Philosophical Intuitions Are Surprisingly Stable Across both Demographic Groups and Situations[1]

Joshua Knobe
Yale University

(In press at a special issue of *Filozofia Nauki*)

In the early days of experimental philosophy, a number of studies seemed to suggest that people's philosophical intuitions were in a certain sense *unstable*. Such studies typically used one of two broad approaches.

First, some studies looked at the impact of demographic factors. In these studies, all participants received the same question, but different participants belonged to different demographic groups. The results appeared to show that people from different demographic groups had radically different patterns of philosophical intuition (e.g., Machery, Mallon, Nichols & Stich, 2004; Weinberg, Nichols & Stich, 2001).

Second, some studies manipulated features of the situation. In these studies, all participants received the same question, but researchers manipulated something about the situation participants were in while answering. The results appeared to show that people in different situations had very different patterns of philosophical intuition (e.g., Swain, Alexander, & Weinberg, 2008; Wheatley & Haidt, 2005).

These early studies attracted an enormous amount of attention in the philosophical community, and they inspired a sustained effort to explore the philosophical implications of instability. The core question guiding this research has been: “If we learn that people's intuitions are unstable, what should we conclude about the use of intuitions in philosophy?” Attempts to answer this question have shown truly impressive levels of sophistication and ingenuity.

In more recent years, however, there has been an explosion of new empirical research about philosophical intuitions. These more recent studies offer a very different picture of people's philosophical intuitions. The evidence now suggests that *philosophical intuitions are surprisingly stable*. Indeed, the available evidence suggests that philosophical intuitions are surprisingly stable across both demographic groups and situations.

The entire aim of the present paper is to review evidence for this one claim. Of course, if the claim does turn out to be true, it immediately leaves us with some deeper theoretical questions, but I will not be defending any view about those deeper questions here. The reason is not that such questions lie outside the scope of the paper. It is simply that I have not been able to come up with any good answers.

---

[1] For comments on a previous version, I am deeply grateful to Mario Attie, Bob Barnard, Joshua Alexander, Jacob Busch, Ross Colebrook, Matteo Colombo, Noah van Dongen, Ivar Hannikainen, Joachim Horvath, Zach Irving, Dan Kelly, Markus Kneer, Eddy Nahmias, Sara Praëm, Morgan Thompson, Jonathan Weinberg, Alex Wiegmann, David Yaden, Adrian Ziółkowski, and two anonymous reviewers. These comments led to very substantial revisions in the manuscript.

To begin with, we face an empirical question as to *why* people's intuitions are so stable. My only answer is that I have no idea. I will be discussing a whole series of experiments in which researchers manipulate some factor and find that intuitions are remarkably unaffected. In every single one of the experiments reviewed below, I would have mistakenly predicted that the manipulation would have a large effect on intuitions, and even now that I know the actual results, I am completely confused about how to explain them.

Perhaps more importantly, we face philosophical questions regarding what to conclude in light of these findings about the role of intuitions in philosophy. Here again, I have not been able to make any substantial progress. In what follows, I will be including a few preliminary reflections, but basically, the answer is that I don't know.

My apologies in advance for the inconclusive character of this paper. Like many experimental philosophers, I fully expected that we would find considerable instability both across demographic groups and across situations, and I was completely caught off guard when this instability failed to materialize. I don't have any explanation for the results we have been obtaining, but at this point, I do think we need to start engaging in a serious way with the fact that things are not coming out the way we expected.

1. Ways of Being Wrong

Recent debates regarding the purported instability of intuitions are best understood as one part of a much larger debate, spanning many centuries, about the role of intuitions in philosophy. Before getting into the details about the instability claim in particular, it might therefore be helpful to situate our question within this much larger debate.

Let's start by articulating a broad line of thought that I suspect philosophers of many different stripes would be willing to endorse. Here it is:

> BROAD LINE OF THOUGHT. If you look in detail at people's intuitions about philosophical problems, you will often find that people have different intuitions that are in tension with each other. There is just no possible way that these different intuitions can all be true. Thus, it has to be the case that some of people's philosophical intuitions are wrong.

When I speak here of “tensions” within people's intuitions, I mean to be grouping together a number of different phenomena.

- Sometimes different people have opposing intuitions about the very same question.
- Sometimes a single individual has a set of different intuitions that are mutually inconsistent.
- Sometimes a single individual has different intuitions about the same question on different occasions.

These phenomena are different in certain respects, but there is also an important respect in which they are similar. Specifically, they all involve people having a pattern of intuitions such that it is not possible that all of the intuitions are true.

In my view, the broad line of thought is obviously correct, and it should be common ground among philosophers who might in other ways have radically different views about the role of intuitions. Some philosophers think that we can use intuitions to make progress on philosophical problems; others think that intuitions cannot play this role. This is a very difficult question, and I will not be trying to resolve it here. The point I am making is just that both sides of this debate should agree on the broad line of thought. In other words, if you think that we should be using intuitions in philosophy, you should not defend your view by flat-out denying that there are tensions in people's intuitions. (This is not a plausible defense.) Instead, you should argue that we have methods that allow us to start with these intuitions and gradually work through the relevant philosophical problems in a way that resolves the tensions.

With all of this in the background, let's now consider the claim that philosophical intuitions are unstable. We might formulate this claim roughly as follows:

> INSTABILITY. Philosophical intuitions vary dramatically across demographic groups and across situations. People from certain demographic groups or in certain situations tend to show one pattern of intuitions, while those from other demographic groups or in other situations tend to show a very different pattern of intuitions.

If this instability claim is true, it follows that the broad line of thought has to be true as well. (For example, if people from different cultures have opposing intuitions, then it has to be the case that there are tensions among people's intuitions.) But that is not what makes the instability claim so interesting. After all, we would presumably have concluded that the broad line of thought was true even if we determined that the instability claim was false.

What makes the instability claim so interesting is that it goes beyond the broad line of thought to say something far more provocative and controversial. It doesn't just tell us that there are tensions between people's intuitions; it tells us something very specific about the character of those tensions. In what follows, I will have a lot more to say about the claim of instability and how it differs from other possible ways of understanding the tensions in people's intuitions.

More importantly, I will be arguing against the instability claim and in favor of the claim that people's philosophical intuitions are surprisingly stable. So I will be arguing for the claim:

> STABILITY. Philosophical intuitions are surprisingly stable across both demographic groups and situations. Even when we look at people from very different demographic groups, or people in very different situations, we find surprisingly similar patterns of intuition.

Almost the entirety of the present paper will be taken up just in reviewing empirical evidence for this one claim. Before getting to that evidence, however, I thought it would be helpful to provide a better sense of how it relates to the broad line of thought.

The key thing to notice is that my claim about stability does not in any way call into question the broad line of thought. Again, it is simply true that people's intuitions are sometimes in tension, and none of the evidence I provide below will indicate otherwise. The point I am making is that the instability claim gives us an incorrect picture of the nature of these tensions.

If we conclude that intuitions are surprisingly stable but that the broad line of thought is still on the right track, we end up with a different picture of the nature of the tensions. This picture goes something like this:

> BROAD LINE OF THOUGHT WITH STABILITY. There are indeed tensions between different philosophical intuitions, but these tensions are themselves stable. Even across very different demographic groups and very different situations, one finds extremely similar tensions between philosophical intuitions.

This claim does not bear directly on the overall degree to which people's intuitions are in tension, or the degree to which such intuitions are wrong, but it does leave us with a very different picture of the nature of the tensions and the *way* in which intuitions are going wrong.

All of this has been very abstract. To get a better grip on these issues, it will therefore be helpful to pick out one specific example. Let's focus on a question about free will. Is it possible to have free will in a deterministic universe? Compatibilists say that the answer is 'yes'; incompatibilists say that the answer is 'no.' We want to know whether people's ordinary intuitions about this question are in tension and, if they are, whether the tensions are best understood in terms of instability.

In an important paper that we will be discussing further below, Hannikainen and colleagues (2019) report the results of a massive cross-cultural study. Participants were given a description of a deterministic universe and asked whether it was possible for an agent in this universe to have free will. All raw data from the study were made freely available, and I was therefore able to run some additional analyses.[2] These analyses showed that, overall, participants were approximately evenly split. Around half (47%) gave the compatibilist answer, while the other half (53%) gave the incompatibilist answer.

Clearly, if this is an accurate estimate of the percentages in the population, there is just no way it can turn out that the vast majority of participants are getting the right answer. To arrive at this conclusion, we don't need to look at stability across cultures, situations, or anything else. The point just follows trivially from the percentage of participants giving each response. Regardless of whether the right answer is compatibilism or incompatibilism, we can be sure that approximately half of the participants are getting the wrong answer.

A further question now arises about the stability of people's intuitions in this case. One possible view would be that they are *unstable*. For example, it might be that intuitions vary dramatically from one culture to the next. Perhaps people in certain cultures have overwhelmingly compatibilist intuitions, while people in other cultures have overwhelmingly incompatibilist intuitions. A very different view would be that these intuitions are *stable*. For example, it might be that there is something very fundamental within us that leads us to find the

[2] All R code for the new analyses reported here is available at https://osf.io/rdcw6/

problem of free will deeply confusing. As a result, it might be that people from all sorts of different groups, in all sorts of different situations, will be highly divided when it comes to this problem. Neither of these views calls into question the broad line of thought, but all the same, the two views are different. They give us deeply different pictures of the precise nature of the tensions between people's intuitions.

Similar remarks apply to many other core questions of philosophy. Consider the debate about consequentialism in normative ethics, or about skepticism in epistemology, or about relativism in metaethics. In each case, if you collected a whole bunch of different intuitions from a whole bunch of different people, you would probably find a lot of tensions. The key issue is just about whether claims of instability give us the right picture of these tensions.

If the tensions are not best understood in terms of instability, an obvious question arises as to how they *are* best understood. In the remainder of this section, I discuss two other possible accounts. The hope is that discussing these alternatives in more concrete detail will give us a better sense of what is at stake in the question regarding the instability claim.

1.1. Conflicting intuitions

One possible view would be that the tensions are to be understood in terms of a conflict between different psychological processes within people's minds. On this view, we can explain certain tensions using a hypothesis that has broadly the following form: Process X is drawing people toward one intuition, but Process Y is drawing them toward the exact opposite intuition. The tension in people's intuitions is due to the conflict between these two different processes.

Hypotheses of this type have played an enormously important role in research in experimental philosophy. They have been proposed as explanations for the tensions in people's intuitions about numerous different topics, including free will, intentional action, personal identity and causation (e.g., Alicke et al., 2011; Cushman & Mele, 2008; De Brigard, 2010; Greene et al., 2001; Murray & Nahmias, 2014; Nichols & Knobe, 2007; Tierney et al., 2014).

Yet, though there are lots of papers that explore the application of this approach to one or another specific question, there has been remarkably little work devoted to understanding the nature of the approach itself. It might therefore be helpful to say a few words on a more general level about how the approach usually works, what meta-philosophical questions it raises, and how an explanation that involves conflict differs from one that involves instability.

Experimental philosophy research on conflicting psychological processes usually aims to make both an empirical contribution and a distinctively philosophical contribution. At an empirical level, the goal is to use the methods of cognitive science to understand the actual psychological processes drawing people toward each of the different intuitions. At a more philosophical level, the goal is to use facts about the way these processes work to make positive progress on the original philosophical question. That is, the goal is to leverage facts about the psychological processes underlying people's intuitions about a question to help make positive progress in addressing the question itself.

To illustrate, consider again the question of free will. When you first encounter this question, you might feel confused. Perhaps you find that something is drawing you toward compatibilism, but you also find that something is also drawing you toward incompatibilism. In

the moment you experience this conflict, you presumably have no introspective access to the nature of the psychological processes drawing you in these opposing directions. You don't know what precisely it is that is drawing you toward compatibilism and what is drawing you toward incompatibilism – you simply find yourself having intuitions that pull you in these opposing directions.

Experimental philosophy research on this topic aims to understand in detail the psychological processes that draw people toward compatibilism and those that draw people toward incompatibilism. There is now a very active debate between opposing hypotheses regarding this empirical question. Some researchers argue that people's compatibilist intuitions are rooted in an error or confusion (e.g., Nadelhoffer, Rose, Buckwalter & Nichols, in press); others argue for the opposite position, suggesting that it is actually people's incompatibilist intuitions that are rooted in an error or confusion (e.g., Murray & Nahmias, 2014).

Then, at a more philosophical level, this research aims to use these empirical facts to make positive progress on the question as to whether free will actually is compatible with determinism. For example, suppose you conclude that this empirical research provides evidence for the view that your own compatibilist intuitions are produced by a psychological process that is making a certain kind of error. This might lead you to discount those intuitions, which could make you more inclined to think that incompatibilism is the correct view. Thus, facts about these psychological processes could prove helpful to you in arriving at an answer to a question about the nature of free will.

Difficult meta-philosophical issues arise about whether this sort of argument actually works. At the core of these issues lies the question: “Can we use facts about the psychological processes underlying people's philosophical intuitions to make positive progress in philosophy?” In what follows, I will not be defending any answer to that question. My point is just that if experimental philosophy research uncovers important empirical facts about the conflicts in people's intuitions, this is the meta-philosophical question that most naturally suggests itself.

By contrast, research on instability tends to take a different form. There is not usually any claim that the instability provides positive support for one specific view about the original philosophical issue, and there is not usually any claim that the philosophical implications emerge only from a detailed account of the underlying psychological processes. Instead, the goal of this research is usually a completely different one. Most typically, researchers suggest that the instability provides reason to reject a certain philosophical method, namely, a purely armchair method in which one takes the intuition that p to be evidence for the claim that p.

Thus, suppose we find cross-cultural instability in people's intuitions about the compatibility of free will and determinism. In some cultures, people have overwhelmingly compatibilist intuitions, while in other cultures, people have overwhelmingly incompatibilist intuitions. One would not be naturally drawn toward the idea that this result could provide positive support either for compatibilism or for incompatibilism. However, one might well think that this result could provide reason to be suspicious of armchair methods that rely on intuitions as to whether free will is compatible with determinism.

Again, difficult meta-philosophical issues arise as to whether this argument actually works. Broadly speaking, these issues hinge on the question: “Can instabilities in people's intuitions provide reason to reject certain armchair methods?” I won't be defending any specific

answer to that question either. My point is just that findings of instability don't naturally tend to raise the same questions that are raised by findings of conflict. If we learn that the tensions in people's intuitions are due to conflicting psychological processes, it would be natural to start wrestling with a certain set of meta-philosophical questions, but if we instead learn that these tensions are due to cross-cultural instability, it would be natural not to focus on those questions but instead on a completely different set of questions.

With all of this in the background, we can now introduce the empirical question that will be our focus in what follows. The notion of conflict seems to offer us a way of explaining the tensions in people's intuitions without invoking instability. Suppose we find that people have opposing intuitions when it comes to some philosophical question. One possible explanation would be that the opposing intuitions are due to a conflict between distinct psychological processes. There would then be a further question as to whether this conflict involved any instability. One possible view would be that the workings of the processes varies dramatically either between demographic groups or between situations. In that case, people's intuitions might be unstable. Alternatively, however, it might be that the processes that generate the conflict are rooted in fundamental facts about the human mind and that they arise robustly even across very different demographic groups and situations. In that latter case, there would be a tension between opposing intuitions, but the tension would itself be highly stable.

This is the question we will be exploring in detail below. In cases where there are tensions in people's intuitions, are those tensions unstable across demographic groups and situations? Or are the tensions themselves stable across demographic groups and situations?

1.2. Inconsistent intuitions

A second possible view is that people have *inconsistent intuitions*. Suppose people have the intuition that *p*, and that they also have the intuition that *q*. It might then turn out that *p* and *q* are inconsistent. In that case, at least one of these two intuitions has to be wrong.

Note that inconsistency need not itself involve either instability or conflict. The intuition that *p* might be completely stable across demographic groups and situations, and people might not be in any way conflicted about it. Similarly with the intuition that *q*. The problem does not lie in either of these two intuitions considered separately but rather in the fact that they are inconsistent with each other.

Here again, the problem of free will is a helpful example. Many studies of free will intuitions proceed by giving participants a vignette about a completely deterministic universe and then asking them some questions about this universe. Studies show that people's responses to these questions are sometimes inconsistent. For example, in one study, all participants received a vignette about a deterministic universe called “Universe A” (Nichols & Knobe, 2007). Participants were then randomly assigned to receive an *abstract* question or a *concrete* question. Participants in the abstract condition got the question:

> In Universe A, is it possible for a person to be fully morally responsible for their actions?

Meanwhile, participants in the concrete condition were introduced to a specific individual named Bill who lived in Universe A and told that this specific individual killed his own wife and children. They were then asked:

> Is Bill fully morally responsible for killing his wife and children?

In the abstract condition, most participants said that it was not possible for anyone to be morally responsible for actions performed in Universe A, whereas in the concrete condition, most participants said that Bill was morally responsible for the actions he performed in Universe A. This result provides evidence for the claim that people's intuitions about free will are inconsistent.

The key thing to notice about this study is that it demonstrates an inconsistency without demonstrating any sort of instability. The study looks at intuitions about an abstract principle and intuitions about a concrete case. It then shows that there is an inconsistency between these two intuitions. Of course, one might then ask a further question as to whether each of these intuitions is itself stable across demographic groups and across situations, but that would be a completely separate question – one that is not in any way addressed just by the finding about inconsistency.

In what follows, I will not be contributing anything new to the study of inconsistency. The point is just that we need to distinguish instability from inconsistency so that we can arrive at a better understanding of instability. Instability is importantly different, both philosophically and empirically.

From a philosophical perspective, it is important to note that the problem of inconsistent intuitions is a well-known problem that has been discussed for far longer than philosophers have been using systematic empirical methods. Philosophers have long recognized that there are sometimes inconsistencies between different intuitions, and we already have a rich tradition of thinking about what to do when we identify tensions between different intuitions (e.g., Aristotle, 340 BCE/1999; Rawls, 1971). The tensions observed in the free will study described above – between intuitions about an abstract principle and intuitions about a concrete case – fall very squarely within this existing tradition.

However, most of this more traditional philosophical work was concerned specifically with inconsistency and not with instability. If we find an inconsistency between two intuitions, we can turn to a sophisticated literature within more traditional philosophy for guidance about how to resolve the tension, but if we find that people's intuitions can be shifted around by factors in the external situation, most of this literature would not apply. Instability would present us with a completely different sort of philosophical problem.

From an empirical perspective, the two are importantly different in that they are explored using studies with very different designs. Specifically, in studies designed to test for inconsistency, participants in different conditions receive *different questions*. The aim is then to see whether participants's answers to these different conditions are inconsistent with each other. By contrast, in studies designed to test for instability, all participants receive *exactly the same question*. The aim is to see whether some other variable shifts around people's responses to this question.

For example, in a study designed to test for instability, researchers might expose participants in one condition to a manipulation that makes them feel disgust and participants in another condition that makes them feel calm. Then all participants would receive exactly the same question (with not a single word changing from one condition to the other). The goal is not to see whether people's intuitions about two different questions are inconsistent with each other but rather to see whether people's intuitions about the very same question are unstable.

With this framework in place, we can now formulate more clearly the thesis that I will be defending in the remainder of the paper. I am not saying that people's philosophical intuitions are surprisingly consistent. On the contrary, people have inconsistent intuitions about many philosophical questions, and these inconsistencies are part of what makes philosophical problems so puzzling. Rather, I am saying that people's philosophical intuitions are surprisingly *stable*. If you give different people the exact same question but then vary certain factors in their external situation, those factors end up having surprisingly little impact on philosophical intuitions. To the extent that people's intuitions are inconsistent, the claim I am making entails that the inconsistency should itself be stable.[3]

1.3. Summary

In this section, we have been concerned with three phenomena: unstable intuitions, conflicting intuitions, and inconsistent intuitions. All three involve a tension between different intuitions, and all three would provide evidence that people's intuitions are often wrong. Still, there are important differences between these different phenomena, and it would be a big mistake just to lump them all together and ignore the distinctions between them.

In the remainder of this paper, I will be arguing that people's intuitions are surprisingly stable. This does not mean that there are no tensions between different intuitions, and it does not mean we have no evidence that people's intuitions are often wrong. But if true, it would provide valuable insight into the nature of the tensions and the precise way in which people's intuitions are going wrong.

---

[3] In this connection, it seems important to mention research in experimental philosophy that explores effects from the judgment and decision-making tradition, such as framing effects and effects of irrelevant options (e.g., Sinnott-Armstrong, 2008; Wiegmann, Horvath & Meyer, 2020; but see Demaree-Cotton, 2016). These effects show that people's intuitions are inconsistent. Importantly, however, a growing body of evidence suggests that these judgment and decision-making effects are not merely due to an idiosyncrasy of one particular culture but instead emerge across a wide range of different cultures (e.g., Ruggeri, et al., 2020). Thus, we have some reason to think that the inconsistency is itself stable.

2. Stability across demographic groups

Let's begin with the evidence that philosophical intuitions are surprisingly stable across demographic factors, such as culture, gender and age. This evidence takes two basic forms.

First, in some cases, studies find that people in a particular demographic group generally converge on one specific answer to a particular question. In other words, when we run studies on these people, we find that a strong majority endorse one specific answer, and only a minority choose any other answer. A question now arises as to whether people from other demographic groups will share this same intuition.

Research on this topic has arrived at a surprising conclusion. In numerous cases, it turns out that such intuitions are widely shared across different demographic groups. For example, in cases where people from Western cultures mostly agree on a particular question in metaphysics or epistemology, we often find that people from numerous other cultures share the same intuition (e.g., Cova et al., 2019; De Freitas et al., 2018; Hannikainen et al., 2020; Yuan & Kim, in press; Rose et al., 2019; Sarkissian et al., 2010). This is an important form of evidence, but I have already discussed it in previous papers (Knobe 2019, 2020), and I have nothing new to say about it here.

Second, in some cases, people from a particular demographic group show a lot of *disagreement* about a particular question. When we run studies on people from this demographic group, we find that responses are sharply divided. Some give one answer, while others give exactly the opposite answer. A question now arises about whether this disagreement will itself be stable across demographic groups.

It is this second question that I will be exploring in what follows. The central conclusion will be that, here again, research finds surprising stability. When there is a lot of disagreement within one demographic group, we tend to find a strikingly similar pattern of disagreement in other demographic groups.

Before we get into the details, a quick word about the basic vision. Certain philosophical questions probably strike you as *confusing*. When you consider these questions, you may find yourself feeling uncertain, or torn in competing directions, or you might find that the people around you have different intuitions from yours. The core claim is that this puzzlement is not due to some idiosyncrasy of the particular demographic group to which you belong. It appears to be due to something more fundamental – something that emerges also in people from numerous other demographic groups.

2.1. Studies that find similarities across demographic groups

It might be helpful to begin just by briefly recapping some of the evidence I discussed in previous papers (Knobe 2019, 2020). If you are already familiar with those papers, please feel free to skip the next three paragraphs.

First, some early studies suggested that there might be dramatic differences between demographic groups in their philosophical intuitions about certain cases. However, more recent studies indicate that there is actually a great deal of stability across demographic groups *even*

when it comes to those exact cases. Here is a list of cases in which recent studies seem to overturn earlier claims regarding instability:

- Cultural differences in intuitions about Gettier cases (Kim & Yuan, 2015; Machery et al, 2017a; Machery et al, 2017b; Seyedsayamdost, 2015)
- Cultural differences in intuitions about Truetemp, the cancer conspiracy case, and the zebra case (Seyedsayamdost, 2015)
- Gender differences in intuitions about Gettier cases, compatibilism, dualism, Twin Earth, the violinist case, causal deviance, the trolley problem, the Chinese Room, the Plank of Carneades, and the magistrate and the mob case (Adleberg et al., 2015; Seyedsayamdost, 2015)
- Age differences in intuitions about the fake barn case (Busch, Bergenholtz & Praëm, in press)

Second, in studies on Western participants, experimental philosophers have uncovered numerous striking patterns that one would not have expected to find based on prior philosophical work. Again and again, subsequent studies have revealed that those effects also arise for people from other cultures. Here is a quick list of effects that have been shown to arise across multiple cultures:

- Gettierized beliefs judged to be knowledge when morally bad (Yuan & Kim, in press)
- Certain forms of law judged more actual than possible (Hannikainen et al., 2020)
- Aesthetic values judged non-objective (Cova et al., 2019)
- Incompatibilist intuitions when people consider a deterministic universe in the abstract (Sarkissian et al., 2010)
- Knowledge attributions unaffected by stakes (Rose et al., 2019)
- The impact of similarity between judges on intuitions about moral relativism (Sarkissian et al., 2011)
- Certain mental states judged to be knowledge but not belief (Yuan & Kim, in press)
- Assertion judged to be appropriate in cases where the speaker has justification but the content is false (Kneer., 2021)
- Intentional action attribution impacted by moral judgment (e.g., Lin, Yu, & Zhu, 2019)
- Even within cases involving harm, intentional action attribution impacted by the severity of the harm (Kneer et al., 2021)
- The true self judged to be morally good (De Freitas et al., 2018)
- The impact of improvement vs. deterioration on personal identity judgments (Dranseika et al., unpublished data)
- People with positive emotion but morally bad lives judged to be not truly happy (Yang et al., in press)
- Larger effect of outcome on deserved punishment judgments than on wrongness judgments in within-subject moral luck cases (Kneer et al., in prep. a)
- Larger effect of action/omission distinction on deserved punishment judgments than on wrongness judgments (Kneer et al., in prep. b)

Third, many of the effects observed in experimental philosophy studies also emerge in young children. Here is a list of effects that have also been found in such developmental studies:

- Intuitions about metaethical status of a judgment impacted by degree of consensus (Heiphetz & Young, 2016)
- Intuitions about generics impacted by valence (Tasimi et al., 2017)
- Impact of physical contact on intuitions about moral dilemmas (Pellizzoni et al., 2010)
- Intentional action attribution impacted by moral judgment (Leslie, et al. 2006; Michelin et al., 2010; Pellizzoni et al., 2009; Rakoczy et al., 2015)
- Causal judgments impacted by prescriptive norms (Samland et al., 2016)
- People with positive emotion but morally bad lives judged to be not truly happy (Yang et al., in press)

In short, there is really *a lot* of evidence for stability across demographic groups. It is not just that research has failed to find certain sorts of exciting differences. It is that there is a large body of research suggesting extremely striking and surprising similarities.

With this quick summary complete, we can now turn to the issue that will be most central in the present paper. Some of the questions that appear in experimental philosophy studies are quite difficult or confusing. When you receive one of these questions, you may feel pulled in competing directions, and different people ultimately end up providing different answers. We can now ask whether this confusion and disagreement is itself stable across demographic groups. If there is a whole lot of disagreement with regard to a certain question within one demographic group, will we find that same pattern of disagreement within other demographic groups?

Consider again people's intuitions about free will. As I noted above, Hannikainen and colleagues (2019) conducted a massive cross-cultural study in which participants from numerous different cultures were given a vignette about a deterministic universe and asked whether people in the universe could be free. Figure 1 shows the pattern of responses across cultures.

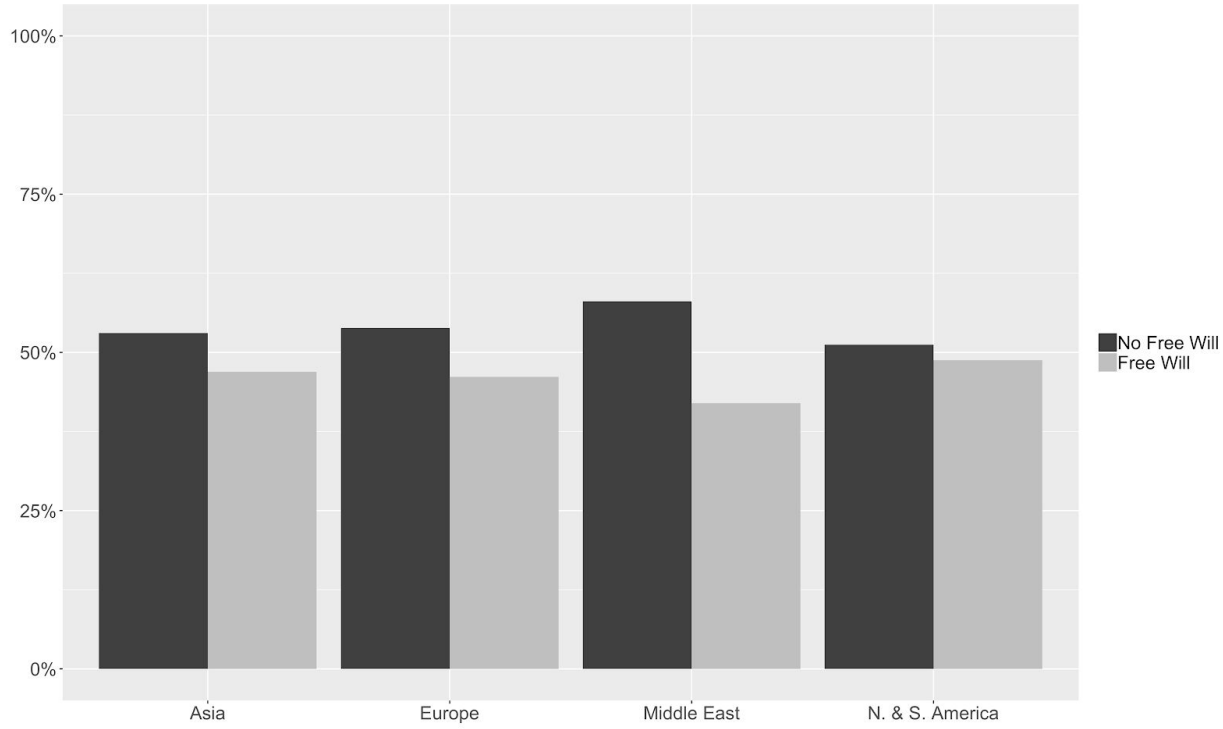

Figure 1. Judgments about whether agents in a deterministic universe can have free will, broken down by world region. Based on the raw data from Hannikainen et al. (2019).

Notice first the result for participants from the Americas. There, we find a whole lot of disagreement. Some participants give compatibilist responses, others give incompatibilist responses. This finding is not itself a surprise. We already knew from previous studies that American participants can be drawn both toward compatibilism and toward incompatibilism, and it was precisely those results that motivated existing theories according to which opposing psychological processes are drawing people in conflicting directions .

But now look at all of the other world regions. Here one finds a result that is truly surprising. The disagreement observed among the American participants appears to be shockingly stable across cultures. In each of the world regions included within the study, one finds the same basic pattern. In fact, there is no significant difference between regions.

Note that this finding is not predicted just by the basic claim that people have conflicting intuitions. Within existing research, there are theories according to which people are drawn toward compatibilism by one type of psychological process and toward incompatibilism by another. But that claim does not, just in itself, predict stability across cultures. For the pattern of people's intuition to be stable across cultures, it would also have to be the case that each of these conflicting processes is found in numerous different cultures, and takes more or less the

same form in each of them. Future research should develop and test specific hypotheses designed to explain this surprising stability.[4]

Similar findings emerge when one looks at the pattern of philosophical intuitions across gender. In an important recent paper, Adleberg and colleagues looked at fourteen different cases in which researchers had specifically suggested that there might be gender differences in people's philosophical intuitions. Famously, the results suggested truly overwhelming stability in philosophical intuition across gender differences. The key point to note about these findings for present purposes is that the stability across gender was not confined just to cases in which participants tended to agree with each other. In some cases, there was a lot of disagreement about what the right answer was, but in those cases, the disagreement itself was stable across gender. The authors made all of their data available, and in what follows, I provide some additional information about their results in cases of this latter type.

One of the questions concerned the problem of free will. This time, participants were given a description of a deterministic universe and then asked whether an agent in that universe would be free to commit murder. As Figure 2 shows, when the questions asked in this way, the percentage of participants giving each answer is a bit different. More importantly, however, this pattern of disagreement is also remarkably stable.

[4] Prior to the advent of systematic cross-cultural studies on free will intuitions, I specifically thought that we would obtain the opposite result. My sense was that American participants were drawn toward incompatibilism by a judgment derived from abstract theoretical cognition and that people from other cultures would not be drawn toward this same judgment. I teamed up with a number of other researchers to explore this question, but our actual results did not support my hypothesis. Instead, the results indicated that the tendency toward incompatibilist intuitions is stable across cultures (Sarkissian et al., 2010). This finding suggests that perhaps there was something fundamentally mistaken in the way I was thinking about people's ordinary incompatibilist intuitions. A good account needs to posit something more fundamental that draws people in numerous different cultures toward incompatibilism (see, e.g., May, 2014).

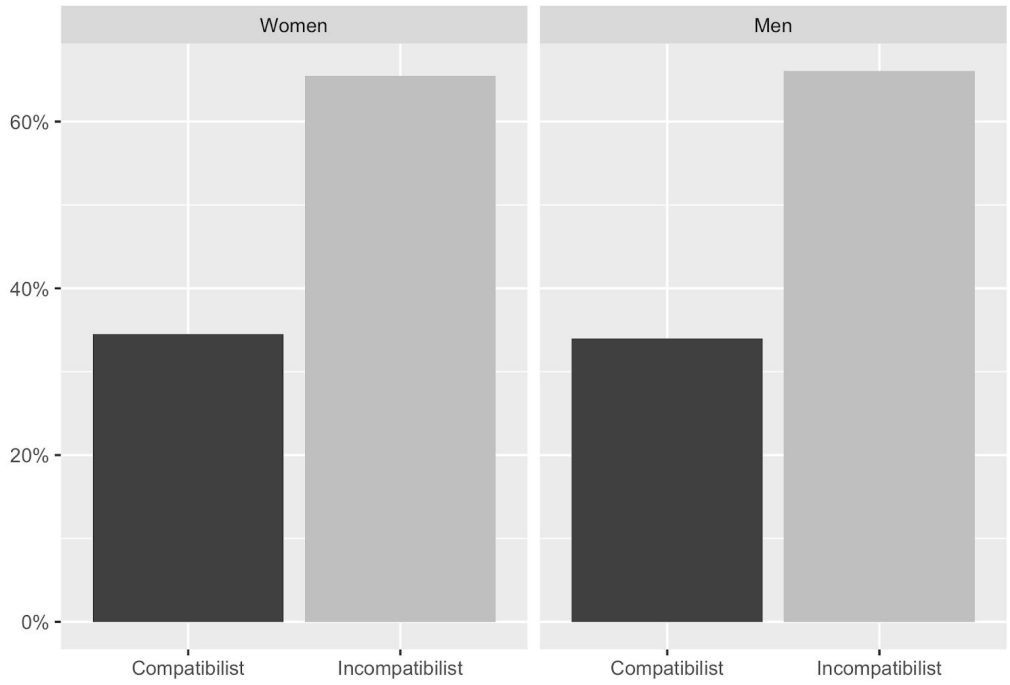

Figure 2. Free will judgments by gender. Based on data from Adleberg et al. (2014).

Another question was about the Plank of Carneades thought experiment. In this thought experiment, a shipwrecked sailor pushes another sailor off of a floating plank so that he can survive. This action causes the death of the other sailor, but it might also be seen as in some sense an act of self-defense. Is the sailor blameworthy for performing this action? On this question, participants did not simply receive a dichotomous yes/no choice but instead gave responses on a scale. Figure 3 shows the distribution of responses broken down by gender. The question is a difficult one, and different participants gave very different answers. But, as the figure shows, the pattern of disagreement was again highly stable across gender.

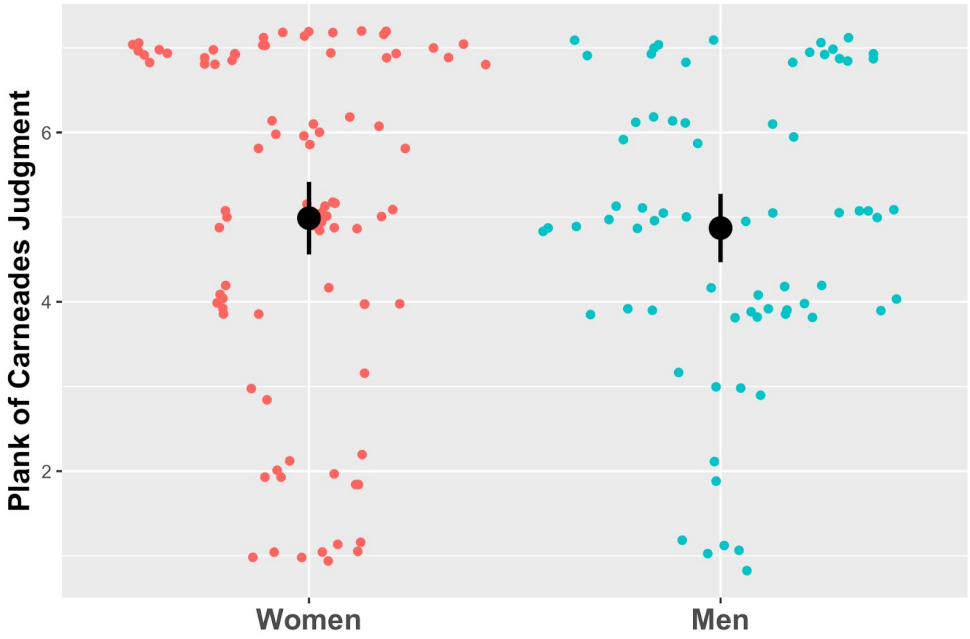

Figure 3. Sinaplot showing responses on the Plank of Carneades thought experiment. Higher numbers indicate higher degrees of blameworthiness. Colored points show the ratings of each individual participant; grey circles show the overall mean for each condition. Error bars show 95% confidence intervals. Based on data from Adleberg et al. (2014).

Now consider the question they asked about Twin Earth. Imagine a person on Earth who utters the word “water.” Now imagine a person on Twin Earth who utters the word “water.” Do these two people mean the same thing, or two different things? Figure 4 shows the distribution of participants' responses. The distribution here is different from the one we saw above for the Plank of Carneades, but here again, one finds that the pattern of disagreement is extremely stable across gender.

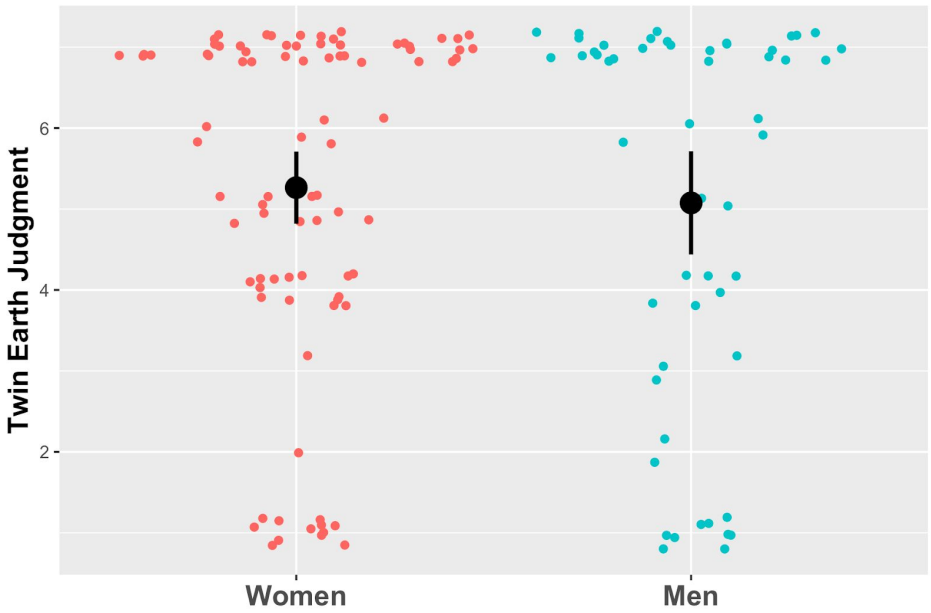

Figure 4. Sinaplot showing responses on the Twin Earth thought experiment. Higher numbers indicate higher degrees of agreement with the claim that when Oscar and Twin-Oscar say 'water,' they mean the same thing. Colored points show the ratings of each individual participant; grey circles show the overall mean for each condition. Error bars show 95% confidence intervals. Based on data from Adleberg et al. (2014).

Finally, we can ask whether philosophical intuitions are stable across age groups. In cases where there are tensions in people‘s intuition, do older people differ from younger people, or are the tensions themselves stable?

Consider intuitions about the fake barn thought experiment. In this thought experiment, a person is looking directly at a building that looks like a barn and forms a belief that it is a barn. As it happens, she is completely right: it is in fact a barn. However, the nearby surroundings are littered with barn facades that look exactly like barns but are not actually barns at all. The question is whether the person *knows* that she is looking at a barn. Some philosophical work argued that there would be an age difference in responses to this thought experiment, and Busch and colleagues therefore conducted an experiment to explore the question. The results are shown in Figure 4. As the figure shows, there is considerable disagreement about whether or not the person knows that she is looking at a barn, but that disagreement is completely stable across age.

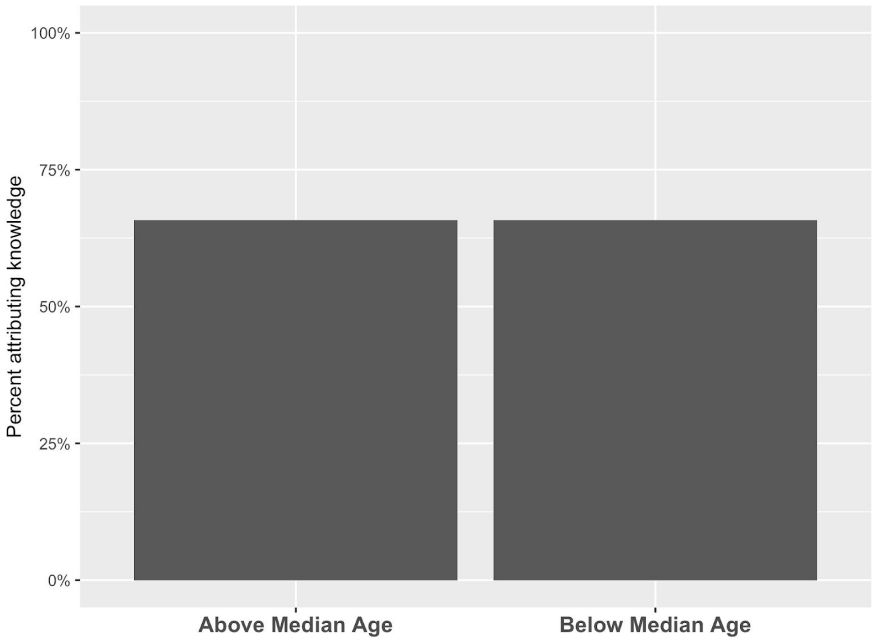

Figure 5. Barchart showing percentage of participants attributing knowledge by age in Busch et al., in press.

Recall what is supposed to be surprising about all of these results. We are looking at cases in which previous philosophical work specifically suggested that there might be a difference between different demographic groups. However, even when we look at those exact cases, what we find is surprising stability across groups. People from different groups all seem to see these same cases as confusing, and in much the same way.

2.2. Studies that find differences between demographic groups

Thus far, I have been focusing on studies that find similarities between the intuitions of participants from different demographic groups, but there are also studies that find differences between participants from different demographic groups. For example, there is a cross-cultural difference in intuitions about Frankfurt examples (Hannikainen et al., 2019), a gender difference in intuitions about brain in a vat cases (Buckwalter & Stich, 2014),[5] and a cross-cultural difference in intuitions about reference (Machery et al., 2004). All of these differences have been observed in rigorous empirical studies, and there is strong reason to believe that these are all real effects.

[5] In informal conversation, I have sometimes heard researchers express skepticism about this finding. My view is that this skepticism is unwarranted. Adleberg et al. (2014) find an effect in the same direction, $t(129.74) = 3.06$, $p = 0.003$. DeRose (2017) asks a similar question using a dichotomous measure and again finds the effect, $\chi 2(1, N = 720) = 13.4$, $p < .001$. Of course, we should always be open to the possibility that future research will overturn any given finding, but in this case, I don't see any particular reason for doubt.

In my previous papers on this topic, the point I made was just that most cases are not like this. For example, Adleberg and colleagues (2014) looked at fourteen different cases in which philosophers had suggested there might be gender differences in intuition. They found a gender difference in *one* of these cases (the brain in a vat case), but they found striking similarity between genders in the other *thirteen*. Thus, even though there are indeed cases in which studies find genuine differences between demographic groups, the overall message coming out of research in this area is how similar the intuitions of people from different groups really are.

In this new paper, I want to look in a little more detail at the results coming out of research that does show differences between demographic groups. I will not be criticizing any of this research. On the contrary, my view is that it has been rigorously conducted and successfully shows exactly what it purports to show – I am a fan. The goal, then, is just to arrive at a better understanding of what this work has actually found.

As a case study, I discuss the stream of research kicked off by Machery and colleagues' (2004) fascinating paper on intuitions about reference. This stream of research has been consistently excellent, and it has taught us something genuinely important about the tensions within people's ordinary intuitions. I worry, however, that philosophers might not have a fully adequate understanding of what this research has shown. In fact, I think what this research has shown is almost exactly the opposite of what many philosophers think it has shown.

A recent paper by van Dongen and colleagues provides a meta-analysis of existing studies on this topic (van Dongen et al., in press). Van Dongen and colleagues made all of their data freely available, and I will be drawing heavily on those data in the analyses that follow.

One of the main findings of the van Dongen et al. meta-analysis is that different studies of this phenomenon are arriving at very different results. More specifically, some of the studies have made use of what are usually called “Gödel vignettes” (first introduced in Kripke, 1972), while other studies have used various other vignettes. It turns out that there is a big difference between the results of studies using the Gödel vignettes and the results of studies using other vignettes.

Gödel vignettes involve a person who receives credit for a discovery he did not actually make. Different studies using this type of vignette have used slightly different materials, but all of them use some variation on the following case (from Machery et al., 2004):

> Suppose that John has learned in college that Gödel is the man who proved an important mathematical theorem, called the incompleteness of arithmetic. John is quite good at mathematics and he can give an accurate statement of the incompleteness theorem, which he attributes to Gödel as the discoverer. But this is the only thing that he has heard about Gödel. Now suppose that Gödel was not the author of this theorem. A man called “Schmidt”, whose body was found in Vienna under mysterious circumstances many years ago, actually did the work in question. His friend Gödel somehow got hold of the manuscript and claimed credit for the work, which was thereafter attributed to Gödel. Thus, he has been known as the man who proved the incompleteness of arithmetic. Most people who have heard the name “Gödel” are like John; the claim that Gödel discovered the incompleteness theorem is the only thing they have ever heard about Gödel.

When John uses the name “Gödel”, is he talking about:
(A) the person who really discovered the incompleteness of arithmetic? or
(B) the person who got hold of the manuscript and claimed credit for the work?

In this case, (A) would be considered the 'descriptivist' response, while (B) would be considered the 'causal-historical' response.

Figure 6 displays the results for studies using the Gödel vignettes among Western participants. Each bar corresponds to one study, and the length of each bar shows the percentage of Western participants in that study who gave the causal-historical response. The black bars show the results from the original study that kicked off this whole research program (Machery et al., 2004). The gray bars show the results from all of the subsequent studies.

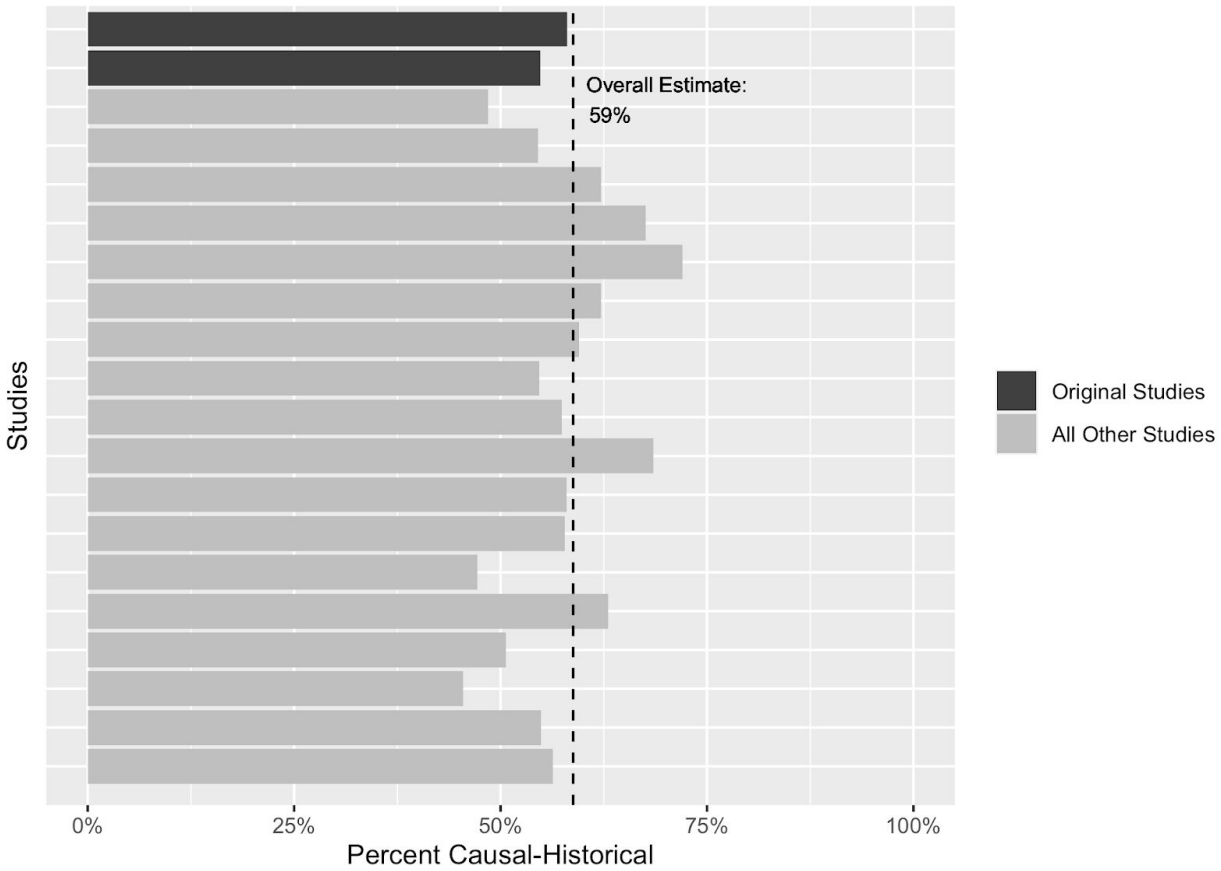

Figure 6. Results from participants from Western cultures in studies on Gödel vignettes. Overall estimate shows the percentage of all participants from Western cultures in these studies who gave the causal-historical response. Based on data from van Dongen et al. (in press).

These results show something truly surprising. Philosophers might well disagree about deeper theoretical issues involving the nature of reference, but there is quite a lot of agreement among philosophers that the correct answer in the Gödel case is the one we are labeling

'causal-historical.' A natural prediction would therefore be that Western participants would pretty much converge on that one answer. Strikingly, that is not what happened. Instead, as the figure shows, existing studies find a large amount of tension in people's intuitions. Overall, a little more than half of participants give the causal-historical response, while a little less than half give the descriptivist response.[6]

From an empirical perspective, this finding leaves us with a difficult question as to how to explain this tension. Why are Western participants so divided in their intuitions about this case? I certainly would not have been able to predict this level of disagreement, and I'm sure many other philosophers are in the same position. We face a difficult question about why participants are responding in the way they are.

Perhaps more importantly, the finding raises some difficult philosophical questions. If we find this much disagreement, what do we thereby learn about the nature of reference itself? There is at least some reason to think that we learn something pretty revolutionary. After all, if there is a single right answer in the Gödel case, it has to be that a large percentage of participants are failing to get that answer. Yet, it might seem implausible to suggest that a large percentage of native speakers could be wrong about such a basic question regarding their own language. So this result seems to provide at least some evidence against the view that there is a single right answer.

A question now arises about the pattern among Asian participants. Do Asian participants also show a large amount of disagreement about this question? Or do they tend to agree with each other? The results for Asian participants are displayed in Figure 7.

[6] It is important to distinguish between what the studies directly show and various further claims one might make about the larger philosophical implications of the findings. What the studies directly show is that if you give participants this exact measure, there is a great deal of variance in their answers. A further claim one might make is that this variance shows that people are actually torn between descriptivism and the causal-historical view. This further claim is a controversial one. (For example, Devitt & Porot, 2018, provide evidence that if you just ask participants to *use* the relevant name, their use of the name is straightforwardly causal-historical.) I do not mean to be taking a position on these larger philosophical issues. Regardless of what one thinks about those issues, a question arises as to whether there is a cross-cultural difference in people's responses to this particular measure, and it is that question that I take up here.

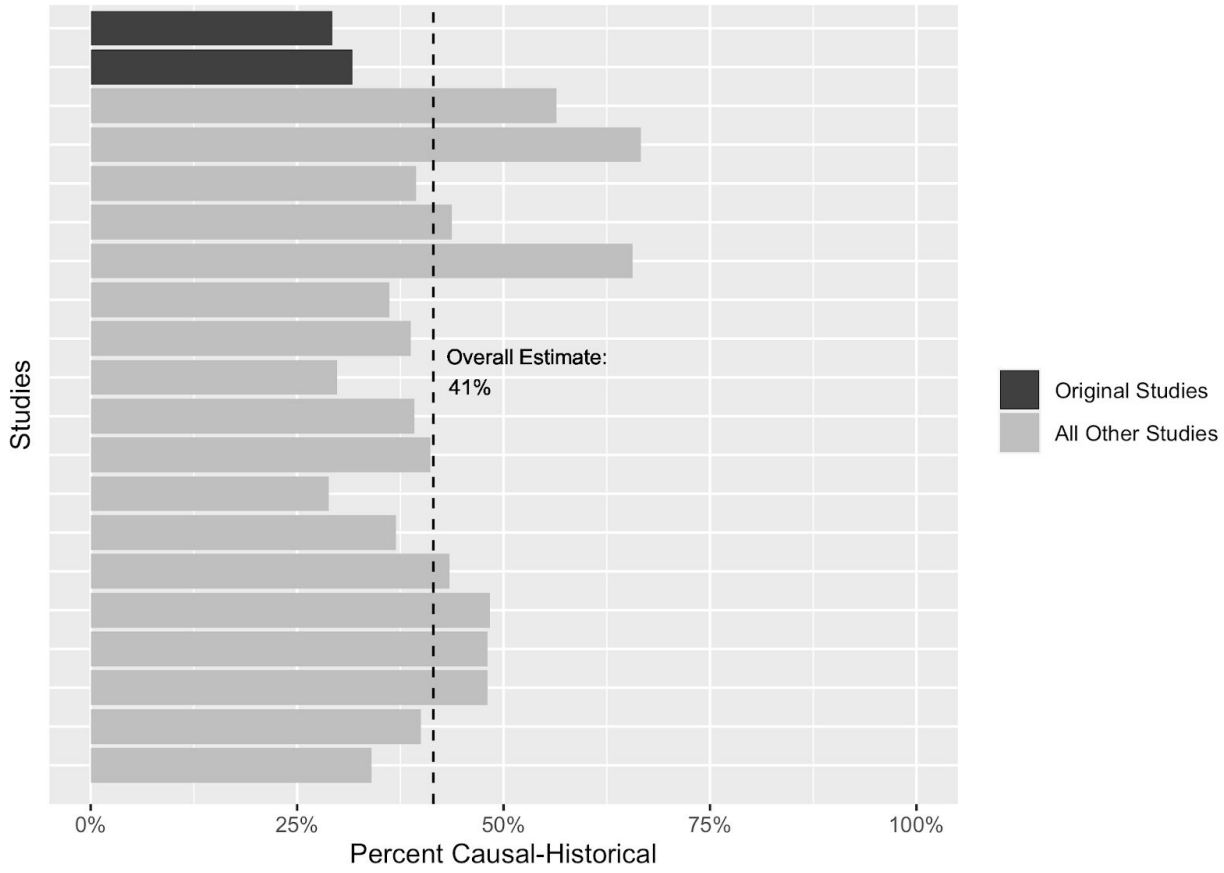

Figure 7. Results from participants from Asian cultures in studies on Gödel vignettes. Overall estimate shows the percentage of all participants from Asian cultures in these studies who gave the causal-historical response. Based on data from van Dongen et al. (2020).

Look first at the results from the original studies (the black bars at the top). These early results seemed to indicate that Asian participants have a completely different pattern of intuitions. While the Western participants show a lot of disagreement amongst themselves, the Asian participants seemed to show at least a fair amount of agreement with each other. A strong majority give the descriptivist response, while only a minority give the causal-historical response.

When these results first came out, it seemed like we had reason to adopt a certain view about the pattern of disagreement. The view went something like this:

> Western participants show a lot of disagreement amongst themselves, but that disagreement is not itself stable across cultures. Rather, Asian participants show a fair amount of agreement amongst themselves.

> Thus, the large amount of disagreement found among Western participants might not reflect any more general truth about intuitions concerning reference. Perhaps it is instead just due to some idiosyncratic fact about Western culture.

Back when these were the only results we had, it really did look like that this view might be right, and philosophers understandably began considering the idea that the cross-cultural difference in intuitions about reference might have some very radical implications.

But now consider the results from all of the subsequent studies. Looking at the full pattern, it is clear that things didn't turn out quite the way we expected. There is indeed a real cross-cultural difference, but it is much smaller than we thought. Participants from Asian cultures are also approximately evenly split, but they lean slightly in the opposite direction. A little more than half of give the descriptivist response, while a little less than half give the causal-historical response.

Taken as a whole, then, the total body of research on this topic gives us a very different picture of the pattern of disagreement. The picture looks more like this:

> There is indeed a certain amount of disagreement between cultures, but the main source of disagreement is the disagreement *within* each culture. Participants from Western cultures show a surprisingly large amount of disagreement among themselves, and participants from Asian cultures also show a large amount of disagreement among themselves.
>
> Thus, the surprising amount of disagreement observed in Western culture *does not* simply reflect an idiosyncrasy of one specific culture. The surprising amount of disagreement is something found across cultures.

This result is extremely interesting in itself. Research with Western participants finds a great deal of disagreement in a place where we would not have expected it. Research with Asian participants shows that this unexpected disagreement also arises in another culture. Overall, then, the results suggest that this unexpected level of disagreement might be a cross-culturally robust aspect of people's responses to this question.

In addition, the studies find a difference between cultures. This, too, is an interesting finding. One possible hypothesis would be that the cross-cultural difference found for this one vignette is part of a far more general cross-cultural difference in intuitions about reference. (Maybe participants from Western cultures show more causal-historical intuitions across the board.) A second hypothesis would be that the difference reflects something about *this specific vignette.* The Gödel vignette involves certain elements that have nothing to do with reference per se (e.g., an issue about getting proper credit for a discovery), and it might be that participants from different cultures are responding differently to those vignette-specific elements.

The best way to answer this question is to look at studies that use other vignettes. Do we also find the very same cross-cultural difference with those vignettes? Or is the difference specific to the Gödel vignette in particular? Figure 8 shows the pattern across vignettes. Each row shows a meta-analytic estimate of the effect size for one type of vignette. Points on the right indicate a big effect such that Western participants are more inclined toward causal-historical

intuitions than are Asian participants. Points on the left indicate an effect in the opposite direction. Points close to the 0 line indicate that participants from the two cultures had similar intuitions.

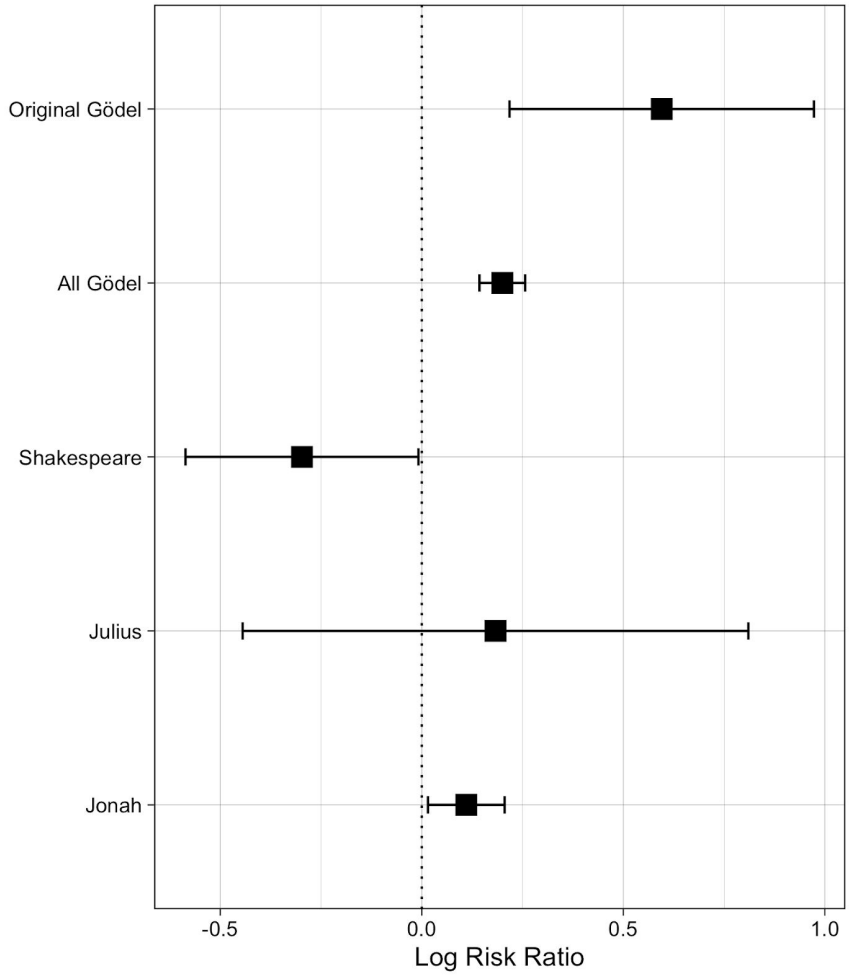

Figure 8. Meta-analytic estimates of effect size by vignette for intuitions about reference. Based on the raw data from van Dongen et al. (in press). Analyses conducted using random effects meta-analysis with the metafor package in R.

The first two rows show the findings we have already been discussing. The earliest study of Gödel vignettes found a large difference between cultures, while the results from existing studies taken as a whole find a much smaller difference.

But now consider all of the other rows. Results from the other vignettes do not simply recapitulate the results found for the Gödel vignette. Instead, they are all over the place. Some show an effect that is in the same direction but substantially smaller; some show no significant effect; some show an effect in the opposite direction. This is itself an interesting and quite striking pattern.

To better understand what is going on here, we can visualize the data in a different way. Instead of looking at meta-analytic estimates of effect size, we can just ignore the distinction between vignettes and look at the raw percentages within each culture when we put all of the data together. Figure 9 shows those results.

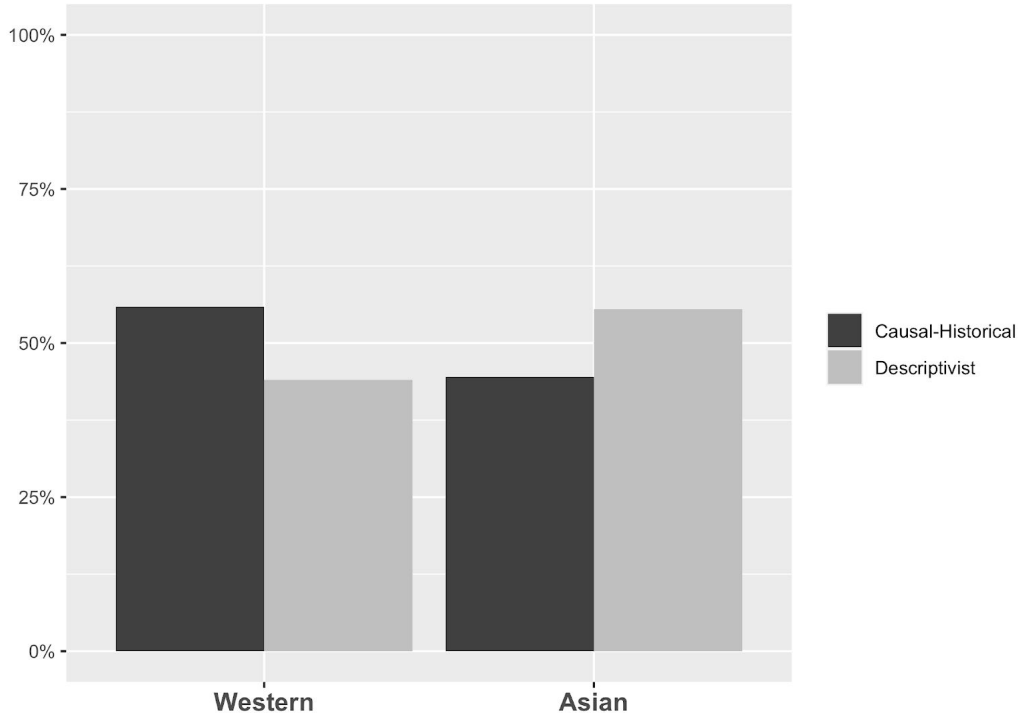

Figure 9. Percent of participants giving causal-historical vs. descriptivist responses by culture, collapsing across vignette. Based on raw data from van Dongen et al. (in press).

First consider the amount of disagreement within each separate culture. Both within Western culture and within Asian culture, these results indicate that the amount of disagreement is *enormous*. For example, within Western culture, 56% of responses are causal-historical, while 48% are descriptivist. If this is an accurate portrayal of the actual pattern of intuitions in the population, then if there is just a single correct answer to this question (either causal-historical or descriptivist), the maximum possible percentage of Western participants who could be getting the correct answer is 56%. It would at least be plausible to claim that this finding should lead us to radically rethink our views about the theory of reference.

What exactly are the philosophical implications of this tension in intuitions within each culture? I am not at all sure. One obvious suggestion would be that the tension found within each culture has exactly the same implications that instability across cultures would have had. In other words, if we take an existing account of the implications of cross-cultural instability (e.g., Mallon, Machery, Nichols, & Stich, 2009), perhaps we can simply repurpose it as an account of the implications of tension within each culture. This suggestion certainly sounds promising.

Another possible suggestion would be that the tension found within each culture is best understood within the framework of the 'conflicting intuitions' research program described above. On that approach, our aim should be to understand the psychological processes within people's minds that draw them toward causal-historical intuitions and the processes that draw them toward descriptivist intuitions. Once we have a better understanding of each of these processes, we might be in a better position to know which intuition is more worthy of our trust. This alternative approach also sounds promising. Future research should continue to explore this question, and to ask which of these approaches, If either, best helps to bring out the philosophical implications of the tension found within each culture

Now compare the results from Western participants to those from Asian participants. Overall, there is indeed a difference, but the difference is *small* (56% causal-historical vs. 45% causal-historical).

This difference is real and certainly worthy of further investigation, but it doesn't seem plausible to say that a difference like this one should lead us to start radically rethinking our whole approach to philosophy. On the contrary, one striking aspect of this result is how stable people's intuitions about reference are across cultures. Across very different cultures, we find that people quite generally find these cases to be confusing – being drawn in part to causal-historical responses, in part to descriptivist responses – and indeed, people in very different cultures are confused by these cases to a roughly equal degree.

2.3. Personality and cognitive style

In their response to my recent papers on stability across demographic differences, Machery and Stich (2020) provide a long list of studies that do seem to show instability. One core theme coming out of this list is that a number of studies find correlations between philosophical intuitions and individual differences in personality and cognitive style. This is an excellent point. Although I have argued that philosophical intuitions are surprisingly stable across demographic groups, I certainly would not also argue that philosophical intuitions are surprisingly stable across differences in personality and cognitive style. These correlations are real, and we need to be thinking about what they mean for larger empirical and philosophical questions.

However, I worry that existing philosophical work on these issues might not be adequately getting at the implications of these findings. Too often, such work just focuses on the mere fact that some philosophical intuitions are predicted by individual difference measures. This is not enough. To engage in a serious way with these difficult philosophical questions, we need to look in more detail at what empirical work on this topic has actually shown.

To begin with, it is not as though the finding is that there is some pervasive phenomenon whereby all sorts of different philosophical intuitions are correlated with all sorts of different individual difference measures. Rather, research has uncovered some very interesting and important correlations, but these correlations emerge only in certain specific cases. In other cases, we are arriving at the opposite finding, namely, that philosophical intuitions are *less* correlated with individual difference measures than one might have expected them to be.

For a particularly striking example, consider the Yaden and Anderson ((in press) study on individual differences among professional philosophers. Philosophers were given a series of philosophical questions (about compatibilism, contextualism, moral realism, physicalism, etc.). Those same philosophers were then given a series of individual difference measures (the five factor model of personality, numeracy, cognitive reflection, satisfaction with life, narcissism, life experiences, etc.). The overall finding was that almost all philosophical views were not predicted by these individual difference measures. There were a few specific exceptions (e.g., subjectivism about aesthetics was predicted by recreational drug use), but for the most part, the result is the one the authors describe in their abstract: “We found limited to no support for the notion that personality or demographics meaningfully predict philosophical views” (Yaden & Anderson, (in press).

In saying this, I don't mean to deny that recent research has made important discoveries about philosophical intuitions and individual differences. The goal is just to get clear about what exactly has been discovered. What we are learning from empirical research is not that philosophical intuitions and individual differences are generally *more* correlated than we might have thought. Rather, what we are learning is something about *which* precise correlations there are. So if we want to understand the philosophical implications of recent discoveries, we are going to have to look in more detail at the actual findings.

A number of studies have looked at the relationship between philosophical intuitions and the degree to which people show a disposition to think more reflectively. Most of these studies make use of the Cognitive Reflection Test (CRT) (Frederick, 2005). The CRT consists of a sequence of questions specifically designed in such a way that the answer that seems intuitively correct is actually incorrect. For example, one of the questions is:

> A baseball bat and a ball cost $1.10 together, and the bat costs $1.00 more than the ball, how much does the ball cost?

It might appear at first that the answer is $.10, but more reflective thought shows that the correct answer is actually $.05. Thus, higher scores on this test indicate a greater tendency to give answers based on more reflective thought.

Studies show that this measure of reflectiveness predicts people's intuitions about a number of different philosophical questions. More reflective people are more likely to have the philosophically orthodox intuition about Gettier cases (Machery et al., 2017), more likely to give the philosophically orthodox response when asked to distinguish between moral and non-moral issues (Colebrook, 2020), more likely to be incompatibilist about free will (Hannikainen et al., 2019), and more likely to give characteristically 'utilitarian' responses to trolley dilemmas (Patil et al., 2020).

For another example, consider intuitions about metaethics. Some participants believe in objective moral truths, while others are drawn toward moral relativism. This difference in metaethical judgments turns out to be predicted by a number of individual difference measures. People who believe in objective moral truths are lower in openness to experience (Feltz & Cokely, 2008), worse at providing empirically adequate explanations for disagreements in belief (Goodwin & Darley, 2008), and higher in willingness to harm other people (Zijlstra, 2019).

For a third example, consider research on moral dilemmas and empathy. When participants are given dilemmas designed to pit consequentialism against deontology, there is a correlation whereby participants who are higher in empathy are more drawn toward deontology. (Thus, participants who are high in psychopathy and in Machiavellianism are more inclined to choose the characteristically consequentialist option; Bartels & Pizarro, 2011.) In subsequent studies, researchers have explored this correlation by developing separate measures of the degree to which participants are drawn to consequentialism and the degree to which they are drawn toward deontology. When these two factors are measured separately, there is no correlation between measures of empathy and consequentialism, but participants who are higher in empathic concern appear to be drawn to a greater degree toward deontology (Conway & Gawronski, 2013).

From an empirical perspective, this research generally looks rock solid. Take the finding that highly reflective people are more likely to give incompatibilist responses. For that test, the sample size was enormous (N = 1,614), and the result is highly significant (p = .0000000009).[7] This is not the sort of finding that is likely to inspire a whole lot of doubt. (As we will see below, the results from studies purporting to show effects of situational factors tend to look very, very different.)

So then, what actually are the philosophical implications of these findings? Within existing philosophical work, one common way of addressing that question has been to lump them in with other findings that purport to show instability. Thus, one might start by asserting that knowledge ascriptions have been shown to be unstable across demographic groups, individual difference variables and situational factors. The question then seems to be about the implications of this instability generally.

The claim I am trying to defend here is that this is the wrong question. It is simply not the case that intuitions have been shown to be generally unstable. For example, it is not the case that knowledge ascriptions have been shown to vary dramatically across demographic groups (see section 2.1), and it is not the case that knowledge ascriptions have been shown to vary dramatically across situations (see section 3.2). In short, empirical research has *not* shown that knowledge ascriptions are generally unstable. What research has shown is that knowledge ascriptions are correlated in certain specific ways with individual difference measures. The real question here is therefore the question about the implications of *these specific findings*.

One approach to answering this question would be to look at the connection between these findings and the research program discussed above on what I called “conflicting intuitions.” Recall the basic idea of this research program. The claim is that people's minds contain different psychological processes that draw them in different directions. Thus, it may happen that one process draws people toward one intuition, while another process draws them toward another intuition. Research within this tradition often suggests that a proper understanding of these different processes will allow us to make positive progress in addressing

[7] The original Hannikainen et al. paper reports a more complex analysis that was designed to address a more sophisticated theoretical question. To answer this simpler question, I used a generalized linear mixed effect model, with CRT as a fixed factor and country as a random factor (random intercepts only). There was a highly significant effect of CRT, OR = .71, 95% CI [.64, .80]. The significance value in the text above was derived by comparing this model to one that did not include CRT.

philosophical problems. More specifically, the usual claim is that if we have a better understanding of the different processes that create the conflict, we will have a better understanding of which intuitions are worthy of our trust and which should simply be dismissed.

Existing findings about philosophical intuitions and individual differences have been very tightly connected with this research program. Many of the studies reviewed above were conducted specifically to test hypotheses about conflicting intuitions, and the others can easily be seen as providing evidence for or against such hypotheses. One obvious view, therefore, would be that the philosophical significance of these findings comes in part from their role in this larger research program. On this view, the significance of these findings is indirect. It is not that there is any significance in itself to the fact that certain philosophical intuitions are correlated with certain individual difference measures. Rather, these correlations give us evidence for hypotheses about the underlying processes that drive people's intuitions, and it is those hypotheses that have philosophical significance.

To Illustrate, consider again the correlation between cognitive reflectiveness and judgments about free will. To visualize the findings, I created a simple plot (see Figure 10). This figure reveals two main findings, closely mirroring what we found for intuitions about reference. First, the disagreement about free will is not primarily a disagreement between more reflective participants and less reflective participants. Rather, we find a great deal of disagreement within each level of reflectiveness. But, second, participants at different levels of reflectiveness show different patterns of intuitions. Among those participants who got the lowest possible CRT score, slightly less than half (42%) gave incompatibilist responses, whereas among those who received the highest possible score, more than half (63%) gave incompatibilist responses.

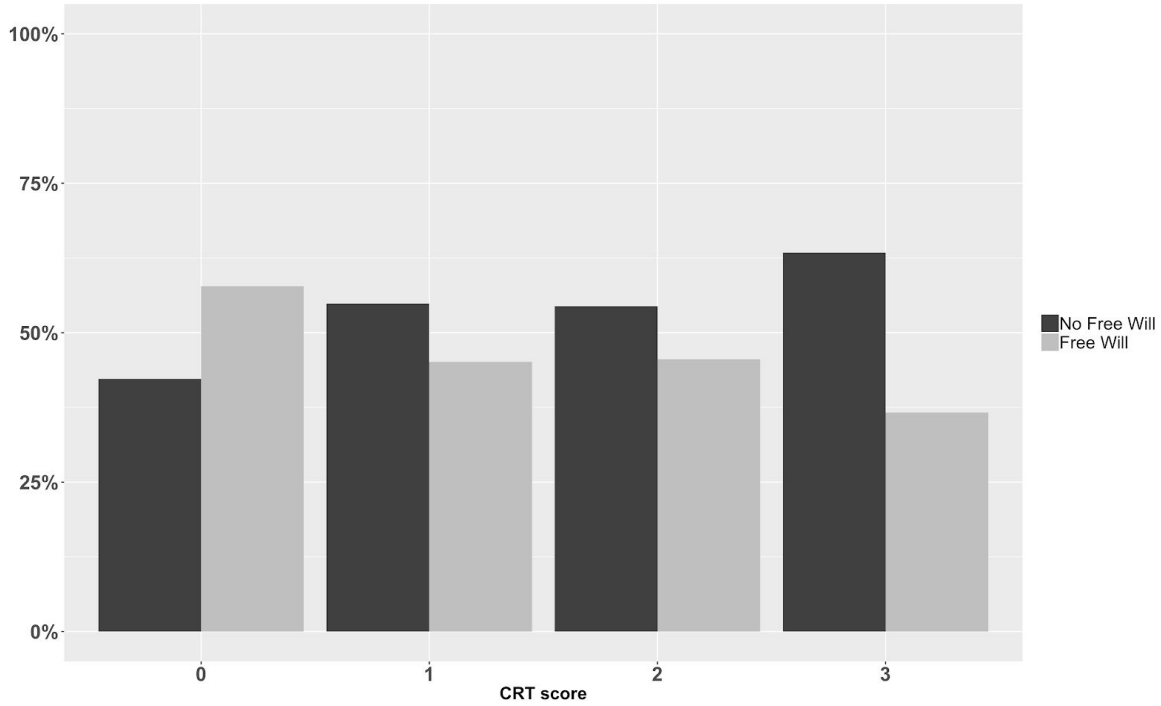

Figure 10. Judgments about whether agents in a deterministic universe can have free will, broken down by CRT score. Based on raw data from Hannikainen et al. (2019).

We can now ask what the philosophical implications of this finding might be. One possible view would be that it provides fuel for the negative claim that intuitions will not help us to resolve the free will debate. To be honest, though, I can't see why anyone would think that this specific finding had that implication. Suppose we already knew that approximately half of all participants had incompatibilist intuitions, but we didn't already know whether the percentage varied by level of reflectiveness. If we now discover that highly reflective people are more inclined toward incompatibilism, how would this discovery give us new evidence for the claim that intuitions are unhelpful in resolving philosophical questions?

An alternative approach – and, I think, a much more natural one – would be to connect this finding with the rich body of existing research on conflicting intuitions. Existing work strongly suggests that there is something within people's minds drawing them toward compatibilist responses but also something drawing them toward incompatibilist responses. However, a great deal of debate remains about how to understand the psychological processes pulling people in these competing directions. Some researchers argue that people's incompatibilist responses are the product of a confusion or misunderstanding (Murray & Nahmias, 2014), while others argue that people's compatibilist responses are the product of a confusion or misunderstanding (Nadelhoffer, Rose, Buckwalter & Nichols, in press). The result we have been discussing here is that people who are higher in a capacity for cognitive reflection are less inclined to give compatibilist responses. Although this result cannot directly settle the debate about which processes involve a misunderstanding, it certainly does seem to provide some valuable evidence.

Of course, if this empirical claim is correct, we immediately face some difficult philosophical questions about how to understand the implications of the conflicting intuitions research program. Is information about the psychological processes that draw us in opposing directions actually valuable in helping us figure out whether or not free will actually is compatible with determinism? For a sophisticated defense of the view that it is, see Murray and Nahmias (2014), but for an equally sophisticated defense of the view that it isn't, see Sommers (2010). This is a complex issue, and I do not have anything further to say about it here. Still, even if we aren't sure whether the argument truly works in the end, we can easily see at least broadly how it is supposed to go. The core claim is that facts about correlations with individual difference measures are indirectly relevant. If we learn that people who are higher in reflectiveness are less likely to be compatibilist, this gives us evidence in favor of a certain hypothesis about the processes that drive compatibilist intuitions, and that hypothesis in turn gives us reason to trust our incompatibilist intuitions more than our compatibilist intuitions.

Much the same could be said about each of the other correlations we have been discussing. The orthodox intuition about Gettier cases is predicted by higher reflectiveness. The intuition that there are objective moral truths is predicted by lower openness and greater willingness to harm others. Deontological intuitions about moral dilemmas are predicted by higher empathy. In each of these cases, it doesn't even seem prima facie plausible to suggest that the correlations give us reason to be more skeptical about the use of intuitions in philosophy. Rather, in each case, the correlation seems at least at first to be giving us information that can help us make positive progress on the relevant philosophical questions themselves. These correlations seem to be giving us reason to trust one intuition over another, and in each case, it is perfectly clear which intuition the results seem to give us reason to trust and which to distrust.

Although this sort of argument does seem at least plausible, difficult philosophical questions arise in each of them about whether the argument can actually be made to work. I am not trying to take a definite stand either way about the answers to these questions. The point I am making is just that these are the questions we need to be exploring.

2.4. Summary

People's philosophical intuitions appear to be surprisingly stable across demographic groups. When people in one demographic group have a particular intuition, we often find that people from other demographic groups share that same intuition. But that is not all. Conflict and confusion seem also to be widely shared across demographic groups. When people in one demographic group are conflicted and confused about a philosophical question, we often find that people from other demographic groups show that same conflict and confusion.

Existing research has uncovered certain specific cases in which there genuinely is a demographic difference in philosophical intuition. These demographic effects are real and worth discussing. However, even in those very cases, we sometimes find evidence for surprising levels of stability across demographic factors.

Finally, although research finds a surprising degree of stability across demographic groups, it is not the case that research shows a surprising degree of stability across individual

differences, such as differences in personality and cognitive style. There are relationships between philosophical intuition and individual differences, and it seems highly plausible that these relationships have important philosophical implications. Nonetheless, it would be a mistake just to lump research on individual differences in with research on demographic effects. The two appear to be very different, and the relationship between philosophical intuition and individual differences should therefore be taken up as a distinct topic in its own right.

3. Stability across situational factors

Let's turn now to studies on the influence of external situational factors. In such studies, researchers give all participants exactly the same case and exactly the same question, but they manipulate some factor in the external situation. Results from these studies have revealed something extremely surprising. People's philosophical intuitions seem to be remarkably unaffected by situational factors.

Much of the relevant evidence here comes from *replication studies*, and it might therefore be helpful to start with a few words in general about this sort of evidence. Many of the cases we will be exploring follow the same basic pattern: An initial study seemed to find an effect of situational factors. This initial study attracted substantial interest, both philosophically and empirically. Because of this interest, a subsequent study was run using almost exactly the same methods but with a far larger sample size. This replication study then failed to find the effect obtained in the original study. Overall, then, the replication study provides evidence that a certain situational factor does not have the sort of influence we previously thought it did.

It is important to note that the problem arising in these cases is not simply a general problem that holds across all experimental philosophy studies but is rather a problem that arises specifically for certain sorts of experimental philosophy studies and not others. Most experimental philosophy studies do not involve manipulations of external situational factors but instead involve manipulations of something in the actual words of the cases participants receive. A recent replication project tried replicating a whole set of studies of this form and found that *90%* of them successfully replicated (Cova et al., 2018). By contrast, our concern in this section is with studies in which researchers do not manipulate anything in the actual text of the case or question participants receive but instead manipulate some external situational factor. As we will see in a moment, the fate of these studies has been radically different. Of the studies of this type that have been tested in replication studies, *almost none* have successfully replicated .

Clearly, this pattern in the study of philosophical intuitions is closely related to a much broader pattern in contemporary psychology research. Over the past ten years or so, there has been something of a revolution in work on the influence of situational factors. Around a decade ago, the journals were filled with studies about how even the most subtle, seemingly innocuous factors could have enormous effects on people's judgments. Now the tide is turning. Again and again, these studies are failing to replicate. To give just a few examples:

- Exposing people to a subtle image of $100 bills does not make them indicate greater support for capitalism (Rohrer, Pashler & Harris, 2015).

- Showing people a picture of Rodin's sculpture *The Thinker* does not make them temporarily more inclined to endorse atheism (Sanchez, Sundermeier, Gray & Calin-Jageman, 2017).

- Exposing people to hostility-related stimuli does not make them more inclined subsequently to interpret ambiguous behaviors as more hostile (McCarthy et al., 2018).

- Making people put their lips into the shape of a smile does not make them rate jokes as funnier (Wagenmakers et al., 2016).

- Reminding people of their own mortality does not make them more inclined to engage in worldview defense (Klein et al., 2020).

Here again, this pattern seems to arise only for certain kinds of studies and not for others. For example, an attempt to replicate studies in personality psychology found that 87% successfully replicated (Soto, 2019), whereas an attempt to replicate studies in social psychology found that only 25% successfully replicated (Open Science Collaboration, 2015). There is perhaps much to be gained by connecting the findings I will be reviewing in this section to the larger discussion of replication failure within contemporary psychology.

The core difference lies in the *point* of reviewing these results. Within contemporary psychology, the main point is usually a methodological one. The fact that so many studies are failing to replicate is taken as evidence for the claim that we are doing something seriously wrong. The goal is then to figure out exactly what we are doing wrong so that we can fix it. This sort of claim will play no role at all in the present paper. I am not trying to assert, or even to suggest, that anyone has done anything wrong, and the question as to whether anyone has done anything wrong is completely irrelevant to the question we will be investigating.

Instead, the goal is to use evidence from replication failures to investigate a question about the degree to which people's intuitions are influenced by situational factors. In all of the cases I will be discussing, the researchers who conducted the original experiments had very good reason to think that situational factors would influence people's judgments, and in each of these cases, I myself would have predicted an effect of the situational manipulation. But, as we will see, in almost every case, the original effect failed to replicate. This pattern seems to be revealing something important about people's philosophical intuitions. It suggests that people's philosophical intuitions are surprisingly stable across changes in the external situation.

3.1. Moral judgment and disgust

A variety of recent studies suggest that the emotion of *disgust* plays an important role in people's moral judgments (e.g., Haidt, Koller & Dias, 1993; Inbar, Pizarro, & Bloom, 2009; Rozin, Lowery, Imada, & Haidt, 1999). Much of the key evidence here comes from correlations between intuitions and individual difference measures. For example, there is a correlation whereby people who are higher in the dispositional tendency to feel disgust are more inclined to

see gay marriage as morally wrong (Inbar et al., 2009).[8] A question now arises about the larger philosophical significance of these findings.

Some philosophers have suggested that information about the role of disgust can help us to make positive progress in thinking about moral questions (Kelly, 2011; Nussbaum, 2009). Thus, suppose you find yourself feeling conflicted about whether or not gay marriage is morally wrong. There is something in your mind giving you the intuition that it is wrong, but also something giving you the intuition that it is not wrong. Now suppose you learn something new: You start reading the literature in cognitive science, and you find evidence that the intuition that gay marriage is wrong is driven in part by disgust. One view would be that this discovery should give you reason to put less faith than you otherwise would have in the intuition that gay marriage morally wrong. This might then shift you toward a positive conclusion, namely, that gay marriage is actually not morally wrong. Although the philosophical issues here are quite complex (see, e.g., Kahan, 1999), this does at the very least look like a promising line of argument.

A second possible view would be that information about the role of disgust gives us a reason to suspect that moral judgment will be susceptible to the influence of external situational factors. After all, if the emotion of disgust is playing a role in the way people make moral judgments, it seems natural to guess that external factors that lead people to feel greater amounts of disgust should influence the judgments they make. Thus, if we simply manipulate certain situational factors (the smell in the room, the sounds people are hearing), we should be able to change the degree to which they feel disgusted and thereby change the moral judgments they make.

Initially, attempts to test this hypothesis appeared to be meeting with a great deal of success. Studies seemed to show that people's moral judgments could be altered by hypnotically induced disgust (Wheatley & Haidt, 2005), by having participants imbibe a disgusting drink (Eskine, Kacinik, & Prinz, 2011) and by exposing participants to the smell of a “flatulence spray” (Schnall, Haidt, Clore & Jordan, 2008). There was even evidence that people made different moral judgments if they had an opportunity to wash their hands after being exposed to disgusting stimuli (Schnall, Benton & Harvey, 2008).

These findings sparked considerable interest in the philosophical community. A number of philosophers argued that they provided evidence for the claim that people's moral intuitions are unstable, shifting around depending on the presence or absence of certain external factors (Machery, 2015; Stich & Tobia; 2015; Weinberg, 2009; but see May, 2018).

In what follows, I will be arguing for precisely the opposite view, but before proceeding further, I want to emphasize that nothing I say here should be seen as a criticism of previous philosophical work. At the time that previous work was written, there was a relatively limited amount of empirical evidence available. Philosophers looked at the available evidence and, very reasonably, arrived at certain conclusions. The point is just that we now have additional

[8] Some of my own previous work claimed to show a similar effect, but that claim turned out to be false. We looked at the relationship between individual differences in disgust sensitivity and intentional action intuitions. It appeared at first that there was a significant relationship (Inbar, Pizarro, Knobe & Bloom, 2009), but subsequent work has shown that this individual difference measure does not predict intentional action intuitions in the way we claimed it did (Klein et al., 2018).

evidence that wasn't available at that time. In light of this additional evidence, I am arguing that we now have reason to adopt exactly the opposite conclusion from the one that many philosophers defended in previous work.

The first doubts about these effects emerged from a widely discussed replication failure of the hand washing studies. In the original study, all participants watched a disgusting scene from the movie 'Trainspotting' involving a person falling into a toilet. Participants were then randomly assigned either to a control condition or to a cleanliness condition in which they had an opportunity to wash their hands. In the original study, this manipulation produced a substantial effect on people's moral judgments (d = -0.85) (Schnall et al., 2008). By contrast, in the replication, this effect disappeared (d = 0.01) (Johnson et al., 2014). In other words, the replication study results indicated that handwashing did not impact people's moral judgments, at least in this experimental paradigm.

To further explore these issues, Landy and Goodwin (2015) conducted a meta-analysis of all available studies of the impact of incidental disgust on moral judgment. This meta-analysis looked at 51 studies with a total of 5,102 participants. There were studies in which participants were asked to sit in filthy work areas, studies in which participants listened to the sounds of people vomiting, even a study in which participants watched a video of a woman vomiting in a dirty bathroom, then retrieving her gum from the toilet and chewing it. The question was whether these manipulations actually have any effect on people's moral judgments.

A random effects meta-analysis revealed only a small effect (d = .11). However, this analysis showed strong signs of publication bias. After correcting for publication bias, the effect completely disappeared (d = -.01). In other words, the evidence from the 51 studies, taken as a whole, indicated that manipulations of disgust have basically no effect at all on moral judgment.

Although this meta-analysis found no effect on the whole, it should be noted that there was a large effect in one particular study (namely, Eskine, Kacinik and Prinz, 2011). In that study, participants were randomly assigned to drink a sweet beverage, a neutral beverage, or a disgustingly bitter beverage. This manipulation had a very substantial impact on moral judgments, with participants in the disgusting beverage condition giving much harsher moral judgments than those in the sweet beverage condition (d = 1.22). In short, even if there is no effect overall – looking across the many different studies that have examined this phenomenon in different ways – one might well think that there is indeed an effect of this specific manipulation.

In an attempt to get clear about the impact of this specific manipulation, Ghelfi and colleagues (2020) conducted a replication. The replication study used the same experimental design but a much larger sample size. In the original study, there were a total of 57 participants (divided across the three conditions), whereas in the replication study, there were 1,137 participants. Using this much more rigorous approach, the replication study found an effect size that was negligible (and actually in the opposite of the predicted direction). In other words, there was no tendency for participants drinking a disgusting beverage to give harsher moral judgments and, if anything, a tiny tendency for them to give less harsh moral judgments.

What are we to make of all of this? When philosophers talk about these results in informal conversations, I often hear them suggest that this whole research program has basically been a failure. The narrative goes something like this: Initially, it seemed like research

in this area was pointing to something really interesting about the impact of disgust on moral judgment, but in the end, the whole thing fizzled, and it turned out that there wasn't anything interesting going on here after all.

In my view, this narrative misses the point. Research in this area has successfully discovered something of great theoretical and philosophical importance. It has shown that people's moral judgments are strikingly stable across manipulations of incidental disgust. The studies reviewed here examine a wide variety of methods for inducing disgust, and in many cases, manipulation checks indicate that these methods were indeed successful in shifting people's emotional state. Yet this shift in emotional state had little or no impact on moral judgments. This is an important finding that is worthy of further philosophical exploration.

Of course, this claim of stability must always be a matter of degree. Existing research could not possibly show that there is literally no effect of incidental disgust on moral judgment. All it can show is that such an effect, if it exists at all, would have to be extremely small. Similarly, existing research cannot possibly show that there are no situations at all in which the impact of incidental disgust is large. All it shows is that when researchers created situations specifically for the purpose of finding a large effect of incidental disgust, what they found instead was little or no effect. In short, the finding is not that there is no effect at all in any situation; it is that there is little or no effect in the situations that researchers have dreamed up thus far.

Still, this is quite an interesting and puzzling finding. Existing research provides at least some evidence that (a) people's intuitions about the moral status of an action can be influenced by disgust. But we also seem to be finding increasing evidence that (b) people's intuitions are remarkably unaffected by manipulations of the external situation that lead them to feel disgust about unrelated objects. One important task moving forward will be to develop an account that explains this stability.

3.2. Truetemp and order effects

The Truetemp thought experiment was first introduced by Lehrer (1990) as part of an argument against reliabilism. The thought experiment runs as follows:

> Suppose a person, whom we shall name Mr. Truetemp, undergoes brain surgery by an experimental surgeon who invents a small device which is both a very accurate thermometer and a computational device capable of generating thoughts. The device, call it a tempucomp, is implanted in Truetemp's head so that the very tip of the device, no larger than the head of a pin, sits unnoticed on his scalp and acts as a sensor to transmit information about the temperature to the computational system of his brain. This device, in turn, sends a message to his brain causing him to think of the temperature recorded by the external sensor. Assume that the tempucomp is very reliable, and so his thoughts are correct temperature thoughts. All told, this is a reliable belief-forming process. Now imagine, finally, that he has no idea that the tempucomp has been inserted in his brain, is only slightly puzzled about why he thinks so obsessively about the temperature, but never checks a thermometer to determine whether these thoughts about the temperature are correct. He accepts them unreflectively, another effect of the tempucomp. Thus, he thinks and accepts that the temperature is 104 degrees. It is. Does he know that it is? (Lehrer 1990:163-164)

On externalist theories like reliabilism, it seems that the answer should be that Truetemp does know, whereas on internalist theories of the type Lehrer himself favors, the answer is that Truetemp does not know. When Lehrer first introduced this case, he suggested that the correct answer was clearly that Truetemp did not have knowledge (Lehrer, 1990).

In an important study, Swain, Alexander and Weinberg (2008) gave participants the Truetemp case and asked them to rate (on a 1-5 scale) their agreement with the claim that Truetemp knows. This study revealed something extremely surprising. It was absolutely not the case that participants generally converged on giving this case a low rating. Instead, participants seemed to treat this case as a confusing one.

More recently, Ziółkowski (in press) conducted a replication of this study. Ziółkowski kindly made his data available, and I put together a figure to show the percentage of participants who gave each response (see Figure 11). As the figure shows, people's ratings are all over the place. Some say that Truetemp does not know, some say that he does know, and many say that they are uncertain.

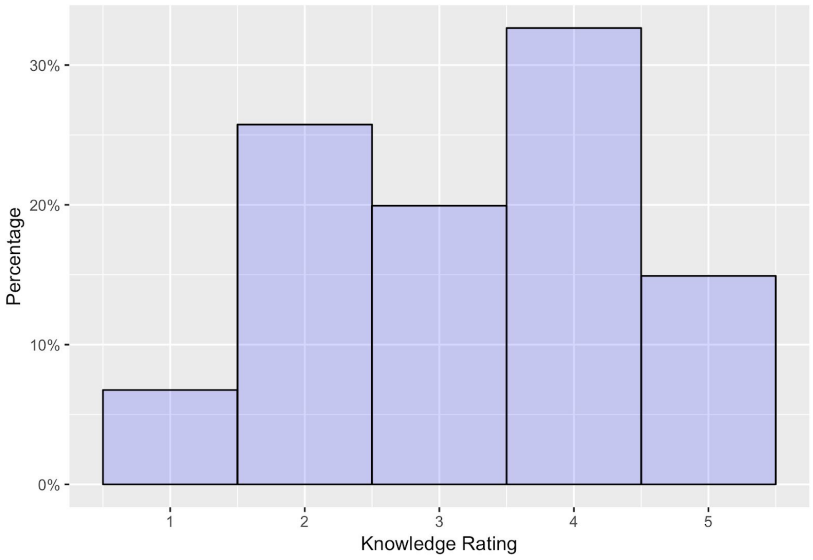

Figure 11. Histogram showing the percentage of participants giving each possible response, collapsing across the three studies reported in Ziółkowski (in press).

A question now arises about how to explain people's intuitions about this case. One obvious suggestion would be that there is something drawing people toward internalism but also something drawing them toward externalism. This possibility has not been systematically explored within existing work, but in my view, it is a promising avenue for future research.

Swain and colleagues (2008) offered a different suggestion that also seemed promising. They suggested that people's intuitions about this case were unstable. More specifically, they suggested that people's intuitions about this case might depend in part on which cases they happen to have considered right before it. If you had just been thinking of a case that is obviously a case of knowledge, the Truetemp case will appear by contrast to be a case of

non-knowledge. But if you had just been thinking of a case that is obviously a case of non-knowledge, the Truetemp case will appear by contrast to be knowledge.

To test this hypothesis, Swain et al. introduced a manipulation that has become a true classic, cited and discussed in numerous papers about the idea that people's philosophical intuition are unstable. Although the design included a few subtle complexities, the basic idea was simple enough. All participants received the Truetemp case and were asked to rate the degree to which they agreed that Truetemp had knowledge. Some participants received the Truetemp case right after receiving a case that was clearly a case of knowledge, some received it right after receiving a case that was clearly a case of non-knowledge, and some received it without receiving any other case previously.

Figure 12 shows the results. There was a significant effect whereby participants' intuitions about the Truetemp case depended on what they had seen previously, and the pattern of the effect seemed to make a lot of sense. Participants were more inclined to attribute knowledge after receiving the non-knowledge case, and less inclined to attribute knowledge after receiving the knowledge case.

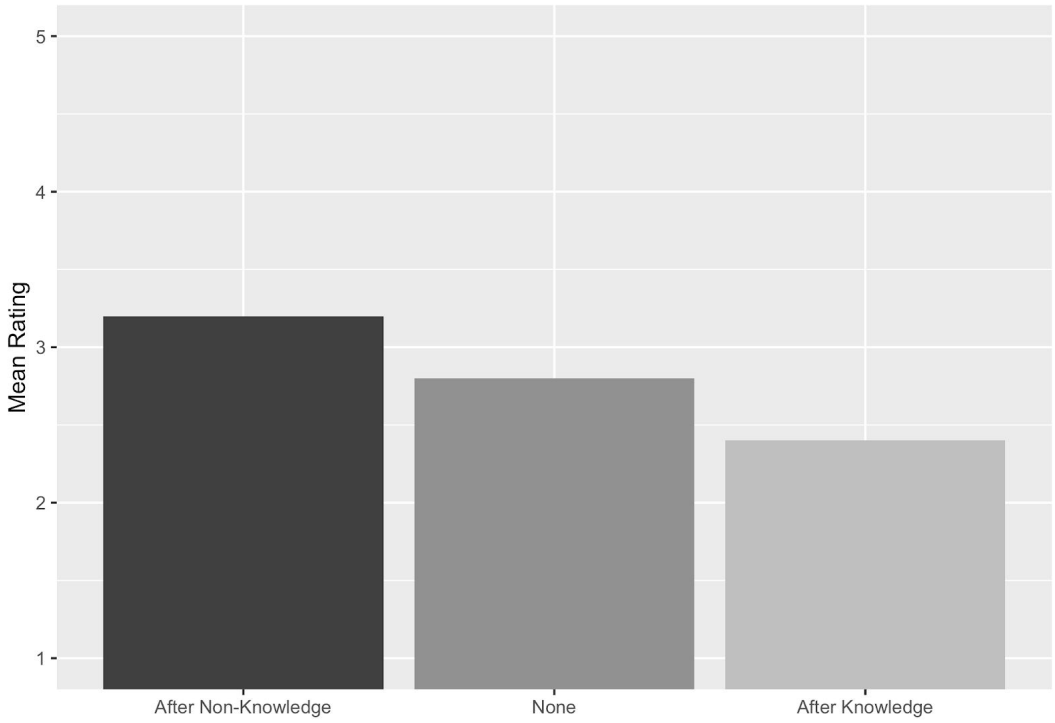

Figure 12. Means by condition in Swain et al. (2008).

From a purely empirical standpoint, the results of the paper do not provide strong evidence of an order effect. The p-values for the effects we have been discussing are: omnibus ANOVA, $p = .048$; after non-knowledge vs. none, $p = .043$; after knowledge vs. none, $p = .054$. Anyone who has been following the replication crisis in psychology will recognize that results like these are generally highly unlikely to replicate successfully.

Still, when I first read this paper, I was completely convinced. The thing that made the paper so compelling was not so much the actual empirical results as the larger theoretical picture. As the authors themselves emphasized, their finding seemed to fit well with a broader pattern of results that were emerging in the empirical literature at the time. Numerous studies seemed to be revealing various ways that people's intuitions were unstable, and it therefore seemed highly plausible that intuitions might turn out to be unstable in this way as well. But it wasn't just that. The real point was that the key prediction seemed to follow naturally from a larger vision that would help to explain the instability. Just put yourself in the position of an ordinary participant in this study. You have never thought before about any of these issues, and suddenly you are faced with a complex question that pits different considerations against each other. How are you going to respond? One obvious hypothesis would be that your response will depend on what you happen to be thinking about the time, and it seems only natural to suppose that what you are thinking about would depend on which case you happened to be considering previously. Putting all of this together, it really felt like there was a lot of good reason to suppose that people would answer this question differently depending on which case they had considered previously.

As we have seen, the overall empirical situation has changed considerably in the years since then, and we now have more reason than we previously did to suspect that people's intuitions might actually be stable in cases like this one.[9] Motivated in part by these larger empirical developments, Ziółkowski (in press) decided to replicate the Swain et al. study. He ran three replication studies in total. These three studies used almost exactly the same methods, except that one involved translating all of the materials into Polish and running it on Polish speakers.

The replication studies find no evidence of an order effect. Using data provided by Ziółkowski, I put together a figure showing the results (see Figure 13). There was no significant effect in any of the individual studies, and a meta-analysis also shows no significant effect overall. In fact, putting all three studies together in a meta-analysis, the impact of the negative prime is not even marginally significant (p = .25). More importantly, the meta-analysis indicates that people's responses were almost exactly the same regardless of condition (Hedges' g = -.07, which, given the sample sizes in each condition, means that Cohen's d = .-07). In short, the results indicate that intuitions are stable across this manipulation of order.

[9] I should note that I myself may have contributed to the problem here. In motivating their hypothesis, Swain et al. draw on my claim that people's intuitions about free will depend in part on whether they receive a case that triggers weak affective reactions (e.g., tax evasion) or strong affective reactions (e.g., murder). Initially, it looked like this claim might be true (Nichols & Knobe, 2007), but like so many of the other claims that motivated this approach, this one turned out to be false (Feltz & Cova, 2014).

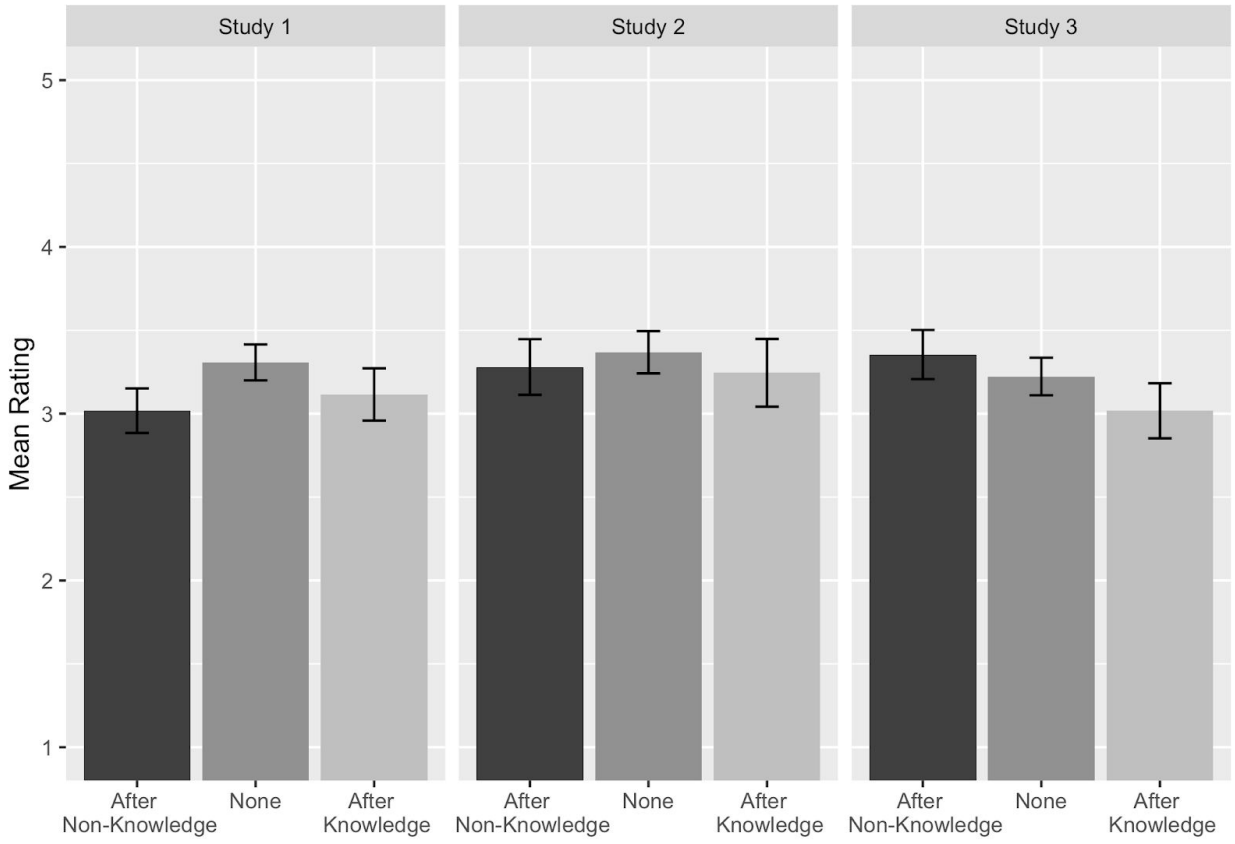

Figure 13. Means by condition in the three studies reported in Ziółkowski (in press). Error bars show SE mean.

Ziółkowski and colleagues (unpublished) then conducted a replication with a much larger sample size (1,626 participants). Figure 14 shows the results. Once again, knowledge intuitions were almost exactly the same across the three conditions.

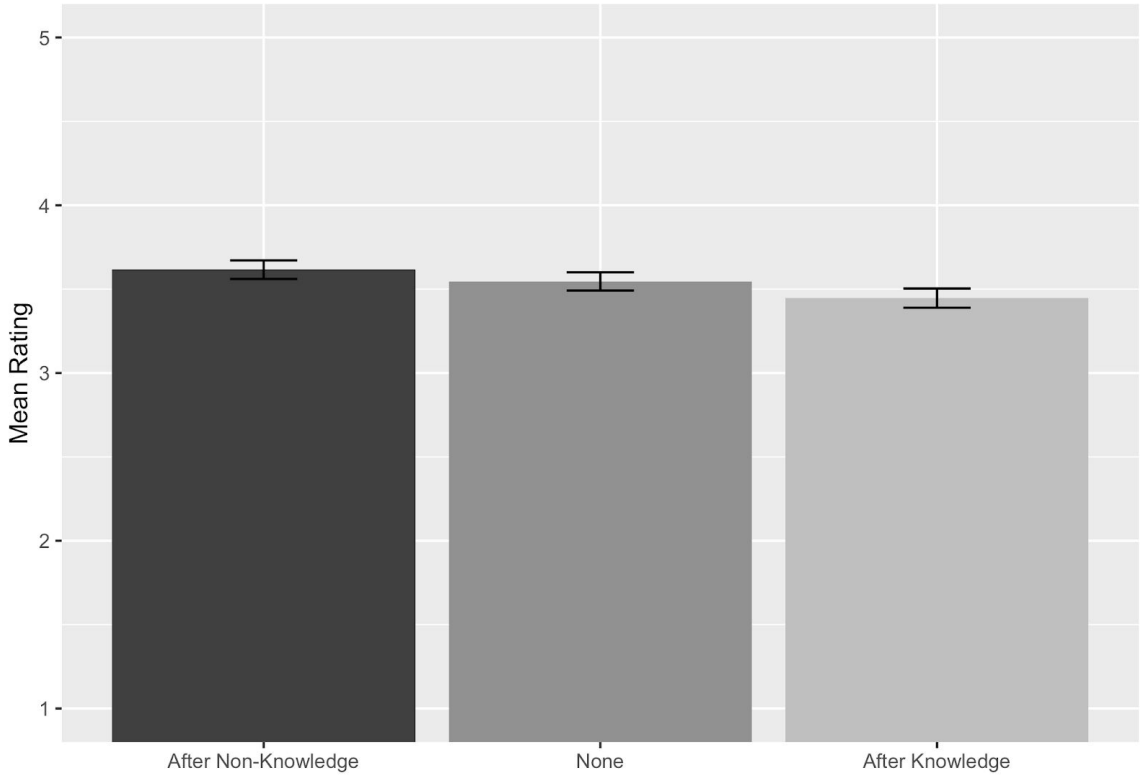

Figure 14. Means by condition in Ziółkowski et al. (unpublished). Error bars show SE mean.

These results not only remove one of the most salient pieces of evidence in favor of the claim that philosophical intuitions are unstable, they actually provide evidence for the claim that philosophical intuitions are surprisingly stable. Existing research shows that (a) this is a confusing case, which leads different people to have different intuitions and leaves many people completely uncertain, but these new studies now suggest that (b) this confusion is itself remarkably stable, with people showing almost exactly the same complex distribution of responses regardless of which case they considered previously. We need to develop an account of people's intuition that can explain this stability.

3.3. Moral dilemmas and manipulations of dual process

Perhaps the most influential of all conflict-based approaches to philosophical judgments is Greene's dual process explanation of judgments about moral dilemmas (Greene, Sommerville, Nystrom, Darley & Cohen, 2001). On this account, people are drawn toward ‘characteristically deontological’ judgments by a more fast, automatic, intuitive process and toward ‘characteristically utilitarian’ judgments by a more slow, controlled, reflective process.

This account raises numerous fascinating questions, but we will be focusing here just on one relatively narrow issue. If it is indeed the case that people are drawn toward utilitarian judgments by a more reflective process, one might predict that it should be possible to shift

people's judgments around by manipulating certain external situational factors. Manipulations that hinder the use of reflective processes should make people less inclined to give utilitarian responses, while manipulations that promote the use of reflective processes should make people more inclined to give such responses.

In one important test of this hypothesis, Greene and colleagues (2008) conducted a study in which they tried to diminish people's capacity to apply more reflective cognition using cognitive load. Participants in the cognitive load condition had to evaluate moral dilemmas while also performing another task at the same time. (A stream of numbers was scrolling across the bottom of the screen, and they had to press a button every time they saw the number '5.') Participants in the control condition simply read and responded to the dilemmas without having to perform any other task at the same time. Prior studies provided strong evidence that performing this other task should decrease people's ability to respond using reflective cognition (Gilbert, Tafarodi & Malone, 1993).

The study examined the impact of this manipulation on two different measures. First, the researchers looked at reaction times. Would the manipulation affect how long it took participants to arrive at a judgment? Second, the researchers looked at people's actual judgments. Would the manipulation lead to a shift in the judgments participants ultimately ended up making?

The results showed an effect on reaction times that was very much in keeping with the dual process theory. Cognitive load led to longer reaction times for utilitarian decisions but not to longer reaction times for deontological decisions. But, strikingly, the manipulation *did not* impact people's actual judgments. Participants were no less likely to make utilitarian judgments under cognitive load than they were in the control condition.

This stability even in the face of a well-validated manipulation of the use of reflective processes is quite a surprising result. Greene and colleagues explain the null effect as follows:

> As noted above, the cognitive load manipulation did not reduce the proportion of utilitarian judgments. One explanation for this is that participants were keenly aware of the interference created by the load manipulation and were determined to push through it. Like motorists facing highway construction, they may have been delayed, but not ultimately prevented from reaching their destinations. (Greene et al., 2008: 1151)

I want to draw attention to these three sentences because they are the only attempt I know of within the existing literature to provide a real, empirically testable explanation for the stability of a philosophical intuition across manipulations of the external situation. The core suggestion is that in cases where the presence of an unusual external factor might naturally tend to shift people away from the judgment they would ordinarily make, a further process kicks in that allows them to continue making that same judgment. This is a very intriguing suggestion, which would be amply worthy of further empirical exploration.

In another important study, Paxton and colleagues tried a manipulation designed to move things in the opposite direction, leading people to use more reflective cognition than they ordinarily would (Paxton, Ungar & Greene, 2012). Participants in one condition evaluated moral dilemmas immediately after going through the cognitive reflection test (CRT). The core idea

here is that the CRT could serve not merely as a measure of reflectiveness but also as a manipulation of reflectiveness. If you go through the CRT and get at least one question right, you will presumably find yourself getting in a mindset in which you are a bit more inclined to question your intuitions. Participants in the control condition went through the steps of the study in the opposite order. They first evaluated the moral dilemmas, then answered the CRT questions.

Paxton and colleagues found a significant effect of this manipulation on people's moral judgments. Participants in the CRT-first condition were more inclined to give utilitarian answers than were participants in the control condition. The p-value for this effect was literally $p = .05$ (which should immediately raise some doubts), but again, this empirical result was embedded in a larger theoretical picture that many of us found extremely compelling.

More recently, however, Cova and colleagues (2018) conducted a replication study with a larger sample size ($N = 729$). All data for this study were made freely available, and I put together a figure that shows the results (see Figure 15). As the figure shows, the distribution of participants' moral judgments was almost exactly the same in the two conditions. The replication study found no significant effect of the manipulation but instead found evidence that participants' moral judgments were remarkably stable ($d = .14$).

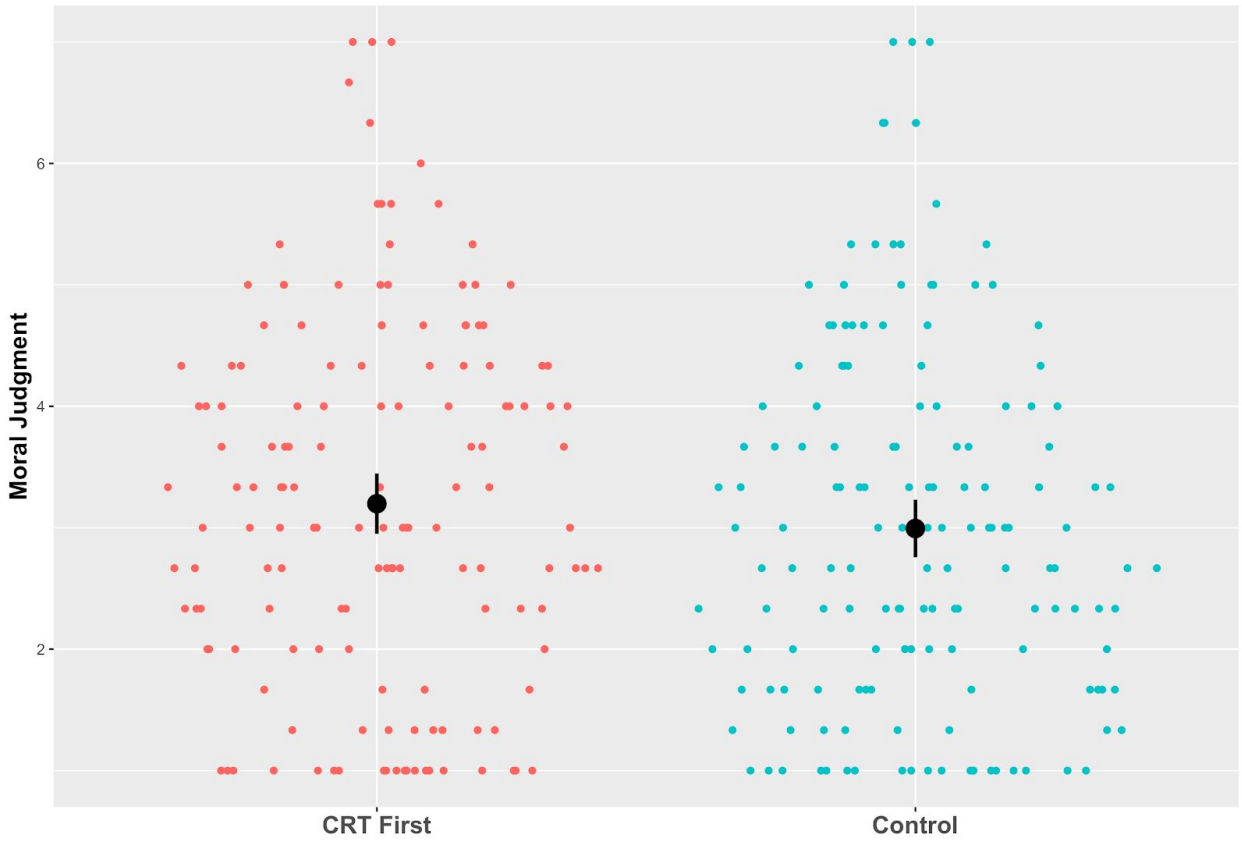

Figure 15. Sinaplot showing the results of the Cova et al. replication study. Each participant answered three moral questions on a 1-7 scale, with higher numbers corresponding to more utilitarian judgments. Colored points show the mean rating for each individual participant. Grey circles show the overall mean for each condition. Error bars show 95% confidence intervals.

Note that the explanation Greene and colleagues (2008) provide for the stability observed in their own study cannot also be applied to the stability observed in this study. It would not make sense to say that when participants encounter a manipulation that promotes reflection, they notice this problem and try to engage an additional process that enables them to arrive at the same judgment they would have made if they had been less reflective. Of course, this does not mean that the Greene et al. explanation is wrong; it just means that if we employ that explanation for the stability observed in the previous study, we will need a completely separate explanation for the stability observed in this one.[10]

[10] I have been focusing here just on intuitions about moral questions, but this striking stability across manipulations designed to increase reflectiveness seems to be part of a more pervasive phenomenon that goes beyond morality in particular. In an important recent paper, Kneer and colleagues (in press) looked more generally at whether manipulations designed to increase reflectiveness ever impact people's philosophical intuitions. They explored four different cases in which previous experimental philosophy studies found intuitions that didn't conform to philosophical orthodoxy (Gettier cases, knowledge without belief, the epistemic side-effect effect, the Gödel case), and they used four different manipulations designed to increase reflectiveness (forced delay, financial incentive, request for justification, analytic

In sum, what we have here is a third example of stability in the face of external situational factors that one might have expected to shift around people's philosophical intuitions. Existing findings provide substantial evidence that (a) people tend to provide more utilitarian answers in moral dilemmas when they use more reflective processes, but we are now getting evidence that (b) manipulations of the external situation that we would expect to hinder or promote reflective processes have very little impact on the degree to which people provide utilitarian responses. To be honest, I have no idea why people's intuitions show this kind of stability. I definitely would not have predicted these results beforehand, and even now that they have been obtained, I still can't help but find them deeply surprising and puzzling.

3.4. Trolley problems and order effects

Finally, let's turn to a case in which empirical research actually has successfully identified an effect of external factors. When participants receive a sequence of different trolley dilemmas, it sometimes happens that participants' intuitions about one dilemma depend on which other dilemmas they received previously. In other words, trolley problem intuitions do sometimes show order effects. There is a wonderfully rigorous empirical literature on these order effects (Lanteri, Chelini & Rizzello, 2008; Petrinovich & O'Neill, 1996; Schwitzgebel & Cushman, 2015; Wiegmann, Okan & Nagel, 2012; Wiegmann & Waldmann, 2014), and we now have a great deal of information about precisely when they do and do not arise.

So then, what has this empirical research actually discovered? Looking at the way it is discussed in the philosophical literature, one might well come away with the sense that the key finding is that the effect of order is far more sweeping and pervasive than one would initially have expected. But this is not at all the case. To a first approximation, the actual finding is exactly the opposite. Empirical research finds that there is only an order effect in certain very specific kinds of cases. This finding has led to a systematic effort to more accurately understand the factors that explain why the effect is limited in this way and why it is that we *don't* find an order effect in so many cases.

A number of different studies have used a design in which each individual participant receives two different trolley dilemmas. In both dilemmas, one has an option that involves killing one person but rescuing five people. One dilemma involves a question about whether one should switch a train on to a sidetrack (the 'sidetrack' dilemma), while another involves a question about whether one should push an innocent person off a footbridge (the 'footbridge' dilemma). These two dilemmas can be presented in either of two possible orders: either sidetrack-footbridge or footbridge-sidetrack. Studies using this design consistently find an order effect (Lanteri, Chelini & Rizzello, 2008; Schwitzgebel & Cushman, 2015; Wiegmann, Okan & Nagel, 2012; Wiegmann & Waldmann, 2014). In the sidetrack dilemma, participants are more inclined to say that it would be wrong to turn the train if they received the dilemmas in the footbridge-sidetrack order than if they receive them in the sidetrack-footbridge order. This is an example of an impact of external situation factors on philosophical intuitions.

---

priming). There was no evidence that any of the manipulations impacted intuitions about any of the cases. I am not sure why this is happening, but it seems like there is some very general tendency whereby people's philosophical intuitions are remarkably unaffected by this sort of manipulation.

A question now arises as to how to understand this effect. One might initially think that it is best understood as just one example of some sweeping and pervasive pattern whereby intuitions about philosophical questions are just generally susceptible to all sorts of different effects of situational factors. As we have seen, there is now evidence against this more sweeping interpretation. Still, one might think that what we have here has got to be an example of something at least relatively general. For example, one might think that it is an example of a more general tendency that emerges across a wide variety of moral judgments. Or, failing that, one might think it is at least an example of more general tendency that emerges across different kinds of trolley problems.

The exciting result coming out of the empirical literature is that this seems not to be the case. Instead, it appears that there is something special about the sidetrack trolley problem in particular that makes it susceptible to order effects. Other trolley problems seem not to show order effects in the same way.

To begin with, existing studies show that the order effect is asymmetric. Receiving the footbridge case first impacts people's intuitions about the sidetrack case, but receiving the sidetrack case first does not have nearly the same impact on intuitions about the footbridge case. In some studies, the impact on footbridge intuitions is much smaller than the impact on sidetrack intuitions (Schwitzgebel & Cushman, 2015), and in many studies, there is basically no impact at all on footbridge intuitions (Lanteri et al., 2008; Wiegmann, Okan & Nagel, 2012; Wiegmann & Waldmann, 2014). A similar pattern emerges when one replaces the footbridge dilemma with a dilemma about whether to kill one person so as to be able to transfer her organs and save five (the 'transplant' dilemma). Studies show that receiving the transplant dilemma first impacts intuition about the sidetrack dilemma but receiving the sidetrack dilemma first does not impact intuitions about the transplant dilemma (Petrinovich & O'Neill, 1996).

Wiegmann and Waldmann (2014) develop a hypothesis to explain this pattern of results. At the core of the hypothesis is the idea that the order effect is due to something about certain idiosyncratic features of the footbridge-sidetrack sequence in particular. The claim is that when people receive the sidetrack dilemma, it is normally possible for them to focus almost exclusively on one specific causal chain (saving the five) and to focus much less on the other causal chain (killing the one), whereas when participants receive the footbridge dilemma, these two events are in the same causal chain (killing the one → saving the five) and focusing on the saving therefore inevitably involves focusing on the killing as well. However, there is also an important sense in which people see the death of the one in the sidetrack dilemma as analogous to the death of the one in the footbridge dilemma. Thus, when people receive the footbridge dilemma, they focus on the death of the one, and when they receive the sidetrack dilemma immediately thereafter, they see the death of the one in that latter dilemma as directly analogous and end up focusing on it as well. This psychological process leads to an order effect in that one sequence.

Importantly, the process posited by this hypothesis would lead to an effect in the footbridge-sidetrack sequence but would not lead to a similar effect in most other sequences of moral dilemmas, or even in most other sequences of trolley dilemmas. The hypothesis therefore leads immediately to some exciting new predictions about other cases in which the order effect should not arise. For example, consider a case in which the agent can save three people by

pushing them out of the way of the train, but doing so would lead to the death of one person (the 'PosPush' dilemma). This dilemma is similar to the sidetrack dilemma in that people are on the whole relatively inclined to say that the action is permissible, but strikingly, the hypothesis predicts that it should not show the order effect observed for the sidetrack dilemma. (The saving of the three is not part of a separate causal chain from the killing of the one.) Wiegmann and Waldmann (2014) look at intuitions about this case, and just as the hypothesis predicts, they find no effect of order.

I would not want to put too much emphasis on the idea that there is literally no order effect at all in this case. Even if the study found no effect, it is still possible that there actually is an effect – just one that is too small to be detected with this sample size. Still, the study does provide strong evidence for an important conclusion. It suggests that whatever process generates the order effect in the footbridge-sidetrack sequence, that same process is not at work in the footbridge-PosPush sequence.

Let us now sum up. In the specific case in which participants receive the footbridge dilemma and then receive the sidetrack dilemma, we have strong evidence for an impact of external situational factors: thinking about the first dilemma really does influence judgments about the second. What is this fact teaching us? The obvious first guess would be that it is evidence of a process that leads to some broader form of instability. Perhaps the process affects people's philosophical intuitions more generally, or perhaps just their moral intuitions, or at a very minimum, it surely affects people's intuitions about a wide variety of different trolley problems. The surprising finding coming out of research in this area is that this is *not* what is happening. The effect observed in this one case seems to be highly circumscribed. Not only does it not emerge in cases that are radically different, it doesn't even emerge in cases that might at first seem extraordinarily similar to the footbridge-sidetrack sequence.

Most directly, this finding provides evidence that the process that generates the order effect in this one case does not also generate an order effect in a wide variety of other cases. But it also shows something further. It shows that if there is indeed a process that generates an order effect in all of these other cases, the effect of that latter process is so small that it could not even be detected in the studies we have been reviewing here.

3.5. Summary

In this section, we looked at existing evidence regarding the impact of external situational factors on philosophical intuitions. The results indicated that situational factors have surprisingly little influence. Early studies showing an influence of disgust on moral intuition, order effects on epistemic intuition and dual process manipulation on moral intuition have all failed to replicate. When it comes to intuitions about moral dilemmas, existing research finds no effect of cognitive load and only an extremely circumscribed effect of order.

I do not mean to exaggerate the implications of these findings. If we continue looking, we will surely be able to eventually come up with at least a few experimental paradigms in which people show broad and replicable effects of the external situation on philosophical intuitions. (For example, we will surely be able to identify at least a few further cases in which people show strong and replicable order effects.) Moreover, even in the experimental paradigms reviewed

here, it seems likely that the impact of the external situation on philosophical intuitions is not strictly zero. I would not want to deny any of this.

Still, the results we have been reviewing here are extremely surprising. We have been looking at cases in which existing theory provided strong reason to expect an impact of external situational factors. Moreover, we have been looking at precisely the cases that have most often been invoked by philosophers who claim that there is an effect of such factors. Yet, in each case, what we found was remarkable stability. Regardless of what might be found in future studies that explore other cases, this result is in itself a deeply interesting and important one.

4. Conclusion

I have been arguing that people's philosophical intuitions are surprisingly stable. This surprising stability emerges both across demographic groups and across situations. It emerges for intuitions about everything from morality to epistemology. It emerges for factors tied to emotion, reflection, age and culture. It really does seem to be a quite pervasive phenomenon.

The stability of people's intuitions raises some difficult empirical and philosophical questions, but unfortunately, I do not have answers to any of these questions. In this concluding section, I will therefore be providing a few preliminary thoughts that might at least indicate potential avenues for further inquiry.

4.1. Empirical questions

Within the existing literature, there have been numerous papers that engage in detail with one or another specific form of stability (one paper on the stability of Truetemp intuitions, another on the stability of trolley intuitions, etc.). These more detailed investigations are absolutely essential, but it seems at this point that there is also a need for some broader theoretical work. We want some understanding of why so many different studies, exploring what might seem like very different sorts of intuitions, have ended up with this same result.

On one hand, it would be a mistake to look for an account that is overly general. The stability we find in philosophical intuitions cannot be explained just by introducing some extremely broad theory that would predict stability also for non-philosophical judgments. For example, it would clearly be a mistake to say that non-philosophical judgments also tend to be surprisingly stable across cultures. What needs to be explained, then, is why people's philosophical intuitions differ from other kinds of judgments. What is it about the sorts of intuitions frequently invoked in philosophy that makes them stable in a way that many other aspects of human cognition are not?

To illustrate the question we face here, consider a recent cross-cultural study on intuitions about the self (De Freitas et al., 2018). In this study, the very same participants were given both (a) a straightforward non-philosophical question and (b) a question about a philosophical thought experiment. For the non-philosophical question, participants were simply asked about their own selves. Unsurprisingly, there was a large cross-cultural difference. Participants from Western cultures saw themselves as more 'independent,' while participants from non-Western cultures saw themselves as more 'interdependent.' For the philosophical

question, these very same participants were then given a thought experiment designed to probe their intuitions about the nature of the self. On this latter question, the results showed striking cross-cultural stability. The patterns that had previously been observed in American participants also arose in participants from every other culture. What we need now is a way to explain this stability in philosophical intuitions that would not also mistakenly predict stability in non-philosophical judgments.[11]

On the other hand, however, it would also be a mistake to find an account that is overly narrow. Presumably, we need an account that does not apply just to one particular type of intuition or one particular respect in which it might be stable. After all, the results reviewed above indicate that a whole bunch of different philosophical intuitions are surprisingly stable in a whole bunch of different respects. It seems that we need an account that gives us some understanding of this broad pattern.

Clearly, this is a tall order. The different effects discussed in the previous section were predicted by what appear to be completely unrelated theories.The effect of incidental disgust was predicted by theories about the nature of disgust. Then, quite separately, the effect of cognitive load was predicted by theories about moral judgment. And similarly, each of the other effects was predicted by a separate theory. Yet what we actually find in the existing results is a broader phenomenon whereby almost all of these effects seem not to be emerging. How is this to be explained?

One possible answer would be that there was something wrong in each of the separate theories that predicted each of the separate effects. Perhaps we were mistaken in some way about the impact of incidental disgust, and we were also mistaken in another, completely unrelated way about the impact of cognitive load… and so forth, with a different mistake explaining each of the different cases in which predicted instability failed to emerge. This is certainly one possibility. The difficulty is just that thi*s s*ort of approach fails to explain the more general pattern. It certainly seems as though we need some explanation of the fact that philosophical intuitions turned out to be so stable across all of these different factors. Future work should strive to articulate theories that have some real potential to explain this more general pattern.

---

[11] Just to throw out one possible hypothesis: It might be that a culture has a greater impact on people's judgments in domains that are explicitly discussed within that culture and then it does in domains that are not explicitly discussed within that culture. In a typical culture, there is a lot of explicit discussion of questions about interpersonal obligations (obligations to the family, obligations to the ingroup) but very little explicit discussion of philosophical thought experiments (the Gettier problem, the epistemic side-effect effect), and precisely for this reason, a typical culture will exert a lot of influence on people's intuitions about the former but not about the latter.

Importantly, however, there is at least one culture that actually does engage in frequent explicit discussion of philosophical thought experiments, namely, the culture of academic philosophy departments. This hypothesis therefore predicts that although most ordinary cultures should not differ substantially from each other when it comes to philosophical intuitions, people in philosophy departments might well differ from all of those other cultures. Existing findings provide at least some support for that claim (e.g., Cova et al., 2019; Rose et al., 2019).

4.2. Philosophical questions

At the same time, these findings present us with a new and difficult philosophical question. What does the stability of people's philosophical intuitions teach us about the use of intuitions in philosophy?

Of course, our answer to this philosophical question will depend in large part on how we answer the empirical questions described in the previous subsection. The stability of people's philosophical intuitions seems to be showing us something about the psychological processes that generate those intuitions. Depending on what we end up concluding about these psychological processes, we will presumably arrive at different conclusions about the use of intuitions in philosophy.

Still, it does seem that there are a few things we can say even at this point. That is, there are at least some implications of the stability of people's philosophical intuitions that do not depend on facts about the underlying psychological processes that create that stability.

First, there is an obvious lesson for existing debates about the philosophical implications of instability. If people's intuitions are surprisingly stable, then the whole debate about the implications of instability is *moot*. As we noted at the outset, there has been an enormous amount of research focused on the question: “If we learn that people's intuitions are unstable, what should we conclude about the use of intuitions in philosophy?” Such research has shown impressive levels of sophistication and ingenuity, but if people's intuitions are not in fact unstable, this is simply not the question we face.

Second, these findings of stability of philosophical intuitions seem to Implications for work on *conflicting intuitions*. Suppose you meet someone who is in the grip of this sort of conflict, pulled by one cognitive process toward a certain intuition, but pulled by another cognitive process toward the exact opposite intuition. You might at first think that the tension she is experiencing merely reflects some highly contingent facts about her particular position. For example, you might think that it is merely a product of the culture she happens to inhabit, or of the situation in which she happens to find herself. But recent empirical work suggests that this concern is misplaced. It suggests that the tensions people experience when they grapple with philosophical problems are rooted in something deeper and more fundamental, namely, in factors that are surprisingly stable across both cultures and situations.

Thus, recent work directs us toward a more sustained investigation of the philosophical implications of conflicting intuitions. It forcefully raises the question: “If we learn more about the psychological processes drawing people in each of these conflicting directions, can this knowledge help us make progress in addressing the philosophical issues themselves?” I don't mean to be taking any position about the correct answer to this question. My point is simply that this is a question we actually face.

References

Adleberg, T., Thompson, M., & Nahmias, E. (2015). Do men and women have different philosophical intuitions? Further data. *Philosophical Psychology*, 28, 615-641.

Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, norm violation, and culpable control. *The Journal of Philosophy*, 108, 670-696.

Aristotle. (1999/340 BCE). *Nicomachean Ethics*, Terence H. Irwin (ed./trans.), Indianapolis: Hackett Publishing Co.

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*, 121, 154-161.

Buckwalter, W., & Stich, S. (2014). Gender and philosophical intuition. *Experimental philosophy, 2,* 307-346.

Busch, J., Bergenholtz, C., Praëm, S. K. (in press). Further insights on fake barn cases and intuition variation. *Episteme*.

Colebrook, R. (2020). Folk Classification of Moral Issues. Unpublished manuscript. Baruch College (CUNY).

Conway, P. & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology*, 104, 216.

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., ... & Cushman, F. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*, 1-36.

Cova, F., Olivola, C. Y., Machery, E., Stich, S., Rose, D., Alai, M., ... & Cheon, H. (2019). De pulchritudine non est disputandum? A cross-cultural investigation of the alleged intersubjective validity of aesthetic judgment. *Mind & Language,* 34(3), 317-338.

Cushman, F., & Mele, A. (2008). Intentional action: Two-and-a-half folk concepts? In J. Knobe & S. Nichols (Eds.), *Experimental Philosophy* (pp. 171–188). Oxford: Oxford University Press.

De Brigard, F. (2010). If you like it, does it matter if it's real? *Philosophical Psychology,* 23, 43-57.

De Freitas, J., Sarkissian, H., Newman, G. E., Grossmann, I., De Brigard, F., Luco, A., & Knobe, J. (2018). Consistent belief in a good true self in misanthropes and three interdependent cultures. *Cognitive Science,* 42, 134-160.

Demaree-Cotton, J. (2016). Do framing effects make moral intuitions unreliable? *Philosophical Psychology,* 29, 1-22.

Devitt, M., & Porot, N. (2018). The reference of proper names: Testing usage and intuitions. *Cognitive Science*, 42, 1552-1585.

DeRose, K. (2017). *The Appearance of Ignorance: Knowledge, Skepticism, and Context (Vol. 2)*. Oxford University Press.

Dranseika V., Lauraitytė E., & Experimental Jurisprudence Cross-Cultural Study Swap Consortium (unpublished data). Cross-cultural Replication of Tobia (2016)

Feltz, A. and E.T. Cokely (2008). The Fragmented Folk: More Evidence of Stable Individual Differences in Moral Judgments and Folk Intuitions. In B.C. Love, K. McRae and V.M. Sloutsky (eds.) *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, Austin, TX: Cognitive Science Society, 1771–76.

Feltz, A., & Cova, F. (2014). Moral responsibility and free will: A meta-analysis. *Consciousness and Cognition*, 30, 234-246.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives,* 19(4), 25-42.

Ghelfi, E., Christopherson, C. D., Urry, H. L., Lenne, R. L., Legate, N., Fischer, M. A., … Sullivan, D. (2020). Reexamining the effect of gustatory disgust on moral judgment: A multi-lab direct replication of Eskine, Kacinik, and Prinz (2011). *Advances in Methods and Practices in Psychological Science* 3.1, 3-23.

Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology*, 65(2), 221.

Goodwin, G.P., and J.M. Darley. 2008. The psychology of meta-ethics: Exploring objectivism. *Cognition* 106: 1339–1366.

Goodwin, G.P. and J.M. Darley, 2010. The Perceived Objectivity of Ethical Beliefs: Psychological Findings and Implications for Public Policy, *Review of Philosophy and Psychology*, 1: 161-88.

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(5537), 2105-2108.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, 107(3), 1144-1154.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology,* 65, 613.

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

Hannikainen, I. R., Tobia, K., Almeida, G., Dranseika, V., Kneer, M., Strohmaier, N., Janik, B. M., Próchnicki, M., Bystranowski, P., Aguiar, F., Dolinina, K., Rosas, A., & Struchiner, N. (in prep.). Cross-linguistic evidence of essentialist beliefs about the law.

Heiphetz, L. & Young, L. L. (2017). Can only one person be right? The development of objectivism and social preferences regarding widely shared and controversial moral beliefs. *Cognition*, 167, 78-90.

Inbar, Y., Pizarro, D. A., & Bloom, P. (2009). Conservatives are more easily disgusted than liberals. *Cognition and Emotion*, 23, 714-725.

Inbar, Y., Pizarro, D. A., Knobe, J., & Bloom, P. (2009). Disgust sensitivity predicts intuitive disapproval of gays. *Emotion*, 9, 435.

Johnson, D. J., Cheung, F., & Donnellan, M. B. (2014). Does cleanliness influence moral judgments? A direct replication of Schnall, Benton, and Harvey (2008). *Social Psychology*, 45(3), 209-215.

Kahan, D. M. (1999). The progressive appropriation of disgust. *The Passions of Law*, 63, 63-65.

Kelly, D. (2011). *Yuck!: The nature and moral significance of disgust*. MIT press.

Kim, M. & Yuan, Y. (2015). No cross-cultural differences in the Gettier car case intuition: A replication study of Weinberg et al. 2001. *Episteme*, 12(3), 355-361.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., ... & Batra, R. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443-490.

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C., Nosek, B. A., Chartier, C. R., ... & Cromar, R. (2019). Many Labs 4: Failure to Replicate Mortality Salience Effect With and Without Original Author Involvement. Unpublished manuscript. https://psyarxiv.com/vef2c

Kneer, M., Hannikainen, I.R., Almeida, G., Aguiar, F., Bystranowski, P., Dranseika, V., Janik, B. M., Garcia Olier, J., Güver, L., Liefgreen, A., Tobia, K., Próchnicki, M., Rosas, A., Skoczén, I., Strohmaier, N. & Struchiner, N. (2021). Outcome effects on mental state ascriptions across cultures. Unpublished manuscript. University of Zurich.

Kneer, M. (2021). A cross-cultural exploration of what we expect of each other in linguistic communication. Unpublished manuscript. University of Zurich.

Kneer, M., Colaço, D., Alexander, J., Machery, E. (in press). On second thought: Reflections on the reflection defense. *Oxford Studies in Experimental Philosophy*.

Kneer, M., Olivola, C. Y., Kim, H., Machery, E., Stich, S. P., Rose, D., ... & Zhu, J. (in prep. a). The action/omission distinction across cultures. University of Zurich.

Kneer, M., Machery, E., Stich, S. P., Rose, D., Alai, M., Angelucci, A.... & Zhu, J. (in prep. b). Moral luck across cultures. University of Zurich.

Knobe, J. (2019). Philosophical intuitions are surprisingly robust across demographic differences. *Epistemology & Philosophy of Science*, 56(2).

Knobe, J. (2020). Difference and Robustness in the Patterns of Philosophical Intuition across Demographic Groups. Unpublished manuscript. Yale University.

Kripke, S. A. (1972). Naming and necessity. In *Semantics of Natural Language* (pp. 253-355). Springer, Dordrecht.

Lanteri, A., Chelini, C., & Rizzello, S. (2008). An experimental investigation of emotions and reasoning in the trolley problem. *Journal of Business Ethics*, 83(4), 789-804.

Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect: Theory of mind and moral judgment. *Psychological Science*, 17, 421-427.

Lin, Z., Yu, J., & Zhu, L. (2019). Norm status, rather than norm type or blameworthiness, results in the side-effect effect. *PsyCh Journal*, 8, 513-519.

McCarthy, R. J., Skowronski, J. J., Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., ... & Barbosa, F. (2018). Registered replication report on Srull and Wyer (1979). *Advances in Methods and Practices in Psychological Science*, 1(3), 321-336.

Machery, E. & Stich, S. (2020). Demographic differences in philosophical intuition: A reply to Joshua Knobe. Unpublished manuscript.

Machery, E., Mallon, R., Nichols, S., & Stich, S. P. (2004). Semantics, cross-cultural style. *Cognition*, 92, B1-B12.

Machery, E. (2015). The illusion of expertise. Experimental Philosophy, Rationalism and Naturalism: Rethinking Philosophical Method, Routledge, 188-203.

Machery, E., Stich, S., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., ... & Hashimoto, T. (2017). Gettier Across Cultures 1. *Noûs*, 51(3), 645-664.

Machery, E., Stich, S., Rose, D., Alai, M., Angelucci, A., Berniūnas, R., ... & Cohnitz, D. (2017b). The Gettier Intuition from South America to Asia. *Journal of Indian Council of Philosophical Research*, 34(3), 517-541.

Mallon, R., Machery, E., Nichols, S., & Stich, S. (2009). Against arguments from reference. *Philosophy and Phenomenological Research*, 79, 332-356.

May, J. (2014). On the very concept of free will. *Synthese*, 191, 2849-2866.
May, J. (2018). *Regard for reason in the moral mind*. Oxford University Press.
Michelin C., Pellizzoni S., Tallandini M.A. & Siegal M. (2010) Evidence for the side-effect effect in young children: Influence of bilingualism and task presentation format. *European Journal of Developmental Psychology* 7: 641–652.
Murray, D., & Nahmias, E. (2014). Explaining away incompatibilist intuitions. *Philosophy and Phenomenological Research*, 88(2), 434-467.
Nadelhoffer, T., Rose, D., Buckwalter, W., & Nichols, S. (in press.) Natural Compatibilism, Indeterminism, and Intrusive Metaphysics. *Cognitive Science*.
Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs*, 41, 663-685.
Nussbaum, M. C. (2009). *Hiding from humanity: Disgust, shame, and the law.* Princeton University Press.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.
Patil, I., Zucchelli, M. M., Kool, W., Campbell, S., Fornasier, F., Calò, M., ... & Cushman, F. (2020). Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. *Journal of Personality and Social Psychology.*
Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science*, 36, 163–177.
Pellizzoni S., Siegal M. & Surian L. (2009) Foreknowledge, caring, and the side-effect effect in young children. *Developmental Psychology* 45: 289–295.
Pellizzoni, S., Siegal, M., & Surian, L. (2010). The contact principle and utilitarian moral judgments in young children. *Developmental Science*, 13(2), 265-270.
Rakoczy, H., Behne, T., Clüver, A., Dallmann, S., Weidner, S., & Waldmann, M. R. (2015). The side-effect effect in children is robust and not specific to the moral status of action effects. *PLoS ONE*, 10, e0132933.
Rohrer, D., Pashler, H., & Harris, C. R. (2015). Do subtle reminders of money change people's political views? *Journal of Experimental Psychology: General,* 144(4), e73.
Rawls, J. (1971). *A theory of justice*. Harvard University Press. Cambridge, MA.
Rose, D., Machery, E., Stich, S., Alai, M., Angelucci, A., Berniūnas, R., ... & Cohnitz, D. (2019). Nothing at stake in knowledge. *Noûs*, 53, 224-247.
Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology*, 76, 574.
Ruggeri, K., Alí, S., Berge, M. L., Bertoldo, G., Bjørndal, L. D., Cortijos-Bernabeu, A., ... & Gibson, S. P. (2020). Replicating patterns of prospect theory for decision under risk. *Nature Human Behaviour*, 1-12.
Samland, J., Josephs, M., Waldmann, M. R., & Rakoczy, H. (2016). The role of prescriptive norms and knowledge in children's and adults' causal selection. *Journal of Experimental Psychology: General*, 145, 125.

Sanchez, C., Sundermeier, B., Gray, K., & Calin-Jageman, R. J. (2017). Direct replication of Gervais & Norenzayan (2012): No evidence that analytic thinking decreases religious belief. *PLoS ONE*, 12(2).

Sarkissian, H., Chatterjee, A., De Brigard, F., Knobe, J., Nichols, S., & Sirker, S. (2010). Is belief in free will a cultural universal? *Mind & Language*, 25, 346-358.

Sarkissian, H., Park, J., Tien, D., Wright, J. C. & Knobe, J. (2011). Folk moral relativism. *Mind & Language*, 26(4), 482-505.

Schnall, S., Benton, J., & Harvey, S. (2008). With a clean conscience: Cleanliness reduces the severity of moral judgments. *Psychological Science,* 19, 1219–1222.

Seyedsayamdost, H. (2015). On normativity and epistemic intuitions: Failure of replication. *Episteme*, 12, 95-116.

Sommers, T. (2010). Experimental philosophy and free will. *Philosophy Compass*, 5(2), 199-212.

Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? The Life Outcomes of Personality Replication Project. *Psychological Science*, 30, 711-727.

Stich, S. & Tobia, K. P. (2015) Experimental Philosophy's Challenge to the 'Great Tradition' *Analytica: Revista de Filosofia*,

Schwitzgebel, E., & Cushman, F. (2015). Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition*, 141, 127-137.

Sinnott-Armstrong, W. (2008). Framing moral intuitions. In Sinnott-Armstrong, W. (Ed.). *Moral Psychology: Vol. 2. The cognitive science of morality*, 47–76. Cambridge, MA: MIT Press.

Swain, S., Alexander, J., & Weinberg, J. M. (2008). The instability of philosophical intuitions: Running hot and cold on truetemp. *Philosophy and Phenomenological Research*, 76(1), 138-155.

Tasimi, A., Gelman, S. A., Cimpian, A., & Knobe, J. (2017). Differences in the evaluation of generic statements about human and non-human categories. *Cognitive Science,* 41, 1934-1957.

Tierney, H., Howard, C., Kumar, V., Kvaran, T., & Nichols, S. (2014). How many of us are there? *Advances in Experimental Philosophy of Mind*, 181.

van Dongen, N., Colombo, M., Romero, F. Sprenger, J. (in press). Intuitions about the Reference of Proper Names: A Meta-Analysis. *Review of Philosophy and Psychology*.

Wagenmakers, E. J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R. B., ... & Bulnes, L. C. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11(6), 917-928.

Weinberg, J. M. (2009). On doing better, experimental-style. *Philosophical Studies*, 145(3), 455-464.

Weinberg, J. M., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics*, 29, 429-460.

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science*, 16, 780-784.

Wiegmann, A., Okan, Y., & Nagel, J. (2012). Order effects in moral judgment. *Philosophical Psychology*, 25, 813-836.

Wiegmann, A., Horvath, J., & Meyer, K. (2020). Intuitive expertise and irrelevant options. *Oxford Studies in Experimental Philosophy Volume 3*, 275.

Wiegmann, A., & Waldmann, M. R. (2014). Transfer effects between moral dilemmas: A causal model theory. *Cognition*, 131, 28-43.

Yaden, D. B. & Anderson, D. E. (in press). The Psychology of Philosophy: Associating Philosophical Views with Psychological Traits in Professional Philosophers. *Philosophical Psychology*.

Yang, F., Knobe, J. & Dunham, Y. (in press). Happiness is from the soul: Origins of an evaluative view of happiness. *Journal of Experimental Psychology: General*.

Yuan, Y. & Kim, M. (in press). Cross-cultural Convergence of Knowledge Attribution in East Asia and the US. *Review of Philosophy and Psychology.*

Zijlstra, L. (2019). Folk moral objectivism and its measurement. *Journal of Experimental Social Psychology*, 84, 103807.

Ziółkowski, A. (2019). The Stability of Philosophical Intuitions: Failed Replications of Swain et al. (2008). *Episteme*, 1-19.

Ziółkowski, Z., Wiegmann, A., Horvath, J. & Machery, E. (unpublished): Are intuitions about Truetemp unstable? A high-powered replication.