Difference and Robustness in the Patterns of Philosophical Intuition Across Demographic Groups[1]

Joshua Knobe

(In press at *Review of Philosophy and Psychology*)

ABSTRACT: In a recent paper, I argued that philosophical intuitions are surprisingly robust both across demographic groups and across development. Machery and Stich reply by reviewing a series of studies that do show significant differences in philosophical intuition between different demographic groups. This is a helpful point, which gets at precisely the issues that are most relevant here. However, even when one looks at those very studies, one finds truly surprising robustness. In other words, despite the presence of statistically significant differences between demographic groups, a core finding coming out of those studies is that philosophical intuitions are surprisingly robust across demographic groups.

Keywords: Experimental philosophy, cross-cultural, metaphilosophy

The emergence of experimental philosophy has led to a rapid growth in our understanding of the patterns of philosophical intuitions across demographic groups. We now have far more detailed knowledge than we did even a few years ago of precisely which types of philosophical intuitions differ across demographic groups, and of the size and direction of those differences. We also have far more detailed knowledge of the ways in which certain intuitions are robust across demographic groups. In many cases, experimental philosophers have identified a peculiar pattern of intuitions within a sample of American adults and then gone on to show that the same peculiar pattern arises in participants from numerous other cultures or in very young children.

Alongside this rapid growth in our empirical knowledge has arisen a lively metaphilosophical literature, written primarily by non-experimental philosophers, that aims to explore the deeper implications of these empirical findings. I have been very impressed with the philosophical arguments developed within this literature, but I sometimes worry that it mischaracterizes the empirical findings whose implications it aims to bring out.

Within the metaphilosophical literature, there is virtually no discussion at all of the many studies showing surprising robustness of intuition across groups. Instead, the discussion is entirely about the implications of studies showing *differences* between groups. This characterization makes it seem as though the main result coming out of actual empirical work on these issues in experimental philosophy has been about differences, with maybe a few occasional exceptions in which studies fail to find such differences.

[1] I am grateful to Edouard Machery and Stephen Stich for sharing an earlier draft of their reply and for helpful discussions of these issues. Thanks also to Ivar Hannikainen and Vilius Dranseika, who kindly provided detailed information on the methods and results of their studies.

This is not an accurate characterization of the field. In the earliest years of experimental philosophy, there were a few blockbuster studies showing extraordinary demographic differences, and these studies kicked off a trend within the metaphilosophical literature. However, as empirical research continued, we began to have a substantial body of studies demonstrating robustness. In such studies, it is not just that researchers fail to find differences between groups; they actually uncover truly surprising ways in which patterns of intuition observed within one group also arise, against expectation, in a variety of other groups.

In a recent paper, I therefore argued that it is time for a shift of focus within the metaphilosophical literature (Knobe, 2019). Existing empirical research, I suggested, indicates that philosophical intuitions are surprisingly robust across demographic groups, and we need to be thinking about the deeper metaphilosophical implications of this robustness.

Machery and Stich (2019) now offer a detailed and extremely helpful reply to that paper. They provide information about a large number of different studies indicating differences in philosophical intuitions between different demographic groups, and they make a variety of important and incisive comments about these differences. All of this is an extremely welcome corrective to what is lacking in the existing metaphilosophical literature. I have been troubled by a tendency to introduce broad generalizations about what experimental philosophy has shown without properly engaging with the details of the actual empirical findings. This reply does exactly the opposite; it helps to initiate a more serious, empirically grounded debate on precisely the issues we most need to be discussing here. Still, I am not entirely in agreement with the picture they paint of the present state of the field, and I thought it might be helpful to continue on with another round of this discussion.

In what follows, I will be presenting a great deal of detailed information about various studies and their results, but first, a brief clarification of the claim I am trying to defend. The basic form of the claim I am making is not that experimental philosophy has failed to find any evidence for something (e.g., that it has failed to find evidence for differences in intuition between demographic groups). Rather, the claim is that experimental philosophy actually *has* found evidence for something. Specifically, the claim is that experimental philosophy has found evidence for an extremely surprising and important tendency whereby intuitions observed in members of one demographic group tend to emerge also in members of other demographic groups.

This point comes out especially clearly in the context of comparisons between adults and children. Clearly, I am not suggesting that experimental philosophy has failed to uncover any evidence of differences in philosophical intuitions between children and adults. Rather, I am suggesting that experimental philosophy has found evidence for something very surprising. It has uncovered a truly shocking tendency whereby patterns of intuition observed in adults show up even in very young children.

Not only is this claim compatible with the claim that there are differences between children and adults, it seems clear that the very same result can provide strong evidence for both. That is, looking at a single experimental result, we might arrive at both of the following conclusions:

(a) This result indicates that the intuitions of young children are not exactly the same as the intuitions of adults.

(b) Nonetheless, the most important aspect of the result is precisely that it indicates that children show – to a truly surprising degree – the pattern originally observed in adults.

I assume that none of this will be controversial. This is the sort of thing one hears all of the time in developmental psychology.

The claim I am making about cross-cultural studies is of exactly the same form. The suggestion is not that experimental philosophy has failed to find any evidence of differences in intuition between different cultures. Rather, it is that experimental philosophy has successfully found evidence for something – an extremely surprising tendency whereby intuitions that were originally observed in one culture tend to show up in other, very different cultures.

In what follows, I proceed in two steps. First, I look at some of the recent studies providing evidence for robustness. Then, second, I turn to some of the studies that Machery and Stich cite as revealing differences between demographic groups. I argue that a more detailed examination of those studies reveals that, in many cases, the most important result coming out of them is actually the surprising level of robustness they reveal. In other words, even when we look at studies that do successfully find differences between demographic groups, sometimes the most surprising and interesting result coming out of those studies is the degree to which people from different demographic groups show the same intuitions. (I do not mean this as an exciting new revisionist take on existing findings. The points I will be making are very straightforward and very much in keeping with the ways in which these findings have been understood within existing work in the field.)

Before moving onward, I want to emphasize that nothing I say below should be taken as a criticism of the work of Machery and Stich. In my opinion, their studies are excellent, their review of the empirical literature does a great deal to move the conversation forward, and the more philosophical points they make in their paper are genuinely insightful and helpful. Of course, it seems that we have arrived at opposite opinions about this one question, but all the same, I very much look forward to further discussion with them on these issues.

Studies showing robustness

In my original paper, I argued that philosophical intuitions are surprisingly robust across demographic groups. I fear that the argument in that original paper may have been a bit too quick, so I'd like to start by presenting a few individual examples in more detail.

The heart of the argument is simple. Experimental philosophy has discovered many surprising patterns in people's intuitions. In some cases, subsequent studies have asked whether those same patterns arise in people from other cultures and in very young children. Again and again, such studies have found that the patterns do arise in these other populations. This is a very surprising

result, which calls for explanation at an empirical level and also for further reflection at a more purely philosophical level.

Part of the puzzle here arises from the nature of the effects explored in these studies. Suppose we discovered, for example, that people in all cultures believe that human beings are more obligated to care for their own children than for complete strangers. Although this might be an interesting fact, it surely would not come as a complete surprise. After all, this moral judgment plays an absolutely central role in the way people ordinarily live their lives. The key point now is that experimental philosophy studies are not like that. Very often, they are concerned with seemingly abstruse topics in metaphysics and epistemology, and within those topics, they typically uncover extremely quirky and unexpected patterns. Yet existing work indicates that even *those* patterns emerge within participants from other cultures and young children.

As an example, consider intuitions about *Gettier cases*. In a beautiful piece of cross-cultural research, Machery and colleagues (2017a) presented two different Gettier cases to participants from four different cultures. The key result was robustness. Participants from all four cultures showed a tendency to deny knowledge in Gettier cases. (For a beautiful replication and extension of these results, see Machery et al., 2017b.)

In informal conversations, I have sometimes heard people express disappointment about this result. Typically, the reaction goes something like this: It might have been interesting if there had been a large cross-cultural difference, but since all four cultures showed more or less the same pattern of intuitions, the actual results were not especially exciting or philosophically important. This reaction is, I think, completely mistaken. The Machery et al. result is a deeply interesting and important one. It shows robustness in a place we might not have expected to find it and thereby teaches us something important about people's philosophical intuitions.

Of course, one might argue that the Gettier intuition is a central aspect of our ordinary concept of knowledge and, hence, that this finding of robustness is not especially impressive. But the same basic finding arises for other effects that seem far less central. In studies with Western participants, Buckwalter (2014) uncovered a very surprising effect that is now known as the Gettierized Epistemic Side-Effect Effect (GESEE). At the essence of GESEE is an impact of moral considerations on people's willingness to ascribe knowledge in what appear to be Gettier cases.

Thus, suppose a mayor believes that the contract he is signing will have the side-effect of helping people. Unbeknownst to him, someone secretly swapped the contract for a completely different one, but – in the usual Gettier twist – it turns out, by sheer coincidence, that this other contract also ends up helping people. Did he know that he would be helping? Participants tend to say no. But what if we now change the case so that he believes that the contract will *harm* people (and actually does harm them)? In that latter version, participants tend to say yes (Buckwalter, 2014). The moral badness of the side-effect somehow leads them to ascribe knowledge even in what might seem to be a paradigm example of a Gettier case.

Yuan and Kim (in press) wanted to know whether this effect would arise as well for participants from other cultures. They therefore translated the original materials into Chinese and Korean and

ran the study on participants from China and South Korea. Quite strikingly, even this very quirky and surprising effect emerged in both of these other populations.

Yuan and Kim then obtained a similar result for an even more complex and arcane variation on this effect. Buckwalter (2014) also observes a Third-Person Gettierized Epistemic Side-Effect Effect. For example, suppose that a third-person observer is just watching events with the mayor unfold. The observer is not doing anything of any moral significance but is simply observing the mayor's actions. Does the *observer* know that the mayor will be helping or harming the people of the town? Here too, Buckwalter found a difference between help and harm in Western participants, and here again, Kim and Yuan found the same effect in China and South Korea.

I have been focusing thus far on robustness across cultures, but existing studies also find truly surprising levels of robustness when comparing adults to young children. In research with Western adults, I found a pattern whereby people's intuitions about whether an effect was brought about intentionally depend in part on the moral status of the effect itself (Knobe, 2003). Specifically, in cases where an agent knew that she was bringing about an effect but did not care at all about that effect, participants were more inclined to say that she brought about the effect intentionally when the effect was harmful than when the effect was helpful.

A question now arises as to whether children, too, would show the asymmetry. To address this question, Leslie and colleagues (2006) conducted a study with 3-5 year-old children. All participants were given a story in which a boy named Andy brings home a frog. Participants were randomly assigned to be told either that he knows that the frog will make his friend Janine happy or that he knows that the frog will make Janine upset. Either way, they were told that he does not care about the impact on Janine. The question then was whether Andy intentionally made Janine happy or upset.

The goal was to determine whether children believe that an agent can intentionally bring about an effect even if he or she does not care about that effect. For the study to answer this question, however, it was essential that children correctly understand that Andy does not care about the impact on Janine. Leslie et al. therefore introduced a test of this understanding. The test was extraordinarily simple. Children were told:

> Now listen very carefully. Andy does not care that Janine will be happy. He is going to bring the frog over just for himself.

They were then asked:

> Does Andy care that Janine will be happy?

In other words, children were directly told that Andy does not care and then asked whether Andy cares. Somewhat surprisingly, very young children show a strong tendency to respond to this question with the answer "Yes!" Indeed, the majority of three-year-old children gave this mistaken answer, and it was only at age four that children began answering the caring question correctly.

With this in the background, we can now explain the core finding of the study. Only by age four were children even able to understand the notion that an agent could help or harm without caring, but even at that age, they already showed a tendency to say that harming was more intentional than helping. In other words, as soon as children understand the idea that people sometimes don't care about effects, they show the help/harm asymmetry. (Subsequent developmental research has replicated and extended this finding; Michelin et al., 2010; Pellizzoni et al., 2009; Rakoczy et al., 2015.)

I have been dwelling on this one study because it nicely illustrates the sort of finding that has been obtained in other studies of the same basic type. In studies on adults, participants also show a help/harm asymmetry in judgments about causation (e.g., Hitchcock & Knobe, 2009) and about generics (Cimpian et al., 2010). Strikingly, both of these effects also emerge in young children. That is, young children show a help/harm asymmetry both for causal judgments (Samland, 2016) and for generics (Tasimi et al., 2017).

I have been dwelling on this one study because it nicely illustrates the sort of finding that has been obtained in other studies of the same basic type. In studies on adults, participants also show a harm/help asymmetry in judgments about causation (e.g., Hitchcock & Knobe, 2009) and about generics (Cimpian et al., 2010). Strikingly, both of these effects also emerge in young children. That is, young children show a harm/help asymmetry both for causal judgments (Samland, 2016) and for generics (Tasimi et al., 2017).

Existing research on adults has explored these effects in great detail and uncovered many more nuanced patterns within them. A key question for future research will be whether young children also show these more nuanced patterns. For one especially striking example, consider the harm/help asymmetry for generics (Cimpian, Brandone, & Gelman, 2010; Leslie, 2008). Suppose participants are told that there is a certain type of tool called a 'kren,' and that 30% of krens are dangerous. Participants then tend to agree with the generic:

(1) Krens are dangerous.

But if participants learn that there is a type of tool called a 'kren,' and that 30% of krens are wonderful, they are not as inclined to agree with the generic:

(2) Krens are wonderful.

So far, this is just yet another harm/help asymmetry. But here comes the twist. It turns out that this effect arises for categories of tools, and for categories of animals, but that it does not arise for categories of human beings. Thus, suppose participants are told that there is a type of *person* called a 'Kren' and that 30% of them are dangerous/wonderful. In that type of case, the entire effect disappears. When it comes to judgments about human beings, there is no harm/help asymmetry, meaning the participants are no more likely to agree with (1) than with (2) (Tasimi et al., 2017). It is not known why the asymmetry is limited in this way to categories of non-human objects.

Regardless, it turns out that even young children show this pattern. Not only do children show no harm/help asymmetry when they are given categories of nonhuman objects, they also show the characteristic tendency whereby the effect disappears when switching to categories of human beings (Tasimi et al., 2017). In other words, what we have here is a very complex pattern that cognitive scientists themselves struggle to understand, and yet young children seem to be showing it in the same way adults do.

I have been trying to illustrate the evidence for the robustness of philosophical intuition by picking out a very small number of examples and exploring them in at least a little bit of depth. It is important to note, however, that findings like these are coming out in the experimental philosophy literature on a very regular basis. To get a sense for the larger patterns, consider a few experimental philosophy projects that have been completed just in the months since my original paper was published.

⋆ In a series of studies on Western adults, Hannikainen and Donelson (2018) explored intuitions about whether laws, by their very nature, fulfill certain requirements (being public, being prospective, etc.). Participants in one condition were asked whether laws *actually do* have these qualities. For example, they were asked whether they agree with the statement: "Some laws are kept secret from those who are subject to them." Participants in the other condition were asked whether it was even *possible* for a law not to have these qualities. For example, they were asked whether they agree with the statement: "There could be laws that are kept secret from those who are subject to them." In an extremely surprising finding (which I certainly would never have predicted), the results indicated that participants were actually *more* inclined to think that some actual laws don't have these qualities than they were to think that it was possible for a law not to have these qualities. In my view, this seemingly paradoxical finding has important implications for the philosophy of law, but leaving those aside for the moment, I hope it is clear that the results obtained here are highly unexpected.

Hannikainen et al. (in press) asked whether these surprising patterns would be robust across cultures. To address this question, they ran the study in Brazil, Lithuania, Poland, the United States, Colombia, Netherlands and Spain. In all seven of these populations, participants showed an effect in the same paradoxical direction, agreeing more with the claim that some actual laws do not have these qualities than with the claim that it would be possible for laws not to have these qualities. Thus, whatever is driving this effect, it appears to be something that is found in numerous different cultures.

⋆ Tobia (2015, 2016) discovered an important asymmetry in ordinary intuitions about personal identity. Suppose that John gets into a car accident, suffers brain damage and is rushed to the hospital and undergoes a procedure. The person in the hospital bed after this procedure has a very different personality from John before the accident. Is this person still John? It turns out that people's answers depend on the direction of change. If the person after the accident is more morally good than the original John, participants tend to say that he is still John. By contrast, if the person after the accident is more morally bad, participants tend to say that John no longer exists and that the person after the accident is best understood as a different person entirely.

Dranseika et al. (unpublished data) ask whether this effect is robust across cultures. They ran the study in six different cultures, and here again, the results provided strong evidence of cross-cultural robustness. In five of the six cultures, there was a significant effect in the same direction found in the original Tobia study. In the sixth (the Netherlands), the effect was not significant, but was in the original direction. However, within that sixth culture, Dranseika et al. also recruited a sample of lawyers, and in that sample, the effect was significant. In short, Tobia's surprising asymmetry seems to be emerging in numerous different cultures.

⋆ Phillips et al. (2017) conducted a series of studies on ordinary intuitions about happiness. The obvious view would be that being happy is simply a matter of having certain psychological states (high positive affect, low negative affect, high life satisfaction). Phillips et al. found that participants' judgments were actually impacted as well by another factor, namely, *morality*. In other words, even when a person was described as having all of the psychological states associated with happiness, participants were reluctant to say that the person was truly happy if she had a morally bad life.

Yang et al. (2021) ask whether this effect is robust across demographic differences. Translating the vignettes into Chinese and running the study on Chinese participants, they find that the same effect emerges. More surprisingly, using a child-friendly version of the vignettes, they find the exact same effect emerging among four- and five-year-old children.

In sum, there appears to be a very general tendency whereby surprising effects that are observed in experimental philosophy studies on Western adults also emerge in studies of adults from other cultures and also in young children. This tendency calls for an explanation. That is, we need some account of why people of different demographic groups show such similar patterns of intuition even about questions that might seem highly abstract and removed from the practical questions we face in everyday life.

One obvious explanation would be that these patterns of intuition have an innate basis. I find this explanation to be plausible and promising, but surely it will be possible to come up with other plausible explanations as well. Future work should more directly pit different explanations against each other and look for empirical evidence that might support some over others.

In any case, it seems likely that the correct explanation – whatever that explanation turns out to be – will have important metaphilosophical implications. For example, suppose it turns out that our intuitions about personal identity are explained in part by innate features of our cognition. What would this tell us about the role these intuitions should play in philosophy?

Thinking about these matters even for a moment, one can easily see at least the outlines of arguments for a number of radically opposing views. On one hand, defenders of conceptual analysis might take innateness to be evidence that there really is a single unified concept of personal identity, widely shared by people of many different cultures, and that ordinary intuitions give us insight into that concept. On the other, philosophers who are more skeptical of the role of intuition (such as Machery and Stich themselves) could argue that these intuitions would have seemed a lot more trustworthy if they had been acquired through a process of learning or reflection rather than being the product of our innate cognitive endowment. After all, if intuitions

about these metaphysical questions are indeed products of innate mechanisms, it seems highly likely that such metaphysical intuitions are best understood as simply being side-effects of mechanisms that exist primarily to achieve some other function. At this point, I don't have any definite views about which answer to this question is correct – my point is just that this is the sort of question we need to be asking.

There is a large literature by non-experimental philosophers about the meta-philosophical implications of experimental philosophy, and within this literature, there has been a great deal of discussion of studies that look at the pattern of philosophical intuitions across demographic groups. Yet none of this literature has explored any of these questions. Indeed, as far as I can tell, there is not a single paper within this literature that explores the meta-philosophical implications of the many studies showing robustness across demographic groups. Instead, the entire literature on this topic seems to be concerned with the implications of studies finding *differences* between demographic groups. I have sometimes even heard it said that the principal finding of the entire field of experimental philosophy is the degree to which intuitions differ across such groups. As a researcher who spends most of his time engaged in the day-to-day work of doing experimental philosophy, I find the portrayal of my field within this literature to be completely unrecognizable.

In my view, this strand of meta-philosophical research needs to change in a very serious way. It should not be considered acceptable to have a literature on the implications of patterns of intuition across demographic groups that simply ignores all of the findings pointing to robustness.

Clarification

Thus far, I have been discussing a range of studies that are widely seen as providing evidence for the claim that people's philosophical intuitions are robust across demographic groups. It is important to emphasize, however, that the claim here does not involve a failure to find statistically significant differences between groups. On the contrary, some of the studies I have been discussing *do* find differences between groups. In other words, these studies do find differences but nonetheless provide evidence for robustness. I assume that this initial point will not be terribly controversial, but it is worth making it explicit, just because I will be using a similar form of argument below in cases where it might be less straightforward.

First, when we look to studies of children, we typically do not find that the pattern of responses given by children is identical to the one given by adults. Rather, there are systematic and statistically significant differences between the children and the adults in almost all such studies. These differences are of course important in themselves and very much worth exploring further. The point is just that even if there is indeed a significant difference between children and adults, it can easily turn out that the most noteworthy result is the degree to which participants from the two different populations respond similarly. This sort of point will play a major role in the argument I develop in the next section.

Second, even in studies that finds that an effect emerges in a number of different cultures, it is sometimes possible to identify *a specific individual vignette* for which the effect does not emerge

in all cultures. For example, I noted above that the Yuan and Kim study finds the Third-Person Gettierized Epistemic Side-Effect Effect in both Korean and Chinese participants. Interestingly, however, the study also finds a cross-cultural difference. The study used a series of different vignettes – one involving a mayor, a second involving a pump, a third involving air quality, and so forth. For most of these vignettes, both Korean and Chinese participants showed the effect observed among American participants. However, for the vignette about the mayor, Korean participants showed the effect, but Chinese participants did not. This cross-cultural difference is not just noise. Yuan and Kim were able to replicate the finding that Chinese participants do not show the effect for that one vignette in the way that American and Korean participants do.

In short, these results do show a cross-cultural difference, but it is not a general difference in the way people from different cultures think about Third-Person Gettierized Epistemic Side-Effects. Instead, it seems to involve a difference in the way people from different cultures think about the events described in this one specific vignette (e.g., the way they think about mayors). This sort of point, too, will play a major role in the argument of the next section.

Studies showing effects of demographic factors

Although I am not entirely in agreement with the claims Machery and Stich defend in their reply, I strongly believe that the reply itself helps to move the larger discussion in precisely the right direction. I have been disappointed that discussions of these issues in the meta-philosophical literature are becoming ever more disconnected from the details of the actual empirical research. This reply does a great deal to remedy that problem. It addresses head-on the empirical questions I regard as most important, and it does a wonderful job of exploring the available data in real detail.

Machery and Stich make a number of excellent points, all of which help to move the conversation forward in very helpful ways. For present purposes, however, it might be helpful to divide the points into three broad families.

1. Machery and Stich considerably expand the scope of the inquiry. In my original paper I focused on demographic groups in the usual, relatively narrow sense (race, gender, age, nationality, etc.). In their reply, Machery and Stich point out that there are also differences in intuition due to other individual differences in personality and cognitive style (working memory capacity, disgust sensitivity, cognitive reflectiveness, etc.). Everything that Machery and Stich say about this strikes me as completely correct.

However, I had never meant to deny that these other individual difference variables impact philosophical intuitions, and I'm not sure that these claims bear in any way on the argument I have been making. The argument is: (a) Philosophical intuitions are extremely robust across cultures and across development. (b) This is a surprising finding, which cries out for explanation. (c) The obvious explanation is that these intuitions are the product of innate cognitive mechanisms. (s) Though other explanations are possible, it seems likely that the correct explanation, whatever it turns out to be, will have important meta-philosophical implications. The key issue now is whether facts about effects of these other individual difference variables bear in any way on this argument.

To get a better handle on this issue, it might be helpful to consider one of the actual findings. A number of recent studies find a correlation between judgments about philosophical questions and scores on the cognitive reflection test. This test aims to measure something about people's ability to overcome their immediate intuition and instead arrive at a judgment based on reflection (Frederick, 2005). In cases in which people face a conflict between intuition and reflection, some people show a greater tendency to decide that their immediate intuitions were actually mistaken and to go instead with a conclusion they arrive at through reflection. Those who show this tendency also tend to show a distinctive profile of philosophical judgments. They are more likely to give orthodox responses in Gettier cases (Machery et al., 2017), more likely to give consequentialist responses to trolley problems (Patil et al., 2020), and more likely to give incompatible responses when asked about a deterministic universe (Hannikainen et al., 2019).

The most natural way of understanding these correlations is that they point to the presence of two conflicting processes in people's philosophical judgments. For example, the correlation observed for incompatibilism is most naturally interpreted as follows: When people receive a case about a deterministic universe, the judgments they make are the result of two conflicting processes. A more intuitive process is drawing them toward the conclusion that people in the deterministic universe can still have free will, while a more reflective process is drawing them toward the conclusion that people in this universe cannot have free will. The judgment they make in the end depends in part on which of these two processes emerges victorious.

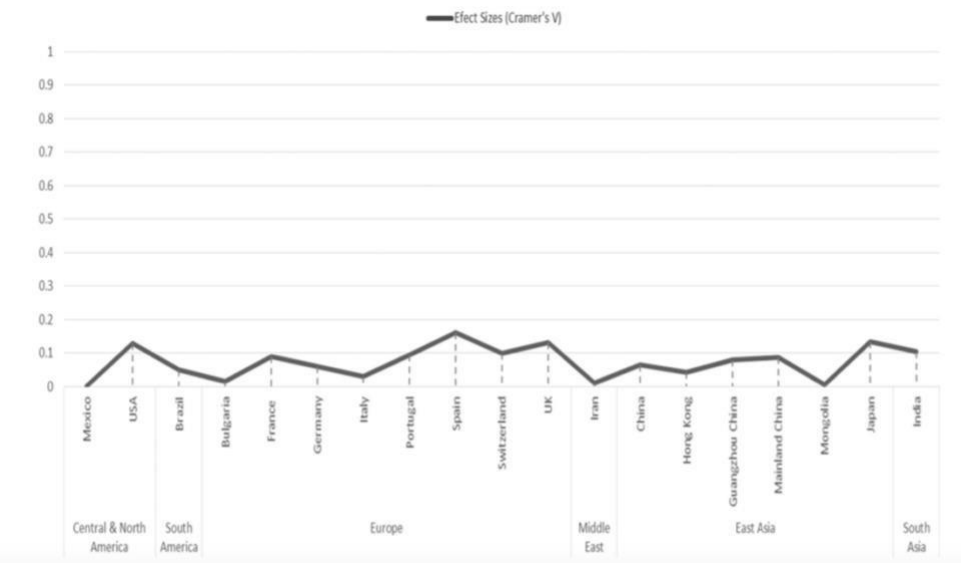
Notice now that, just taken in itself, this sort of claim does not provide evidence either for the conclusion that the underlying representations are learned or for the conclusion that they are innate. For example, if we find that people in Western cultures experience a conflict in these cases, what we are learning is just that there is a conflict between two representations within the minds of Western participants. Just taken in itself, this would not tell us anything about whether the representations involved in the conflict were learned or innate.

What we need to know, then, is *whether the correlation itself is robust across cultures*. If we look at other cultures, do we find that philosophical judgments are correlated with individual different variables in the same way? To address this question, Hannikainen and colleagues (2019) conducted a cross-cultural study, and – very strikingly – they find that the same correlation actually arises among Asian participants. In other words, among Asian participants, there is also a tendency whereby those who are higher in cognitive reflection are more likely to give incompatibilist responses. This result indicates not only that the pattern of intuition is robust across cultures but also that the underlying psychological process might be robust across cultures. That is, we are getting at least some initial indication that, across a number of different cultures, people are drawn by a more intuitive process toward compatibilist responses and by a more reflective process toward incompatibilist responses.

2. Machery and Stich look at studies that compare the intuitions of philosophers with those of non-philosophers. Perhaps the most salient example here are the many studies on the relationship between stakes and knowledge attributions. Philosophers generally report having intuitions about knowledge that depend in a certain way on the stakes, but experimental philosophy research

generally finds that ordinary folks often have intuitions that are not affected by stakes in this way (Buckwalter, 2010; Buckwalter & Schaffer, 2015; Feltz & Zarpentine, 2010; May, Sinnott-Armstrong, Hull & Zimmerman, 2010; but see Pinillos, 2012; Sripada & Stanley, 2012). Thus, the intuitions of professional philosophers about these cases appear to depart quite systematically from those of ordinary folks. Machery and Stich point out that this should itself be regarded as a demographic effect on philosophical intuitions. This is a very important point, which I had previously failed to appreciate. I fully concede that Machery and Stich are right about this.

This point then interacts in an interesting way with the idea I was exploring above. For example, Rose and colleagues find that folk intuitions across a wide variety of different cultures do not show an effect of stakes. Below are their results:



Here the size of the effect is measured on a scale that goes from 0 to 1. Across numerous different cultures, one finds that the effect of stakes on knowledge attributions is quite low. As Rose et al. put it, there is a "surprising stability in the lack of a stakes effect across sites" (Rose et al., 2019:234).

What exactly is it that makes this stability so surprising? At least to a large extent, the answer lies in the very fact that folk intuitions from all of these different populations are similar to each other but also different from a widely held view among philosophers. In other words, it is in part the difference between the folk and the philosophers that makes the robustness among the folk so striking and interesting.

3. Finally, Machery and Stich note that even if we confine ourselves to demographic differences in the narrow sense, existing studies have identified a wide variety of cases in which demographic differences are associated with differences in philosophical intuition. In particular, Machery and Stich cite studies showing differences between demographic groups in intuitions about intentional action, trolley problems, free will, relativism, and a number of other topics. Here again, Machery and Stich make a valuable contribution to the discussion.

To this last point, I make two responses. First, it is surely the case that there are at least *some* intuitions that show substantial demographic effects. I have been suggesting that philosophical intuitions are generally surprisingly robust across demographic differences, but there is no real chance that it will turn out that *all* philosophical intentions are robust in this way. We will certainly be able to find at least some cases in which intuitions do vary quite a bit between demographic groups. When we do identify these cases (as, inevitably, we will), a question will arise as to why these cases differ from all of the cases we have been discussing thus far.

Here is one possible hypothesis: It might be that the reason there is so little impact of culture on the intuitions we have been discussing thus far is that we have been focusing on intuitions about questions that people hardly ever discuss explicitly. Thus, when you are bringing up a child, it is highly unlikely that you would ever bring up the topic of third-person Gettierized epistemic side-effect cases and try to convey your culture's distinctive view on those cases. By contrast, people often talk quite explicitly with their children about morality and about religion. It therefore seems plausible that cultural learning might directly impact people's intuitions about those topics in a way that it does not seem to be impacting the sorts of intuitions we've been discussing thus far. This is an important topic for further research.

Secondly – and this is a response we will be exploring in more detail – I will argue that people's philosophical intuitions are surprisingly robust across demographic groups even when it comes to the specific topics Machery and Stich mention in their paper. The issue is not that there are no real differences between demographic groups on those topics; such differences absolutely do exist, and Machery and Stich have done a wonderful job of identifying them. Rather, the issue is that when we look in detail at how those differences work, we come away with a deeper sense of the surprising robustness of intuitions across demographic groups. Indeed, many of the studies that do show significant differences between demographic groups also provide support for the claim that intuitions are surprisingly robust across such groups.

The only way to support this claim is to get into the nitty-gritty details. In what follows, I therefore discuss the details of three specific cases in which existing research finds significant differences in philosophical intuition between different demographic groups.

Relativism

People from Western cultures often hold the view that there is no objective truth concerning certain moral or aesthetic questions. This view is sometimes known as "relativism," and it is frequently derided in rather contemptuous tones by professional philosophers. It is widely associated with watered down postmodern thought or with various forms of flakiness or general hippy-dippyness.

A question now arises as to whether relativist intuitions are found robustly across cultures. Before moving onward, it is worth emphasizing how plausible it would be to claim that this whole phenomenon is culturally local. On one hand, it seems unlikely, at least a priori, that features of our basic cognitive endowment would incline us toward any specific view about

abstract metaphysical issues like this one. On the other, it is an obvious and undeniable fact that relativist views are promulgated quite explicitly within Western culture, and any typical Western undergraduate will have had some exposure to such views. Thus, a natural hypothesis would be that the relativism found in Western cultures does not reflect anything general about human cognition but simply reflects the popularization of a distinctively Western cultural outlook.

In one of the earliest studies on folk relativism, Sarkissian and colleagues (2011) looked at intuitions about status of moral judgments in a sample of American undergraduates. Participants in the 'same-culture' condition were asked to imagine that two different Americans held opposite views about whether a specific token action was morally bad. They were then asked whether, given that the two had opposite views, one of them had to be wrong or whether they could actually both be right. Participants in the 'other-culture' were asked to imagine that an American person held one view about whether the action was morally bad, while someone from another culture held the opposite view. Here again, they were asked whether one of these people had to be wrong or whether they could both be right. The results showed a significant difference between conditions. Participants were more inclined in the other-culture condition than in the same-culture condition to deny that one of the people had to be wrong.

So, is this result just the product of the sort of postmodernism or multiculturalism that is so salient in contemporary American campuses? To address this question, Sarkissian et al. ran the same study in Singapore. The results showed precisely the difference between same-culture and other-culture conditions that had been observed in the American undergraduates. This result provides evidence that even certain aspects of the detailed pattern of relativist intuitions are surprisingly robust across cultures.

In their recent paper, Machery and Stich point out that studies of relativism about aesthetic judgments actually do find a significant difference between cultures. This is completely correct, and it is a finding worth discussing further.

Cova and Pain (2012) conducted a series of studies about whether people see aesthetic judgments as relative or objective. All participants were undergraduates at a French university. Participants were asked to consider something that they themselves regarded as beautiful and then to imagine that someone else held the opposite opinion. They were then asked how to characterize the disagreement:

a. One of the interlocutors is right while the other is wrong.
b. Both are right.
c. Both are wrong.
d. Neither is right or wrong. It makes no sense to speak in terms of correctness in this situation. Everyone is entitled to his/her own opinion.

In the sample of French undergraduates, participants tended to choose option d, indicating that participants from that specific population reject the idea of objective truths in this type of case.

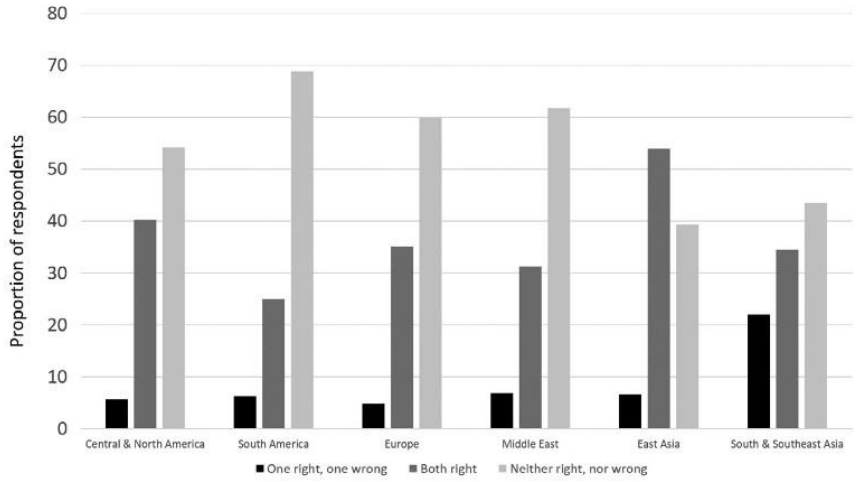
As part of a large cross-cultural project led by Machery and Stich, a study by Cova and colleagues (2019) looked at whether that same pattern of responses would emerge across

cultures. The study used a method derived from Cova and Pain, but this time with over 2,000 participants from 19 different countries. The resulting paper is an absolutely fantastic contribution to the experimental philosophy of aesthetics, and I am delighted to have an opportunity to write about it here.

The response options this time were:

1. One of you is correct while the other is not.
2. Both of you are correct.
3. Neither is correct. It makes no sense to talk about correctness in this situation.

Shown below is the proportion of participants choosing each of these options, broken down by world region.



The thing Cova et al. emphasize about these results is that the intuitions of ordinary folks are so deeply different from the views philosophers most typically hold on these issues. I completely agree that this is the most important upshot.

But consider also the pattern across the different cultures. The study clearly does find a difference between cultures, but all the same, the more striking result is the degree to which participants from all these different cultures show such a *similar* pattern of responses.

In saying this, I don't take myself to be making a controversial claim. I am simply agreeing with what Cova et al. say about their own results. As they put it: "However, despite these differences, there is a consistent pattern: Answer 1 is the least chosen option in all geographic areas. Even in the South and Southeast Asia sample, the proportion choosing Answer 1 was significantly below chance... So, although there is important variation across geographic areas in the percentage of people choosing Answer 1, it nonetheless remains the least chosen answer" (Cova et al., 2019:326).

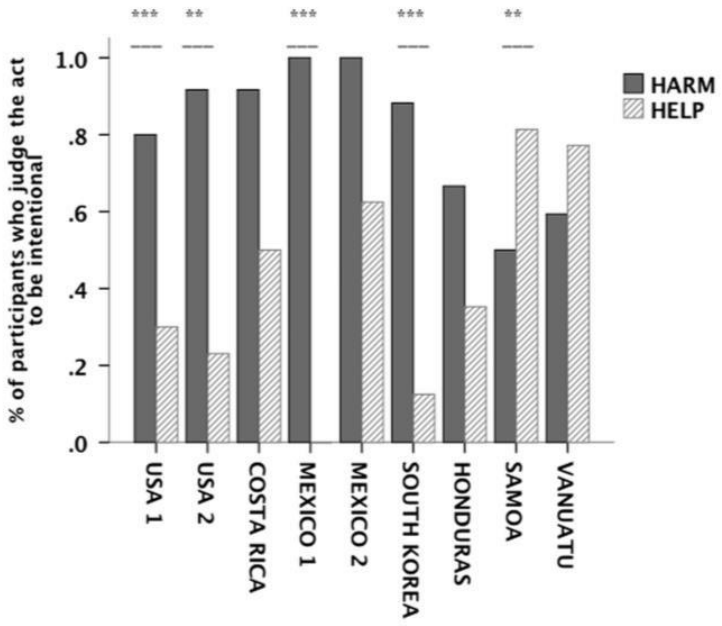
We are therefore left with an important further question: *Why are the intuitions of participants in these very different cultures so extraordinarily similar?* The fact that people's intuitions in these cases depart from those of philosophers only deepens the mystery. The view philosophers tend to hold on this issue is a simple and straightforward one, and yet, somehow, people from all these different cultures depart from that simple and straightforward view in the same direction. We need some explanation of this surprising cross-cultural robustness.

Intentional action

In an earlier section, I used intuitions about intentional action as an example of robustness, noting that even young children show the pattern observed in adult participants. Machery and Stich helpfully point out, however, that existing studies have found substantial cross-cultural differences in intentional action intuitions. This is an excellent point, and one that is worth exploring further.

In an important cross-cultural study, Robbins and colleagues (2017) looked at intuitions about intentional action in 11 different cultures. All participants were assigned either to a case in which the side-effect was helping or a case in which the side effect was harming. Some participants were told that the agent was a CEO (in the United States, Mexico, South Korea, and Honduras), while others were told that the agent was a High Chief (Samoa, Vanuatu, and Costa Rica).

Here are the results:



Clearly, what we have here is a dramatic cross-cultural difference. Not only are there cultures that do not show the pattern observed in American participants, there are cultures that actually show a significant effect in the opposite direction.

But of course, we immediately face a question as to *why* this cross-cultural difference is emerging. Is it that people from certain cultures simply do not show the help/harm asymmetry observed in American participants? Or might the difference here be due to something specific to this one vignette?

Robbins et al. decided to explore the latter possibility. They hypothesized that the cross-cultural difference might not be a matter of how people from different cultures think about intentional action but might instead reflect something about how people from certain cultures think about High Chiefs. More specifically, they hypothesized that people from certain cultures show a distinctive deference toward High Chiefs and would therefore tend to say that a High Chief acted intentionally even when producing help as a side-effect.

To test this hypothesis, Robbins et al. conducted a follow-up study in which Samoan and NiVanuatu participants were randomly assigned to receive either the original High Chief case or a case that was exactly the same except that the agent was described as a commoner. The results showed an extremely strong effect of this manipulation, with participants being far more inclined to regard the help as intentional in the High Chief condition than in the commoner condition. In light of this finding, Robbins et al. conclude that their results, taken as a whole, actually provide evidence *in favor* of the hypothesis that people's basic criteria for intentional action are robust across cultural differences.

In sum, this cross-cultural study yields two very striking findings. First, there is a cross-cultural difference whereby participants in some cultures show a deference to people in positions of authority that is not shown in Western culture. Second, the cross-cultural difference does not seem to involve a difference in people's intuitive criteria for intentional action. Those criteria themselves appear to be surprisingly robust across cultures.

Free will

In my original paper, I used intuitions about free will as an example of a case in which experimental philosophy research genuinely has uncovered important cross-cultural differences. Let's now examine this case in more detail so that we can get a better sense of what things look like when there is, in fact, a major effect of demographic factors.

A large body of research in experimental philosophy has explored people's intuitions about whether free will is compatible with determinism. Research on participants from Western cultures indicates that they are pulled in different directions by different cognitive processes. It appears that there is something within these people's minds drawing them to the view that no one can have free will in a deterministic universe (incompatibilism), but also something within their minds drawing them to the view that even agents in a deterministic universe can have free will (compatibilism). Different researchers have proposed radically different hypotheses about what the relevant cognitive processes actually are (Knobe, 2014; Murray & Nahmias, 2014; Rose et al., 2017), but the details of these different hypotheses will not concern us here.

A question now arises about whether this experimental philosophy research is uncovering something that is robust across cultures. One could imagine a whole continuum of different

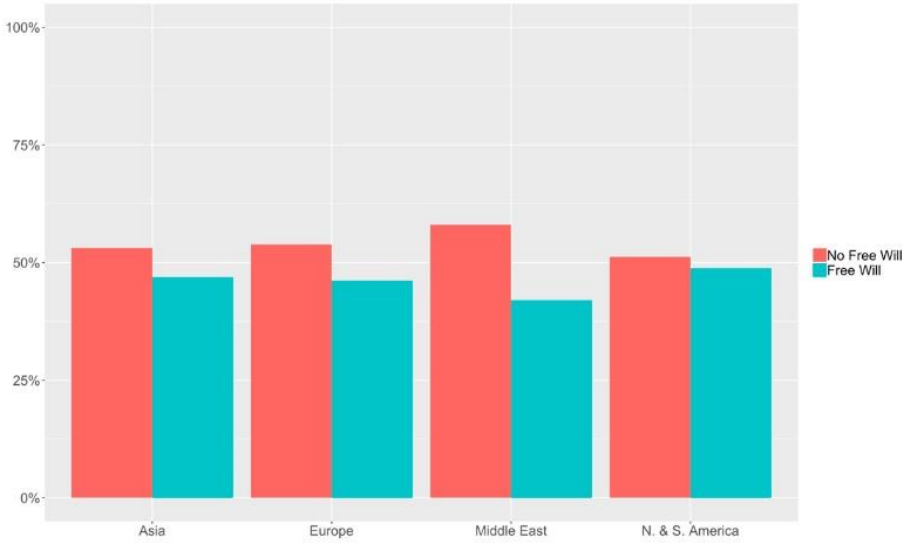
possible answers to this question. On one extreme, it might be thought that the conflict and confusion that one finds among Western participants is due to something culturally local. For example, we might find that there is something drawing people from Western cultures to incompatibilism but that people from Asian cultures will not be drawn to incompatibilism and will therefore see determinism as a non-issue. If this turns out to be the case, people from Asian cultures would, at least to a large degree, tend to think that the so-called "problem of free will" was not actually a problem at all.

Moving now to the opposite extreme, it might be thought that the exact processes that draw Western participants in these two conflicting directions are robust across cultures. For example, it might be thought each of these two processes has an innate basis. On this second view, the problem of free will should itself be a cultural universal, in the sense that people from numerous different cultures should experience the same conflict and confusion one finds in participants from Western cultures.

The best evidence regarding this question comes from a paper by Hannikainen and colleagues (2019), which was part of the massive cross-cultural research led by Machery and Stich mentioned above. The paper is truly a tour de force: 5,268 participants, 21 different locations, from 4 continents. This is a model of what cross-cultural research in experimental philosophy should look like.

The paper uses a complex design in which some participants receives a question about a deterministic universe, while others a question about a Frankfurt-style case. The analysis then asks whether the difference between judgments about these two different cases is itself different in different cultures. The results show that there is indeed a cross-cultural difference in this difference. Participants from Asia show a substantially smaller difference between judgments about a deterministic universe and judgments about a Frankfurt-style case than do participants from other world regions. This is an important finding, which genuinely does provide evidence for a cross-cultural difference in philosophical intuition.

But suppose we focus just on intuitions about whether it is possible to have free will in a deterministic universe. The researchers very helpfully made all of their data freely available, and I was therefore able to put together a simple visualization of the result for that one question. Here are the percentages of participants who ascribed free will to an agent in a deterministic universe, broken down by world region.



What we see here is a truly shocking degree of similarity across cultures in judgments about the compatibility of free will and determinism. The pattern observed among American participants seems to arise, in almost exactly the same form, for participants from numerous other cultures. The results thereby provide at least some initial evidence for the view that the conflict and confusion people experience about the problem of free will is itself robust across cultures.

To really put this issue to the test, one would want to look at cross-cultural robustness of more detailed findings from existing experimental philosophy research. In support of their hypothesis about the process driving incompatibilist intuitions, Murray and Nahmias (2014) conduct a series of studies that demonstrate a phenomenon they call "bypassing." In defense of my own hypothesis, I show that the Murray-Nahmias bypassing effect arises only for intuitions about human beings and not for intuitions about computers (Knobe, 2014). Regardless of whether either of our hypotheses turns out ultimately to be correct, it does seem likely that these two findings teach us something about the processes drawing people in conflicting directions. One way to further explore these phenomena would therefore be to ask whether these findings also emerge in participants from other cultures.

Before moving on, let me briefly clarify my remarks here. I am not trying to diminish or call into question the cross-cultural difference observed in the Hannikainen et al. paper. Quite the contrary: I specifically picked out this paper because it is a beautiful example of an experimental philosophy study that *does* find a cross-cultural difference. The point I am making is that, even in this case where experimental philosophy research has indeed found a cross-cultural difference, the pattern of the results also provides evidence of surprising cross-cultural robustness.

Conclusion

I have been reviewing evidence for the claim that philosophical intuitions are surprisingly robust across demographic differences. This evidence comes from studies that find no significant

differences of intuitions between cultures, from studies comparing children to adults, and even from studies that do find significant differences between cultures. Regardless of what one thinks about the larger theoretical issues, it would be hard to deny that there is *a lot* of evidence for robustness.

One obvious conclusion to draw from this evidence would be that philosophical intuitions are indeed surprisingly robust across demographic differences. My own view is that this is precisely the right conclusion to draw, but I'm sure some readers will arrive at other, quite different conclusions. In this final section, I therefore want distinguish between two different ways of rejecting the conclusion for which I have been arguing thus far.

The first would be to *argue* against this conclusion. For example, it would be possible to argue that some of the evidence that seems to suggest robustness should instead be interpreted in a different way. Machery and Stich's reply develops this sort of argument in some very helpful directions. I was especially impressed with the point they make about differences between philosophers and non-philosophers. Just as they say, some of the studies that find surprising robustness in folk intuitions across cultures also show that the intuitions of philosophers differ dramatically from ordinary folk intuitions. This is an important point.

Similarly, it would be possible to argue that although philosophical intuition happens to be surprisingly robust in the specific case studies I have explored here, there are other cases in which philosophical intuitions vary dramatically by demographic group. This, too, is a point that Machery and Stich pursue very nicely in their reply. I would be extremely open to further discussion of this type of argument as well.

But all of these the very helpful arguments and objections should be distinguished from the approach that has become prevalent within the meta-philosophical literature. Within that literature, it has become common practice just to *completely ignore* all of the evidence for robustness. This approach should not be regarded as at all legitimate or acceptable. If the goal of this literature is to explore the implications of experimental philosophy, it will need to engage far more directly with the findings coming out of actual experimental philosophy research.

References

Buckwalter, W. (2010). Knowledge isn't closed on Saturday: A study in ordinary language. *Review of Philosophy and Psychology*, 1, 395– 406.

Buckwalter, W. & Schaffer, J. (2015). Knowledge, stakes and mistakes. *Nous*, 201– 234.

Buckwalter, W. (2014). Gettier made ESEE. *Philosophical Psychology*, 27(3), 368-383.

Cimpian, A., Brandone, A. C., & Gelman, S. A. (2010). Generic statements require little evidence for acceptance but have powerful implications. *Cognitive Science*, 34, 1452-1482.

Cova, F., & Pain, N. (2012). Can folk aesthetics ground aesthetic realism? *The Monist*, 95(2), 241-263.

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.

Donelson, R., & Hannikainen, I. (2018). Fuller and the Folk: The Inner Morality of Law Revisited. *Oxford Studies in Experimental Philosophy*, 3.

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

Feltz, A., & Zarpentine, C. (2010). Do you know more when it matters less? *Philosophical Psychology*, 23, 683-706.

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

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 press). Are there cross-cultural legal principles? Modal reasoning uncovers procedural constraints on law. *Cognitive Science*.

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.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy*, 106(11), 587612.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, 63, 190-194.

Knobe, J. (2014). Free will and the scientific vision. *Current Controversies in Experimental Philosophy*, 69-85.

Knobe, J. (2019). Philosophical Intuitions Are Surprisingly Robust Across Demographic Differences. *Epistemology & Philosophy of Science*, 56(2).

Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect: Theory of mind and moral judgment. *Psychological Science*, 17(5), 421-427.

Leslie, S. J. (2008). Generics: Cognition and acquisition. *Philosophical Review*, 117, 1-47.

Machery, E., Stich, S., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., ... & Hashimoto, T. (2017a). 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.

Machery, E. & Stich, S. (2019). Demographic Differences in Philosophical Intuition: A reply to Joshua Knobe. Unpublished manuscript.

May, J., Sinnott-Armstrong, W., Hull, J. G., & Zimmerman, A. (2010). Practical interests, relevant alternatives, and knowledge attributions: an empirical study. *Review of Philosophy and Psychology*, 1(2), 265– 273.

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.

Myers-Schulz, B., & Schwitzgebel, E. (2013). Knowing that P without believing that P. *Noûs*, 47(2), 371-384.

Murray, D., & Nahmias, E. (2014). Explaining away incompatibilist intuitions. *Philosophy and Phenomenological Research*, 88(2), 434-46.

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*.

Pellizzoni S., Siegal M. & Surian L. (2009) Foreknowledge, caring, and the side-effect effect in young children. *Developmental Psychology* 45: 289–295.

Phillips, J., De Freitas, J., Mott, C., Gruber, J., & Knobe, J. (2017). True happiness: The role of morality in the folk concept of happiness. *Journal of Experimental Psychology: General*, 146(2), 165.

Pinillos, Á. (2012). Knowledge, Experiments and Practical Interests. In (eds.) Brown & Gerken, *Knowledge Ascriptions* (pp. 192-221). Oxford University Press.

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(7), e0132933.

Robbins, E., Shepard, J., & Rochat, P. (2017). Variations in judgments of intentional action and moral evaluation across eight cultures. *Cognition*, 164, 22-30.

Rose, D., Machery, E., Stich, S., Alai, M., Angelucci, A., Berniūnas, R., ... & Cohnitz, D. (2019). Nothing at stake in knowledge. *Noûs*, 53(1), 224-247.

Rose, D., Buckwalter, W., & Nichols, S. (2017). Neuroscientific prediction and the intrusion of intuitive metaphysics. *Cognitive Science*, 41(2), 482-502.

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(2), 125.

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

Sripada, C., & Stanley, J. (2012). Empirical Tests of Interest-Relative Invariantism. *Episteme* 9, 3-26.

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(7), 19341957.

Tobia, K. P. (2015). Personal identity and the Phineas Gage effect. *Analysis*, 75(3), 396-405.

Tobia, K. P. (2016). Personal identity, direction of change, and neuroethics. *Neuroethics*, 9(1), 37-43.

Yang, F., Knobe, J. & Dunham, Y. (2021). Happiness is from the soul: Origins of an evaluative view of happiness. *Journal of Experimental Psychology: General*, 150, 276–288

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