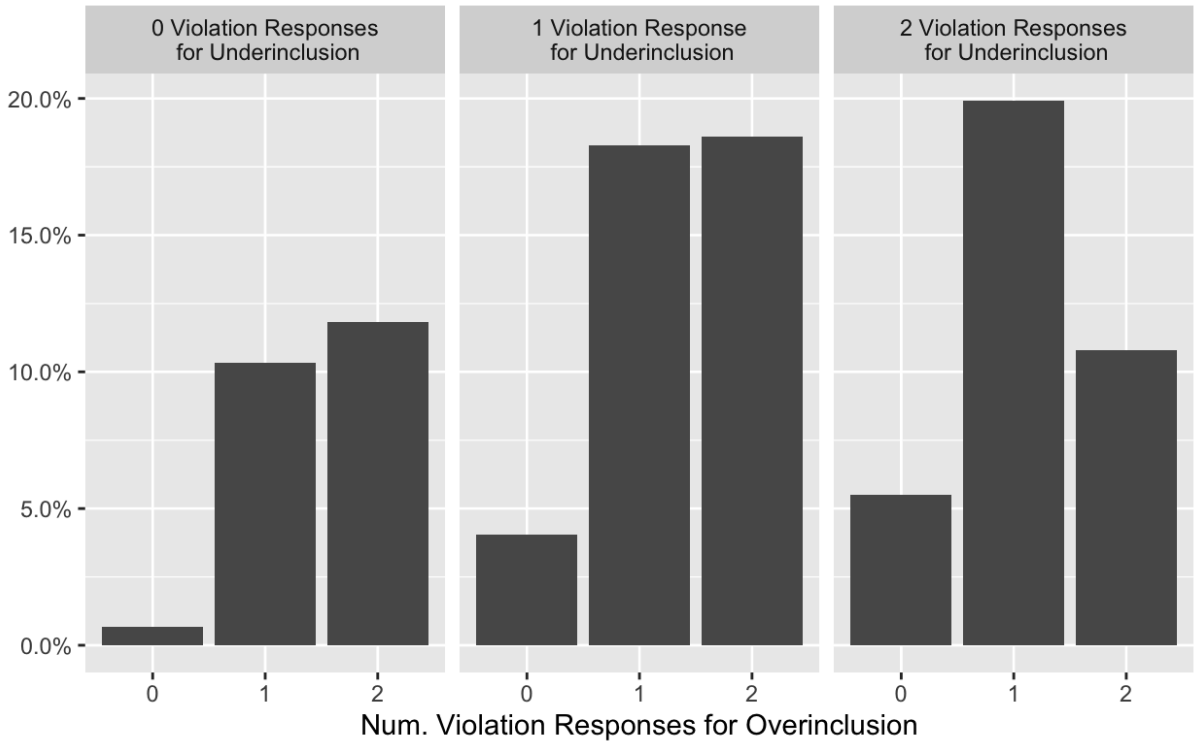

Figure 5. Percentage of participants giving each possible combination of responses on underinclusion cases and overinclusion cases. Based on data from Flanagan et al. (2023).

In short, existing research shows that people sometimes give textualist responses and sometimes give purposivist responses. However, this is not because certain people, in certain

situations, give almost exclusively textualist responses, while other people, in other situations, give almost exclusively purposivist responses. Rather, what is happening is that each individual person in each individual situation is giving a mix of textualist and purposivist responses.

But there is clearly also more to the story. Even though each individual person tends to give a mix of textualist responses and purposivist responses, the results also show that people tend to give *more* textualist responses than purposivist responses. Moreover, this tendency is heightened in participants who have legal training (Hannikainen et al., 2022). That is, participants with legal training show an even greater proportion of textualist responses. Clearly, then, it would not be enough just to say that people have both textualist and purposivist intuitions. We need some understanding of the processes within people's minds that explain their tendency to give more textualist responses.

One possible approach would be to say that different people have different intuitions. For example, one might think that a given individual has, say, probability .65 of having a textualist intuition on any given case and probability. .35 of having a purposivist intuition. Then this individual might have a mix of textualist and purposivist intuitions across a range of cases, but on any individual case, she would either have purely textualist intuitions or purely purposivist intuitions. The difference between different populations (e.g., lawyers vs. laypeople) could then be understood as a difference in the relevant probabilities. Although it is seem possible in principle that this approach will turn out to be correct, I am not aware of any research that aims to spell it out in detail and put it to the test.

Another possible approach would be to explain the phenomenon by positing a psychological process that does not require different people to have different intuitions. On this second approach, there is a psychological process that leads people to give more textualist responses, but this process does not require people to have only the textualist intuition and not also the purposivist intuition. Instead, even if a person had both textualist and purposivist intuitions, this process would lead them to give more textualist responses. Within the existing literature, researchers have developed a specific hypothesis along these lines and tested it experimentally (Hannikainen et al. 2022).

At the core of that hypothesis is the game-theoretic notion of *coordination*. Each individual person wants the law to be interpreted in a particular way, but in addition, a key feature of the way people think about laws is that each person wants to coordinate with others on a shared interpretation. Thus, when I am asked how to interpret the law, my response will be affected in part by a guess about how you will interpret the law. But this immediately creates a complex recursive problem. After all, your interpretation will depend in part on your guess about how I will interpret the law. In cases with this structure existing theories suggest that people tend to coordinate on salient points called *focal points* (also called 'Schelling points'; Schelling, 1980).

This framework allows us to provide an explanation for people's responses. Suppose each participant has both textualist and purposivist intuitions. Now suppose that each participant also has a preference to coordinate with other participants. Suppose further that the textualist response serves as a focal point. Then, to the extent that participants want to successfully coordinate with each other, they will tend to give textualist responses.

To test this hypothesis, Hannikainen et al. (2022) conducted an ingenious experiment. Participants were given cases where textualist and purposivist judgments might be thought to diverge. Then they were told that another participant had also received that exact same case and that they could win money if they gave the same response that the other participant gave. Strikingly, this manipulation shifted participants toward a greater use of textualist responses.

Again, I recognize that this evidence is far from conclusive, and I would be very open to seeing whether it is possible to find evidence on the opposite side. But just looking at the research that is available right now, we are finding some support for a broader psychological theory that would predict conflicting intuitions (Almeida, in press; Almeida et al. in press), some evidence for the claim that each individual participant has both textualist and purposivist intuitions (Flanagan et al., 2023), and at least the beginnings of a theory that would explain why people might show a tendency to give a higher proportion of textualist responses even if they have both textualist and purposivist intuitions (Hannikainen et al., 2022). Taken together, this evidence provides at least preliminary support for a conflicting intuitions explanation in this case study.

3.2. Methods that allow ambivalent responses

One common way of running an experimental philosophy study is to present participants with a case and then ask them whether they think the correct answer is A or B. In such studies, there is simply no way for participants to indicate that they have both an A intuition and a B intuition. However, a small number of studies have used methods that do allow participants to express both intuitions. These studies provide further evidence as to whether participants have conflicting intuitions.

In studies about free will judgments, one common approach is to ask participants to choose between two options, where one option corresponds to compatibilism and the other to incompatibilism. This sort of measure makes it impossible for participants to express conflicting intuitions. Participants can express agreement with compatibilism, or with incompatibilism, but they cannot express agreement with both compatibilism and incompatibilism.

In an innovative twist on this usual approach, Deery and colleagues (2014) conducted a study in which participants were asked independently whether they agreed with compatibilism and whether they agreed with incompatibilism. Each participant received a questionnaire that included a large number of separate statements. Participants were asked to rate their level of agreement or disagreement with each statement. Included within this larger set of statements were four statements that expressed what philosophers call "sourcehood compatibilism" and four that expressed "sourcehood incompatibilism."

Here is one of the statements that expressed sourcehood compatibilism:

> As long as I decide what to do on the basis of my own values, that's enough by itself for me to be the ultimate source of my decisions; in other words, that's enough for my actions to be "up to me."

And here is one that expressed sourcehood incompatibilism:

> Even when I decide what to do on the basis of my own values, that's not enough for me to be the ultimate source of my decisions; I must also have had the final say about what my values were in the first place.

Before moving onward, let me just report my own reactions to these statements. Reading through each statement, my intuitive reaction is that each of them seems true. In other words, I find myself having conflicting intuitions. If I were a participant in this experiment and received both statements, I presumably would not have indicated agreement with both – but that fact about my responses would not directly reflect my intuitions. Instead, it would reflect a further process in which I recognized that my intuitions were contradictory and tried to develop a more consistent position.

The authors made their data available, and I put together a figure to show the pattern of participants' responses. To do this, we can give each participant a score for the mean of the ratings that the participant gave to the four sourcehood compatibilism statements and a separate score for the ratings that participant gave to the four sourcehood compatibilism statements. Figure 6 displays the relationship between those two scores.

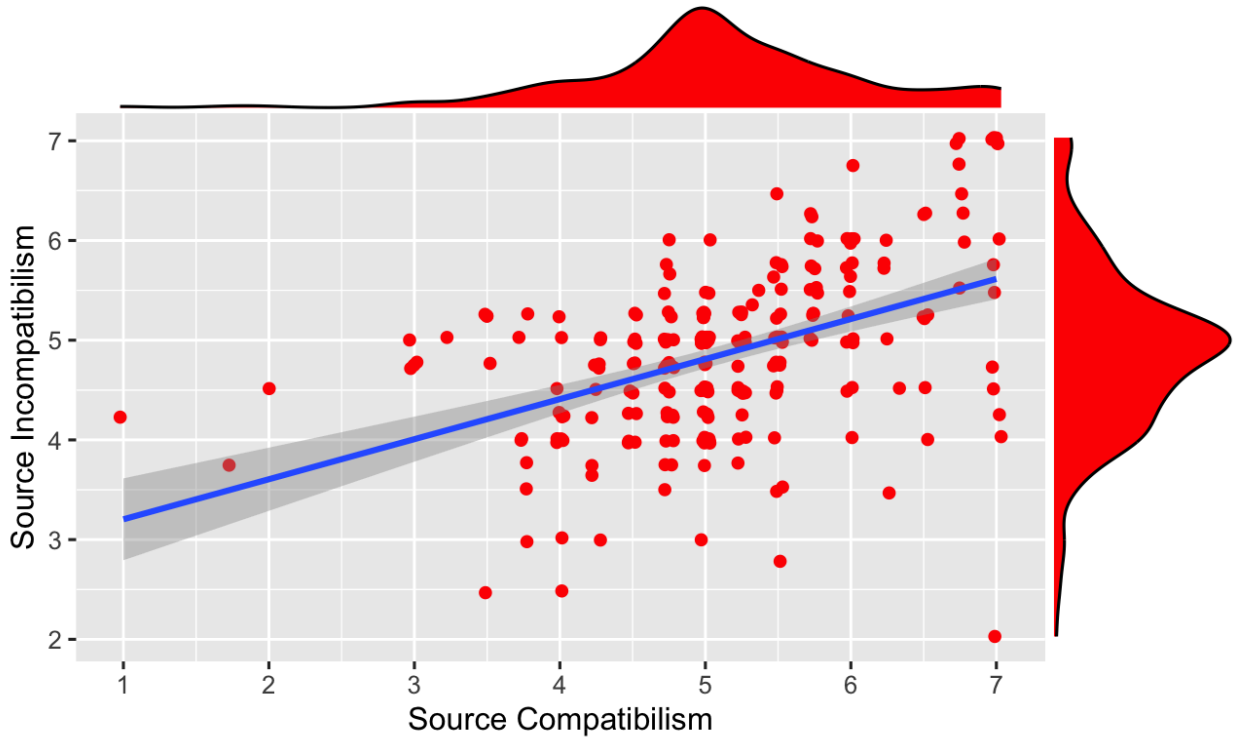

Figure 6. Jittered plot showing the distribution of scores for sourcehood compatibilism and sourcehood incompatibilism in Deery et al. (2015).

As the figure shows, the majority of participants agreed with *both* sourcehood compatibilism *and* sourcehood incompatibilism (and indeed, agreement with these two positions was positively correlated). This result suggests that the disagreement and confusion one finds when it comes to

this question is not the result of different people having different intuitions. It seems instead to be the result of individual people having conflicting intuitions.

For a second example, let's turn to intuitions about the Ship of Theseus problem. The basic problem is simple: Imagine that the Ship of Theseus experiences some wear and tear, and some of the planks have to be replaced. Each time the owner replaces one of the planks, she puts the original plank in a storeroom. After many years of this, none of the original planks are left – all of them have been replaced by new ones. Now imagine that the owner goes into her storage room, finds all of the original planks, and puts those planks together to build a ship. In that case, which ship is the Ship of Theseus – the one created by gradually replacing the original planks or the one created out of the original planks?

Rose and colleagues (2020) conducted a large-scale cross-cultural study to examine people's judgments about this problem. In this study, participants were introduced to the story of the ship (this time called 'Drifter') and then told to imagine a dialogue between two people who had opposite opinions about which answer is correct.

> Suzy and Andy disagree on which of the two rowboats is actually Drifter. Andy thinks that the rowboat just built a month ago is actually Drifter since it has exactly the same planks, arranged in exactly the same way as Drifter originally had. But Suzy thinks that the rowboat that resulted from gradually replacing the original planks used to build a boat thirty years ago is actually Drifter since, even though it has all new parts, this was just the result of normal maintenance.

They were then asked to choose between the following two options:

> [**Replacement**] I agree with Suzy that Drifter is the rowboat that resulted from gradually replacing the original planks used to build a boat thirty years ago and that now has none of its original planks.
>
> [**Original Parts**] I agree with Andy that Drifter is the rowboat built a month ago with the planks and plans that were used thirty years ago.

Overall, the results indicated that there were split responses. The study was run in 25 different samples from 22 different countries and in almost all of those, there was a substantial proportion of participants choosing replacement and also a substantial proportion choosing original parts. For example, in 13 of the samples, the proportion of participants selecting Replacement was in the 60%-70% range, leaving a substantial minority (in the 30-40% range) selecting Original Parts.

A question arises as to how to understand these split responses. Is it that different participants have different intuitions? Or is it that participants tend to have conflicting intuitions? Existing theoretical work on intuitions about persistence over time provides at least some reason to expect that people might have conflicting intuitions in these sorts of cases (Dranseika et al., in press), and the question now is whether they actually do.

To address this question, Dranseika (in press) introduced a new method that allowed participants to more explicitly indicate their ambivalence. In this new method, participants got

basically the same case as in the original Rose et al. study, but now they could choose from four different options.

> [**Replacement**] It only makes sense to say that the ship repaired with new parts is Theseus.
>
> [**Original parts**] It only makes sense to say that the ship built from old parts is Theseus.
>
> [**Both**] It makes sense to say that the ship repaired with new parts is Theseus, but it also makes sense to say that the ship built from old parts is Theseus.
>
> [**Neither**] It does not make sense to say that the ship repaired with new parts is Theseus, and it also does not make sense to say that the ship built from old parts is Theseus.

Results are displayed in Figure 7. As the figure shows, when participants are given the opportunity to express ambivalence, the majority of participants choose the option "Both."

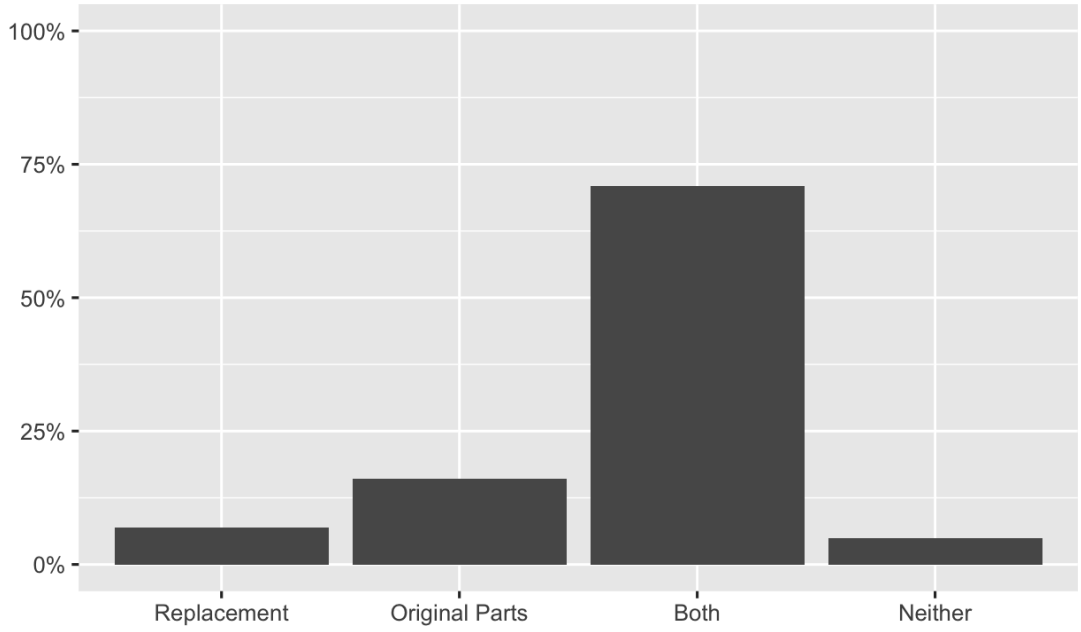

Figure 7. Barchart showing results from Dranseika (in press), Experiment 5.

Dranseika provides transcripts of some of the justifications participants gave for their responses. Participants giving these justifications explicitly state that they feel drawn in one direction and also drawn in the exact opposite direction.

> I completely get both sides, on the one hand, the one built from the old planks is physically drifter - everything's literally identical, and it uses the parts from it. On the other hand, the maintained one has lived the life of John's boat and could be sentimentally considered as Drifter, with all the experiences and events it has been through. (M, 21)

> It really depends on whether the boat itself as a whole or the materials count as the boat. I remain unsure which should truly count but can totally see both points of view so remain undecided. (F, 36)

So, here again, we are getting evidence that the confusion is not a matter of different people having different intuitions but rather a matter of individual people having conflicting intuitions.

We have been looking in detail at just two specific cases in which researchers used methods that allowed ambivalent responding, but such methods have also been applied to a number of other philosophical questions. For example, Fischer and colleagues (2023) look at intuitions about the debate between direct realism and indirect realism in the philosophy of perception. There too, they find evidence of conflicting intuitions. Many individual participants have both direct realist and indirect realist intuitions.

3.3 Stability across groups

A broad array of studies have looked at the way people from different groups respond to philosophical questions of the type we have been exploring here. Such research has looked at patterns across different demographic groups (culture, gender, age) and also patterns across individual difference variables (personality, cognitive style). Drawing on the framework we've been developing here, we can now ask whether this research indicates that people from different groups have different intuitions.

There has been a lively debate about this topic within recent research (Alexander & Weinberg, in press; Knobe 2021; Stich & Machery 2022), but in what follows, I will not be defending any of the claims I made within that existing debate. Rather, I will be exploring the ways in which the framework we have been developing here might shed new light on some of these questions.

Let's begin by distinguishing two types of results. In some studies, people from different groups give very similar responses, while in others, people from different groups give different patterns of responses. We can explore each type of result separately.

First, consider studies in which people of different groups show very similar patterns of responses. To illustrate, Figure 8 displays the pattern of judgments about free will and determinism across world regions. As the figure shows, people in each world region tend to show split responses. But it's more than that: the pattern of responses is actually remarkably similar across world regions. In a previous paper, I described this sort of pattern of results by saying that it indicates that intuitions are *stable* across groups (Knobe 2021). I reviewed numerous different studies pointing to this sort of stability and suggested that it might have implications for various issues that will not concern us here.

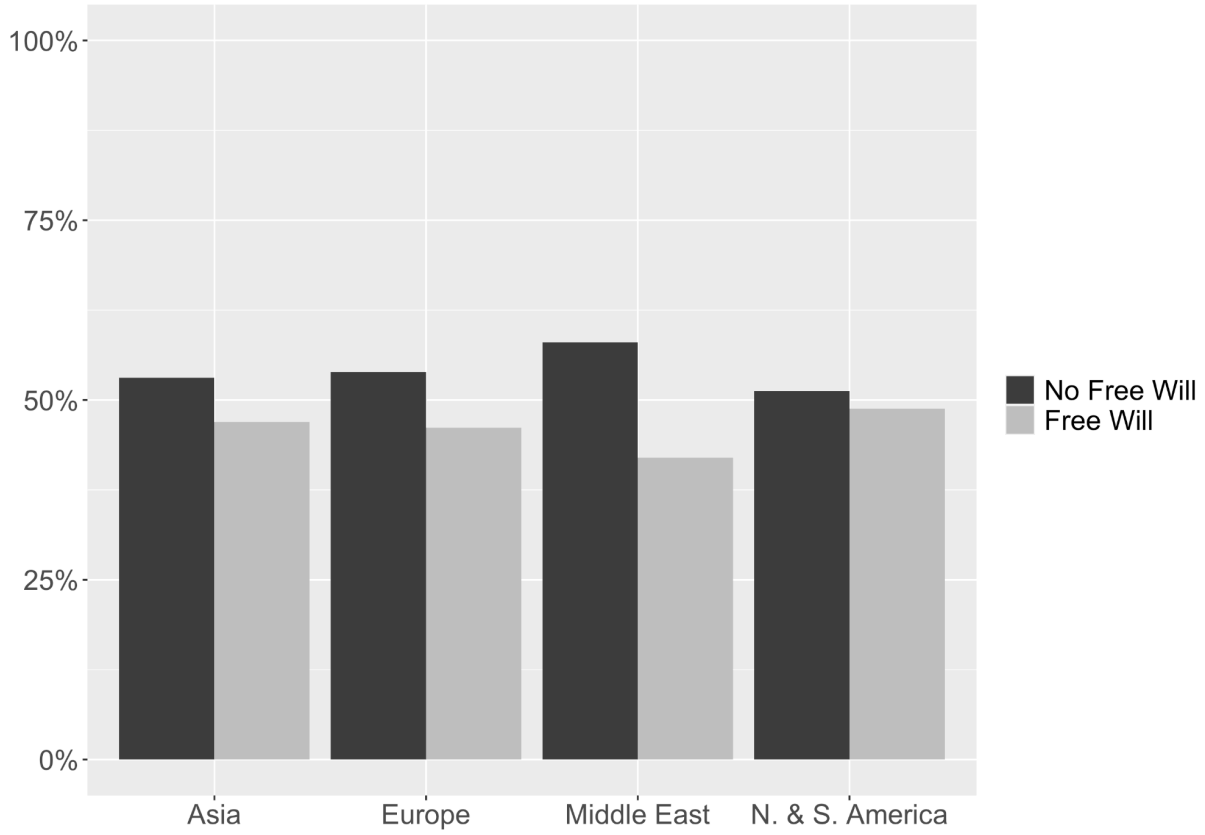

Figure 8. Judgments about whether agents in a deterministic universe can have free will, broken down by world region. Based on raw data from Hannikainen et al. (2019).

We now face a question about what it even means to say that intuitions are stable across groups. Jonathan Weinberg (personal correspondence) points out that there is something fundamentally dissatisfying about the way I explained this claim in previous work. For example, the result in Figure 8 is concerned with one particular variable (culture). If we look just at that one variable, what we find is that the proportion of people giving each response does not differ very much across groups. One possible way of making sense of this result would be to claim that there is a more fundamental pattern such that different groups do not differ in their responses to this question. But even a moment of thought shows that this interpretation makes no real sense. After all, there must be *something* that is making some participants give one response and others give the opposite response. So all the study shows is that we haven't yet identified the factors that are explaining the difference. People's responses might be stable across the particular variable we happened to look at in this study, but once we find the variables that do explain the difference, we can say that people's responses are not stable across *those* variables. This objection is well-taken, and I agree that what I said in previous work was not right.

The present framework makes it possible to understand stability across groups in a different way. The key point is that there are two different possible hypotheses about what is happening

within each group, and these different hypotheses then yield very different understandings of the similarities we observe between groups. In thinking about what is happening within each group, one hypothesis would be that different people within each group have different intuitions. Within each group, some people have one intuition, while others have the exact opposite intuition. If you think that this hypothesis is correct, you will immediately be drawn to a particular understanding of the similarities between groups. That understanding says: Within each group, different people have completely different intuitions, but the proportion of people who have each of the different intuitions ends up being almost exactly the same across different groups.

The conflicting intuitions hypothesis says something very different about what is happening within a single group. For example, if we look just at studies of people from Western culture, the conflicting intuitions hypothesis might say that different people from Western culture don't have radically different intuitions from each other. Instead, they have very similar intuitions (namely, conflicting intuitions). Now suppose we observe that people in other cultures show almost exactly the same pattern of response to the patterns observed within Western culture. The conflicting intuitions hypothesis yields a very different understanding of that similarity. It would say: If people from that other culture show the same pattern of response, the best guess would be that people from that other culture tend to have the very same conflicting intuitions.

Depending on which of these hypotheses turns out to be true, we will get very different answers to our question about the sense in which intuitions are stable across groups. If the first hypothesis turns out to be true, then the result is just that intuitions don't differ very much across the particular groups we happen to be looking at. It would always be possible, at least in principle, to divide people into groups in some other way such that intuitions do differ across those other groups. By contrast, if the conflicting intuitions hypothesis turns out to be true, there is a deeper sense in which intuitions are stable. Despite the fact that half the people give one response and half give the other, people really do share the very same intuitions.

Let's now turn to cases in which the proportion of participants giving each response differs from one group to the next. There is a great deal of debate in the existing literature about the extent to which this actually happens. Researchers on one side of the debate say that experimental philosophy findings point to a pervasive tendency whereby all sorts of different intuitions differ between demographic groups (e.g., Stich & Machery 2022), while researchers on the other side say that early studies that seemed to indicate large differences between demographic groups have mostly failed to replicate and that the main finding coming out of more recent research is that people from different demographic groups tend to show extremely similar patterns of response (e.g., Knobe 2021, 2023). Let's now put this debate to one side. Regardless of which of these views is correct, it is clear that there are at least some cases in which different groups do show different patterns of response, and we want to know how to understand those differences.

Just to get a start on this question, we can consider a few concrete ways in which research has clearly shown differences in responses between groups. Here are some examples:

[**Cognitive reflection and free will**] There is an effect of cognitive style such that people who receive high scores on the cognitive reflection test (CRT) are more likely to give incompatibilist

responses on questions about the relationship between free will and determinism (Hannikainen et al. 2019).

[**Gender and sacrificial dilemmas**] There is a gender difference such that men are more likely than women to give consequentialist responses in sacrificial dilemmas (Friesdorf et al. 2015).

[**Culture and the textualism/purposivism debate**] There is a cross-cultural difference such that people from Poland and Lithuania are more likely to give textualist responses (Hannikainen et al. 2022).

A question now arises about how to understand these findings. What are they teaching us about whether intuitions are stable across groups?

Here again, the present framework suggests that there are two different possible hypotheses. One hypothesis would be that the differences between groups are primarily a matter of people from different groups having different intuitions. Another would be that people in different groups tend to have the same conflicting intuitions, and the difference then arises because people in different groups choose different responses when faced with this conflict.

In deciding between these hypotheses, the best way to proceed is to begin by looking separately at each individual case. For each individual case, we can ask what existing work suggests about how to explain the differences. For the three specific differences used as examples above, I think it's fair to say that existing work suggests that they are not best explained in terms of differences in intuition across groups, but rather in terms of people having conflicting intuitions.

[**Cognitive reflection and free will**] The cognitive reflection test is intended to measure something about what people tend to do when they experience a tension between two opposing processes. Thus, the most natural explanation of existing findings would not be that people with high CRT scores lack compatibilist intuitions. It would instead be that people with high CRT scores respond in a different way when they experience a conflict between compatibilist and incompatibilist intuitions.

[**Gender and sacrificial dilemmas**] Existing research consistently points away from the idea that participants who give consequentialist responses in sacrificial dilemmas do not have deontological intuitions and suggests instead that such participants experience a conflict between two different processes pulling them in two opposing directions (e.g., Greene et al. 2004). Thus, the most natural explanation of this gender difference would not be that men tend to lack deontological intuitions. It would be that men tend to respond differently when they experience a conflict between deontological and consequentialist intuitions.

[**Culture and the textualism/purposivism debate**] As we saw above, existing theoretical research gives us a framework to explain why certain people might be especially inclined to give textualist responses even if they have conflicting intuitions (Hannikainen et al. 2022). Thus, a natural way to understand this cross-cultural difference would be that it is not a cross-cultural difference in the

degree to which people have purposivist intuitions at all but rather a cross-cultural difference in the processes people use to choose a response when they have conflicting intuitions.

Of course, these claims remain tentative and provisional. Further work could potentially overturn existing theories in any of these areas. Moreover, I have been discussing only three group differences, and further work could show that things work out very differently when it comes to other differences.

Still, these points do seem to establish something. If we simply take existing theories in each of these areas and apply them in the most straightforward way, we do *not* get the conclusion that the difference in responses is due to a difference in intuition. Thus, if someone wants to argue that what we are seeing here is a difference of intuition, that person should provide some positive argument either against these existing theories or against the idea that they should be applied to these effects in this straightforward way.

3.4. Cases of different intuitions?

We have been reviewing evidence that seems to support the conflicting intuition hypothesis, but even if this hypothesis does turn out to be true in most cases, it must surely be possible to find at least *some* cases in which different people genuinely do have different intuitions. To properly assess the evidence for and against the conflicting intuition hypothesis, we need to be actively looking for cases in which we have reason to expect that it will not be true. Then we can get a better sense of the larger pattern as to when people have conflicting intuitions and when different people have different intuitions.

One obvious place to look would be in cases where people will only have a particular intuition if they already have a certain body of background knowledge. For example, some philosophical thought experiments rely on basic knowledge of scientific facts (knowledge about genetics, knowledge about neuroscience, etc.). People who do not have the relevant background knowledge might genuinely have different intuition from people who do have that knowledge.

In a nice illustration of this phenomenon, Protzko and colleagues (2023) looked at intuitions about cases in which two people's brains are swapped. The key question was whether a person's obligations follow that person's body or that person's brain. Interestingly, responses to this question were correlated with education: those participants who had lower levels of education were more inclined to say that obligations follow the body. Of course, it might be possible to explain this result in terms of conflicting intuitions, but at a minimum, it does seem that one natural explanation would be in terms of different people having different intuitions. Perhaps people with lower levels of education simply lack certain relevant background knowledge about the role of the brain. If this knowledge is a prerequisite for having a particular intuition, it might then be that people with lower levels of education tend not to have that intuition.

Here we have been looking at one particular study to explore one particular process that might lead to differing intuitions, but the larger point does not depend on this specific case. Even if it turns out that the split responses in the brain swap case are not due to differences in background knowledge, it seems likely that there are at least some cases in which people do have differing

intuitions as a result of different background knowledge. And, more importantly, even if it turns out that this whole point about differences in background knowledge is mistaken, it seems almost certain that there are at least some cases in which different people truly do have different intuitions.

3.5. Summarizing the evidence

In this section, we have been reviewing evidence that bears on the conflicting intuitions hypothesis. Here is a quick summary of that evidence:

- Existing opposing process theories seem naturally to predict conflicting intuitions (i.e., they predict conflicting intuitions unless supplemented with ad hoc auxiliary assumptions that seem designed just to prevent that prediction).

- When participants receive multiple items aimed to address the same philosophical debate, we sometimes find that the majority of individual participants give some responses that fit with one position and other responses that fit with the other position.

- In the textualism/purposivism debate, there is evidence for a specific theory according to which textualist responses are not due to people having only textualist intuitions but rather to people preferring the textualist response even when they have both intuitions.

- In some cases, researchers have used designs that allow participants to openly express their ambivalence, and in such cases, participants do seem to explicitly express conflicting intuitions.

- In some cases where we find differences in responses between groups, existing theories seem to point to the idea that these differences are not explained by differences in intuition but are instead explained by differences in what people would ultimately conclude when faced with conflicting intuitions.

Overall, I would characterize the amount of evidence in favor of this hypothesis as moderate. On one hand, most of the evidence that is available now seems to support the conflicting intuitions hypothesis. On the other, the evidence that is available now is a bit limited and unsystematic.

A key task for further research will be to explore this question across a wider range of different cases. There is no mechanical way to do this. For each separate case, researchers simply have to think of different possible theories. Some of these theories will involve people having different intuitions, while others will involve each person having conflicting intuitions. Then we can conduct experiments to test the different theories in each separate case and thereby arrive at a better understanding of the broader pattern.

4. Philosophical implications

We have been concerned thus far with an empirical question about people's intuitions. Let's now ask whether answers to this empirical question might have philosophical implications.

To do this, we can start by imagining that you are in a certain position within your own philosophical inquiry. Suppose, then, that you are in the following position: A number of years ago, you encountered a thought experiment and immediately found yourself having conflicting intuitions. You were gripped by the deeper philosophical questions that this thought experiment seemed to pose, and you have been working on those deeper questions ever since. After years of reflection, you have arrived at a belief about which answer is actually true.

Now suppose you discover an empirical fact: You start looking at research on reactions to this thought experiment among ordinary folks with no philosophical training, and you find that the conflicting intuitions hypothesis holds true in this case. That is, you find that the split responses observed in ordinary folks are not primarily a matter of people having different intuitions but rather a matter of individual people having conflicting intuitions.

The key question is: If you do find this, what implications would it have for your own philosophical work? This is a difficult question, and we will not be able to completely resolve it here. However, we will be exploring three possible ways in which this empirical finding might be thought to be philosophically relevant.

4.1. Implications of differences in responses

One of the most straightforward and salient pieces of information available from experimental philosophy research is the percentage of participants who select each option. A typical experimental philosophy study might report something like this: "The majority of participants (79%) endorsed view A while a minority endorsed view B (21%)." The conflicting intuitions hypothesis seems to suggest something about the philosophical implications of these percentages. What it suggests is that the percentages will generally not have any deeper philosophical implications.

To see this, consider first what we might conclude if we thought that different people have different intuitions. Suppose you have been working on a philosophical question and have arrived at the belief that view A is the right answer, but then you learn something new: studies show that the most experimental participants think that view B is correct. If you interpret this result as suggesting that most other people have different intuition from yours, there would be at least some reason for you to see this result as a challenge to the view you had been developing. For example, you might think that your philosophical inquiry started out with certain intuitions but that you now have evidence that you should be rethinking these starting-points of your inquiry. Difficult questions arise about whether there really are serious grounds for concern here, but at the very least, one can see broadly how the argument is supposed to go.

Now consider what you might make of this same situation if you assume that the conflicting intuition hypothesis is correct. The whole situation then becomes very different. You start out with both an A intuition and a B intuition, then engage in philosophical reflection and arrive at the belief that B is the correct answer. Now you learn that most experimental participants believe that A is the correct answer. However, you don't thereby acquire any reason to rethink the starting-points of your inquiry. On the contrary, even though most experimental participants arrived at a different belief from the one you have, you are assuming that they share your initial intuitions. Just like you, they start out with both an A intuition and a B intuition. The difference is only that, after reflecting on

the question, they arrived at a different belief about the correct answer. But in this respect, they were faced with the exact same philosophical problem that you face, except that they only spent a few seconds reflecting on it. Why would it matter which answer they tend to choose after this brief moment of reflection?

Of course, it is always possible that we will be able to think of some surprising way in which the percentages do end up having philosophical implications even if the conflicting intuition hypothesis is correct. But at first blush, it certainly seems that the conflicting intuitions hypothesis should tend to orient us toward a different aspect of people's responses. If the conflicting intuitions hypothesis is correct, we should be focusing not on the philosophical implications of which answer participants ultimately select but on the philosophical implications of the conflict itself.

4.2. Could it be that both intuitions are true?

What are the philosophical implications of the conflict itself? At least in some cases, one might think that the fact that people have conflicting intuitions should sometimes change our conception of the debate itself. In particular, it could give us at least some reason to prefer a philosophical theory according to which the *two apparently conflicting intuitions are actually both true*. Existing work in philosophy has led to the development of a number of different frameworks for understanding the idea that apparently conflicting claims can both be true (polysemy, contextualism, relativism, etc.). It might be argued that the empirical finding that individual people tend to have both intuitions regarding a particular question provides support for the idea of applying these sorts of frameworks to that question.

Broadly speaking, the argument might go something like this: First, imagine we discovered that some people have A intuitions and others have B intuitions, but almost no one has both A and B intuitions. Then we might conclude that even if different people have different intuitions in certain respects, there is a certain proposition that fits with almost everyone's intuitions, namely, the proposition that A and B are not both true. Thus, to the extent that we want a philosophical view that accords with people's ordinary intuitions, we would want a view that accords with that proposition.

But this seems not to be what we are observing. Instead, we seem to be finding that individual people tend to have both A intuitions and B intuitions. We might therefore arrive at the opposite conclusion. Any view according to A and B cannot both be true would have to go against most people's intuitions. The only kind of view that could accord with most people's intuitions would be a view on which A and B are both true.

But of course, we need to introduce a further distinction. Even if it turns out that people have A intuitions and people have B intuitions, it's not necessarily the case that people have the intuition that A and B are both true. After all, if we assume that these intuitions are generated by distinct processes, it could easily be the case that people have A intuitions, people have B intuitions, and people also have the intuition that there is no possible way for A and B both to be true. So we now face a further empirical question as to what people think, not merely about each of the two separate claims, but about the conjunction of the two of them.

There has already been at least some research on this question. As we have seen, there are certain cases in which people explicitly endorse statements that seem to involve both of the conflicting intuitions (e.g., the Ship of Theseus problem; Dranseika, in press). In addition, there has been a growing interest in the idea that people might think the two intuitions might be correct in two very different senses. It might be that there is a straightforward sense in which A is clearly true but also a deeper sense in which, ultimately, B is true. For example, for the textualism/purposivism question, people seem to think that the textualist answer is true in some sense, while the purposivist answer is true in a deeper sense. Thus, when participants are given the following two statements, presented back to back, they tend to agree with both of them (Almeida et al., 2023):

> In a narrow sense [the agent] violated the rule.
>
> If you think about what it really means to violate the rule, [the agent] did not truly violate the rule.

Similarly, on the question about personal identity, participants tend to agree with the following conjunctive statement (Knobe, 2022):

> There's a sense in which the man after the accident is clearly still Phineas, but ultimately, if you think about what it really means to be Phineas, you'd have to say that he is not truly Phineas at all.

As we continue to explore these questions, I suspect we will find different results for different philosophical issues. For some philosophical issues, we seem to be finding that people think both intuitions can be true, but for other philosophical issues, we will presumably find that people think it is not possible for both to be true.

In any case, the primary question we face here is not an empirical question but a purely philosophical one. Independent of anything that could be settled through further empirical work on ordinary intuitions, we need to know whether there is some way of working out a philosophical theory that makes sense of what people seem to be saying in these cases. For example, within the philosophy of law, we can ask whether there is any way of developing a theory on which it can be right to say something like: there is a sense in which this action clearly violates the law, but there is a deeper sense in which it does not violate the law. Similarly, within the work on personal identity, we can ask whether there is any way of developing a theory on which it can be right to say: there is a sense in which this person is clearly Phineas, but there is also a deeper sense in which he is not truly Phineas.

Here again, I suspect we will arrive at different answers for different philosophical issues. For some, we may find that it is possible to develop such a theory, and philosophical work along these lines might lead to important new insights. For others, we will presumably arrive at the opposite conclusion. We will find that there is simply no possible way for both intuitions to be true.

4.3. Conflicting intuitions and philosophical problems

Finally, it seems that conflicting intuitions can have philosophical implications even in cases where there is no possible way for both intuitions to be true. Suppose you are exploring a philosophical question. There are two plausible views about a particular case, and you have strong reason to think that it cannot turn out that both views are true. As you consider this case, you find yourself having conflicting intuitions. You might feel that the fact that you have these conflicting intuitions itself has some philosophical significance.

Now suppose that researchers engage in a systematic study of ordinary intuitions regarding this case. This study reveals that it's not as though you just happen to have conflicting intuitions about it; rather, there is a robust and widespread tendency whereby most people have conflicting intuitions. The question now is whether this fact about ordinary intuitions would have any philosophical implications.

A very natural reaction would be that it does have philosophical implications. The thought would go something like this: Suppose that when you first found yourself having conflicting intuitions, you were unsure whether those conflicting intuitions had any larger philosophical significance. If you then discover that your conflicting intuitions are widely shared, you should become at least slightly more inclined to think that they do have a larger philosophical significance. In my view, this is a thought worth pursuing, and we need to ask how it might be spelled out in more detail.

For a fully adequate explanation, we would need a clear answer to the larger question as to how conflicting intuitions can have a larger philosophical significance even when we recognize that there is no possible way that both intuitions can be true. We will only be able to touch very briefly on that larger question here, but perhaps even this brief discussion will shed at least some light on issues about the philosophical implications of the empirical results.

As a first step, we can eliminate one approach that seems like a complete non-starter. In certain cases, philosophers suggest that the fact that people have a particular intuition provides some reason to think that the content of this intuition is *true*. Arguments of this type tend to have at least broadly the following form:

1. Intuitively, it seems that *p*.
2. [Various further considerations]
3. Therefore, we have at least some reason to think that *p* is true.

There has been a huge amount of discussion of arguments of this form, and a great deal of debate about whether they can be made to work. For discussion, see Jackson (1998), Sosa (2007), Machery (2017), Stich and Tobia (2016), Cappelen (2012).

Whatever the merits of these arguments in general, I hope it's clear that they will not prove helpful in the present context. We are assuming that you have found that there is no way that both of your intuitions can be true. Thus, if you think that these intuitions can help us to make valuable progress, it can't be that they will prove helpful as part of an argument of this form. We will need to look elsewhere.

It will be helpful, therefore, to turn to a very different activity in which philosophers are sometimes engaged. One salient way in which people can contribute to philosophy is by identifying and formulating *problems*. In such cases, one looks at an area in which it might initially seem that everything is fine, and one argues that there is actually a deep and important problem in this area. Successfully showing that there is a problem is widely regarded as an important contribution in itself. Indeed, if we think about the most valuable contributions that have been made in the history of philosophy, the ones we would most want students to understand, some of these contributions will consist precisely in identifying a problem.

In arguments for the claim that there is a problem, it is very common to say that a particular proposition is intuitive without suggesting that this proposition is true. For example, one might argue:

1. Intuitively, it seems that *p*.
2. Intuitively, it seems that *q*.
3. But it is impossible for *p* and *q* both to be true.
4. Therefore, there is a problem.

There is a great deal to be said about arguments of this form, but for the moment, let's confine ourselves to two very straightforward points.

First, arguments of this form do rely on claims about what is intuitive. The mere fact that it is impossible for *p* and *q* both to be true is not itself sufficient to show that there is a problem. The conclusion only follows if one also assumes that it seems intuitive that *p* and it seems intuitive that *q*.

Second, although the argument relies essentially on a claim about what is intuitive, it does not proceed by claiming that what is intuitive is true. On the contrary, the whole point of the argument is that what seems intuitive in this case cannot be true.

A question now arises as to whether empirical findings about people's intuition have implications for arguments of this form. This is a difficult question, but at the very least, it certainly seems that facts about the relevant intuitions can teach us something about how important the problem actually is. If we find that only a small minority of people from one very specific culture have the relevant intuition, we might see the problem as a relatively minor one, whereas if we find that there is something more fundamental about the whole way human beings understand the world that leads them to have these conflicting intuitions, we might see the problem as more deeply important.

If we do accept this basic approach to understanding the philosophical importance of conflicting intuitions, we immediately arrive at an understanding of a philosophical question for which these empirical findings might be relevant. Suppose you find yourself having conflicting intuitions about a particular thought experiment. You take those conflicting intuitions as a starting point, and after prolonged reflection, you think you have uncovered an important philosophical problem. But now there is a question as to whether you really have successfully identified anything important. One way to dismiss or diminish the problem would be to call into question the status of the relevant intuitions. It might be said that the conflicting intuitions you are experiencing just reflect

some idiosyncratic facts about you in particular. Or it might be said that these intuitions just reflect something about a small minority of people, or the people in a particular demographic group.

The position we have been developing here is on the extreme opposite side. According to this position, the thing that makes these cases so confusing is an opposition between two fundamental processes within human cognition. The confusion is not just a matter of different people having different intuitions but rather a matter of individual people having conflicting intuitions. And those conflicting intuitions seem to be widely shared across different groups.

In short, if the position we have been developing here turns out to be right, people's conflicting intuitions look very hard to dismiss. On just about every dimension, these intuitions have precisely the characteristics that would allow them to play an important role in philosophical inquiry.

References

Alexander, J., & Weinberg, J. (in press). Practices make perfect: On minding methodology when mooting metaphilosophy. *Oxford Studies in Experimental Philosophy*.

Almeida, G., Struchiner, N., & Hannikainen, I. R. (in press). Rule is a dual character concept. *Cognition*.

Almeida, G. (in press). A Dual Character Theory of Law. *Journal of Legal Philosophy*.

Bengson, J. (2013). Experimental attacks on intuitions and answers. *Philosophy and Phenomenological Research*, 86(3), 495-532

Cappelen, H. (2012). *Philosophy without intuitions*. Oxford University Press.

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., ... & Zhou, X. (2021). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*, *12*, 9-44.

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

Daw, N., & Shohamy, D. (2008). The cognitive neuroscience of motivation and learning. *Social Cognition*, 26, 593-620.

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

Dranseika, V. (in press). Two Ships of Theseus. *Synthese*.

Dranseika, V., Nichols, S., & Shoemaker, D. (in press). The identity of what? Pluralism, practical interests, and individuation. *Philosophy and Phenomenological Research*.

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

Flanagan, B., de Almeida, G. F., Struchiner, N., & Hannikainen, I. R. (2023). Moral appraisals guide intuitive legal determinations. *Law and Human Behavior*, 47, 367-383.

Friesdorf, R., Conway, P., & Gawronski, B. (2015). Gender differences in responses to moral dilemmas: A process dissociation analysis. *Personality and Social Psychology Bulletin*, 41, 696-713.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293, 2105-2108.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44, 389-400.

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.

Hannikainen, I. R., Tobia, K. P., de Almeida, G. D. F., Struchiner, N., Kneer, M., Bystranowski, P., ... & Żuradzki, T. (2022). Coordination and expertise foster legal textualism. *Proceedings of the National Academy of Sciences*, 119(44), e2206531119.

Jackson, F. (1998). *From metaphysics to ethics: A defence of conceptual analysis*. Clarendon Press.

Knobe, J. (2021). Philosophical intuitions are surprisingly stable across both demographic groups and situations. *Filozofia Nauki*, 29, 11-76.

Knobe, J. (2023). Difference and robustness in the patterns of philosophical intuition across demographic groups. *Review of Philosophy and Psychology*, 14, 435–455.

Knobe, J., Prasada, S., & Newman, G. E. (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition*, 127, 242-257.

Knobe, J. (2022). Personal identity and dual character concepts. In K. Tobia (Ed.) *Experimental Philosophy of Identity and the Self*, Bloomsbury Academic, 49-70.

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

Morris, A., Phillips, J., Huang, K., & Cushman, F. (2021). Generating options and choosing between them depend on distinct forms of value representation. *Psychological Science*, 32, 1731-1746.

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

Protzko, J., Tobia, K., Strohminger, N., & Schooler, J. W. (2023). Do obligations follow the mind or body? *Cognitive Science*, 47, e13317.

Reuter, K. (2019). Dual character concepts. *Philosophy Compass*, 14(1), e12557.

Rose, D., Machery, E., Stich, S., Alai, M., Angelucci, A., Berniūnas, R., ... & Grinberg, M. (2020). The ship of Theseus puzzle. *Oxford Studies in Experimental Philosophy*, 3, 158-174.

Sarkissian, H., Park, J., Tien, D., Wright, J. C., & Knobe, J. (2011). Folk moral relativism. *Mind & Language*, *26*, 482-505.

Schelling, T. C. (1980). *The strategy of conflict*. Harvard University Press.

Sosa, E. (2007). Experimental philosophy and philosophical intuition. *Philosophical Studies*, 132, 99-107.

Sousa, P., Allard, A., Piazza, J., & Goodwin, G. P. (2021). Folk moral objectivism: The case of harmful actions. *Frontiers in Psychology,* 12, 2776.

Stich, S. P., & Machery, E. (2022). Demographic differences in philosophical intuition: A reply to Joshua Knobe. *Review of Philosophy and Psychology*, 1-34.

Stich, S., & Tobia, K. (2016). Experimental philosophy's challenge to the 'great tradition'. *Analytica: Revista de Filosofia*, 20, 9-40.

Struchiner, N., Hannikainen, I. R., & de Almeida, G. D. F. (2020). An experimental guide to vehicles in the park. *Judgment and Decision Making*, 15, 312-329.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Sytsma, J., & Snater, M. (2023). Consciousness, phenomenal consciousness, and free will. In P. Henne and S. Murray (Eds.). *Advances in Experimental Philosophy of Action*, 13-32.

Tierney, H., Howard, C., Kumar, V., Kvaran, T., & Nichols, S. (2014). How many of us are there? In J. Sytsma (Ed.). *Advances in Experimental Philosophy of Mind*, Bloomsbury Academic, 181-202.

Tierney, H. (2020). The subscript view: A distinct view of distinct selves. *Oxford Studies in Experimental Philosophy*, 3, 126–157.

Tobia, K. P. (2015). Personal identity and the Phineas Gage effect. *Analysis*, 75, 396-405.