Diametros 17 (2020), 64: 72–89
doi: 10.33392/diam.1499

X-Phi and Impartiality Thought Experiments: Investigating the Veil of Ignorance

– Norbert Paulo –
– Thomas Pölzler –

Abstract: This paper discusses "impartiality thought experiments", i.e., thought experiments that attempt to generate intuitions which are unaffected by personal characteristics such as age, gender or race. We focus on the most prominent impartiality thought experiment, the Veil of Ignorance (VOI), and show that both in its original Rawlsian version and in a more generic version, empirical investigations can be normatively relevant in two ways: First, on the assumption that the VOI is effective and robust, if subjects dominantly favor a certain normative judgment behind the VOI this provides evidence in favor of that judgment; if, on the other hand, they do not dominantly favor a judgment this reduces our justification for it. Second, empirical investigations can also contribute to assessing the effectiveness and robustness of the VOI in the first place, thereby supporting or undermining its applications across the board.

Keywords: veil of ignorance, thought experiments, empirical ethics, experimental philosophy, Rawls, moral epistemology, impartiality, moral point of view

Published online: 12 May 2020

Ethicists and political philosophers regularly make use of thought experiments. Consider the following two famous examples:

- *Trolley*: A runaway trolley heads towards five workers. If nobody interferes it will kill them. Would it be appropriate to divert the trolley onto a different set of tracks where there is only one worker?[1]
- *Pond*: A child is about to drown in a pond. You can rescue it, but this would ruin your shoes and upset the plans that you had for this day. Ought you rescue the child?[2]

Norbert Paulo
Law School, University of Salzburg, Austria
Department of Philosophy, University of Graz,
Attemsgasse 25, 8010 Austria
email: norbert.paulo@uni-graz.at
Thomas Pölzler
Department of Philosophy, University of Graz,
Attemsgasse 25, 8010 Austria
email: thomas.poelzler@uni-graz.at

[1] Foot (2003); Thomson (1976).
[2] Singer (1972).

The precise philosophical function of thought experiments such as these is sometimes unclear.[3] On the face of it, most thought experiments in ethics and political philosophy seem to have been used to *justify* normative claims. Both *Trolley* and *Pond*, for example, might be claimed to provide reasons in favor of utilitarianism. The philosophically interesting second part of *Pond* starts from the assumption that most of us find it unacceptable to allow the child to drown and asks, why many also believe that it is acceptable to spend one's disposable income on luxury products even if this money could effectively feed a starving child in the developing world. Singer suggests utilitarianism as the appropriate position to make sense of the fact that both cases seem to be morally similar. Singer's and other thought experiments have sometimes also been used to *illustrate* abstract arguments or to function as heuristics that generate novel normative claims which are then tested independently. We will come back to these interpretative issues later.

Recently, empirical researchers (psychologists, cognitive scientists, experimental philosophers, etc.) have become increasingly interested in normatively motivated thought experiments as well. A number of studies in particular have attempted to answer two kinds of empirical questions. Focusing on the thought experiments' outcome, researchers have tested to what extent people share the intuitions that philosophers had expressed about these thought experiments (e.g., the intuition that it would be appropriate to divert the trolley or that we ought to rescue the child). Focusing on the underlying cognitive processes, many studies have also addressed the ways in which these intuitions come about (e.g., by emotion or by reason), focusing in particular on the reliability of these processes.

The results of empirical studies on thought experiments are interesting for psychological reasons. Moreover, some scholars have argued that they can also be of normative relevance.[4] In his famous 2008 paper Joshua Greene, for example, famously compared responses to *Trolley* to responses to a similar scenario (*Footbridge*) in which one can only save the five railroad workers by pushing a large man off a footbridge onto the tracks. Subjects' dominantly deontological intuitions about this latter case were found to be caused by emotional aversions against personal (as opposed to impersonal) harm. As the personal/impersonal distinction seems morally irrelevant, Greene argues that we should therefore regard "characteristically deontological" moral judgments as being unlikely to track the moral truth.[5]

So far both empirical research on thought experiments and philosophical research on the normative implications of this empirical research has focused on cases like *Trolley* and *Pond* – thought experiments that have been developed to justify or illustrate very specific normative claims. But ethicists and political philosophers sometimes use thought experiments of a different kind too. These thought experiments serve methodological purposes and can hence be appealed to in arguments for all kinds of normative claims on all levels of abstraction. We have in mind here thought experiments like the following:

[3] See Brun (2017).
[4] Rini (2013); Kumar, Campbell (2012).
[5] Greene (2008). On different versions of Greene's argument see Paulo (2018).

1. John Rawls' *Veil of Ignorance*
2. David Hume's and Adam Smith's *Impartial Spectator*
3. Jean-Jacques Rousseau's *State of Nature*
4. Jürgen Habermas' *Ideal Discourse*
5. Ronald Dworkin's *Hercules*

It is difficult to pinpoint what, if anything, unites this group of thought experiments (apart from their multi-purpose character). Most plausibly, it seems that some of them centrally (1. and 2.), and others at least importantly (3.-5.) serve the purpose of increasing impartiality. The assumption is that in thinking about normative issues we should not favor ourselves or our loved ones. This impartial standpoint – which is sometimes also referred to as the "moral point of view"[6] – is most effectively achieved by disregarding a number of contingent facts about us that seem to be morally irrelevant (e.g., our age, gender, or race). In different ways, the above thought experiments all attempt to enable and motivate such a disregard for these morally irrelevant facts.

As indicated, empirical and empirically-informed research on impartiality thought experiments has so far been rather sparse. Our paper is an attempt to remedy this situation. We will try to investigate whether and in what way empirical investigations of impartiality thought experiments could be normatively relevant. In doing so, we will exemplarily focus on the veil of ignorance (VOI) thought experiment, as it is arguably the most famous and widely used instance of impartiality thought experiments.[7] First, we will address ways in which outcome-focused empirical investigations of this thought experiment could be normatively relevant. Then we will turn to process-focused research. Finally, we will briefly discuss a general objection against the normative relevance of empirical findings about the VOI. It will turn out that there are several ways in which such findings can be important for a number of debates in ethics and political philosophy.

Outcome-Focused Investigations

The most obvious way to empirically investigate any version of the VOI is to test whether it dominantly generates the intuitions the philosopher who originally proposed it thought it would. In what follows we will discuss the normative relevance of such outcome-focused empirical investigations of the VOI. First, we will do so with regard to the thought experiment's original formulation by John Rawls (henceforth: *rVOI*). Then we will consider a generic idea of the VOI, as it has been used by many subsequent political philosophers and ethicists (henceforth: *iVOI*).

[6] For discussion, Jollimore (2018), sec. 2.
[7] See, e.g., Brun (2017).

Rawls' VOI

The rVOI was first proposed by Rawls in his landmark work *A Theory of Justice*.[8] It was developed as part of a broader thought experiment known as the original position (which involves a situation in which parties are to agree on the principles of justice that will determine the basic structure of their society)[9].

According to Rawls, we should accept those principles of justice that would be chosen by rational representatives of free and equal persons in the original position. These representatives are characterized in specific ways. Among others, they are said to be rational (which includes endorsing principles of rational choice and having an idea of a "rational plan of life", i.e., of what defines a good life for them), mutually disinterested (i.e., they are self-interested but free from envy; they don't strive to be better off than others for its own sake), to have a capacity for reasonableness (i.e., a conception of right beyond rationality) and a sense of justice (i.e., the willingness and desire to comply with the demands of justice).

The rVOI is the most important distinguishing feature of the original position. It was supposed to add a "strong impartiality condition".[10] In particular, Rawls asks us to imagine the representatives in the original position as not knowing many things they normally do know, namely a number of contingent facts that Rawls deems to be morally irrelevant. What are these facts? At the very least the veil is supposed to shield the representatives from knowledge about specific facts about themselves such as their age, gender, race, wealth, health, religious or political convictions, etc. (so-called "thin" VOI). In addition, the representatives are also said to not know basic facts about their society – how wealthy it is, which resources it has, what its population is, etc. (the so called "thick" VOI).[11]

We are now in a position to look at the ways in which one can (or cannot) derive normatively significant conclusions from empirical research on the rVOI. The main idea is that the right principles of justice are those that people choose if they engage in the thought experiment. Rawls used this idea non-empirically. When invoking the rVOI, he simply assumes that others would agree with him on his well-known principles of justice. The obvious empirical question to ask is of course: Is this assumption correct? Do people really decide in the way imagined by Rawls when he designed the rVOI thought experiment?

Some researchers have already addressed this question. The most influential outcome-focused experimental investigation of the rVOI is Frohlich and Oppenheimer's.[12] In their experiments Frohlich and Oppenheimer asked subjects to discuss which of four distributional principles to accept. The main result was that most groups agree on principles that maximize average income with a floor constraint that guarantees a certain minimal income. Crucially, no group of test subjects opted for one of the principles Rawls thought would be chosen behind the VOI, namely his well-known difference principle.

[8] Rawls (2005).
[9] For an overview see Hinton (2016).
[10] Freeman (2006): 154.
[11] On Rawls' use of these notions see Gaus, Thrasher (2016).
[12] Frohlich, Oppenheimer, Eavey (1987); see also Frohlich, Oppenheimer (1993).

This result was later successfully replicated,[13] and other researchers have developed similar experimental designs to investigate other aspects of the rVOI.[14]

The philosophical relevance of the above-mentioned empirical investigations seems to be somewhat limited. Among others, this is because the operationalization of the rVOI diverged substantially from how this thought experiment was stated by Rawls.[15] For instance, in most studies subjects have only been explicitly veiled from their (class of) income, (not their gender, race, etc.), and income was taken as a proxy for all other primary goods, including non-material goods such as the social bases of self-respect.[16] Many studies also did not experimentally isolate the VOI. They rather investigated decisions in (approximated) original positions as a whole.[17] Finally, subjects were often asked to make an *actual* decision (that sometimes had real consequences within the experiment) rather than to think about how they or someone else *would* decide; and they were asked which distributive principle *they themselves* (would) prefer, rather than which principles conformed to the preferences of *rational representatives of free and equal persons*.[18]

Yet, even though the available empirical evidence on the rVOI may be of limited normative relevance, it is clear that such research could principally be highly relevant. Suppose a number of studies managed to show which principles of justice people would actually choose behind different approximations to the rVOI, each modelled as closely as experimentally feasible after certain aspects of Rawls' particular understanding of the rVOI. The results of such studies could confirm or disconfirm Rawls' assumption that most people would agree with him on his well-known principles of justice. If Rawls was right to make this assumption, this would support his use of the rVOI in arguing for his principles of justice. But if people did not agree on his principles, this would undermine his use of the rVOI in support of these principles (without thereby necessarily debunking Rawls' principles of justice, because they could well be justified by ways other than the use of the rVOI). Both results would be highly significant for Rawls scholarship.

Critics might object that empirical investigations of the rVOI necessarily miss their target because Rawls never used this thought experiment in an epistemic sense, i.e., as an attempt to provide reasons for certain normative claims. Rather, the rVOI was meant to be a mere illustration of the abstract arguments for his preferred principles of justice. After all, Rawls described the rVOI as a "device of representation, or alternatively, a thought experiment for the purpose of public- and self-clarification."[19] And Freeman explains, "[t]his means that its purpose is not to impose an obligation on us that we do not already have. Its purpose rather is to elucidate the reasons behind our considered convictions of justice."[20]

[13] E.g., Aguiar, Becker, Miller (2013); Bond, Park (1991); Lissowski, Tyszka, Okrasa (1991); Chan (2005).
[14] E.g., Herne, Mard (2008); Herne, Suojanen (2004); Michelbach, Scott, Matland et al. (2003); Bruner (2018); Wolf, Dron (2015); Wolf, Lenger (2014).
[15] See, e.g., all of the references provided in the last three footnotes.
[16] E.g., Bruner (2018); Frohlich, Oppenheimer (1993); Wolf, Dron (2015).
[17] E.g., Bruner (2018); Frohlich, Oppenheimer (1993).
[18] E.g., Bruner (2018); Frohlich, Oppenheimer (1993).
[19] Rawls (2001): 17.
[20] Freeman (2006): 144.

We acknowledge the importance of this point. On the other hand, however, there are also good reasons to believe that the rVOI does have an epistemic function. For, were it merely an illustration it could not play the role it is supposed to play in Rawls' overall theory of justification, "reflective equilibrium".[21] In reflective equilibrium, the features of the choice situation (including the rVOI), together with elements of rational choice theory, and the candidate principles of justice have to be balanced against each other in search for overall coherence. It is the coherence between all of these elements that justifies every single one of them.[22] Moreover, in the general literature on thought experiments most authors ascribe an epistemic[23] or heuristic[24] role to the rVOI, and those who do not ascribe an epistemic role to the rVOI also typically don't understand it as a mere illustration.[25]

Here we do not want to commit ourselves to either the illustration or the epistemic understanding of the rVOI. Above we showed how empirical research on the thought experiment can turn out normatively significant on the basis of the epistemic understanding. We will also focus on this understanding in the following sections. However, even if one assumes that the rVOI merely has an illustrative function there is a limited way in which empirical investigations of it can gain philosophical importance. For instance, and most importantly, if it could be shown that, behind the rVOI, people in fact choose principles different from those Rawls argues for, then one could at least conclude that the rVOI is a bad illustration of Rawls' more abstract reasoning. Undermining this illustrative function of the rVOI would mean that the thought experiment does not aid understanding, and would reduce the rhetorical force and intuitive appeal of Rawls' argumentation.[26]

The „Idea" of the VOI

After Rawls had first introduced the VOI the thought experiment has been taken up by many other political philosophers and ethicists. These scholars mainly have not used the VOI in the exact same way as Rawls.

In what follows we will refer to these subsequent usages of the VOI as the *idea of the VOI* (iVOI). The iVOI has been spelled out in many different ways, depending on the particular argumentative goals of the philosophers using it. What most importantly distinguishes the iVOI from the rVOI are the following typical features: (1) The iVOI has not only been applied to judgments about distributive justice (for which the rVOI was designed), but also to judgments about many other normative matters. (2) Above we

[21] For a recent overview see Cath (2016).
[22] For a detailed discussion of this view see Hübner (2017).
[23] E.g., Gähde (2000); Celikates (2012).
[24] Gendler (2007); Miscevic (2017).
[25] See Cohnitz (2005): 145–52; Brun (2017): 201. Unlike the epistemic and the illustrative, the *heuristic* function of ethical thought experiments is open-ended. When thought experiments are used heuristically, they are not intended to provide reasons for or against a theory or to illustrate it. Rather, they should help to generate new hypotheses about or to get a better understanding of the implications or differences between moral statements, principles or theories – see Paulo, Pölzler (under review); Brun (2017).
[26] On the importance of understanding in contrast to knowledge see Stuart (2017).

noted that the rVOI may be interpreted in both an illustrative and an epistemic sense. The iVOI, in contrast, has mostly been used epistemically. (3) The iVOI does not include all the other conditions of the Rawlsian original position, with all the specific criteria for the representatives mentioned above (e.g., the idea of a "rational plan of life," their mutual disinterestedness, etc.). (4) The iVOI is mostly only a "thin" VOI or even only part of a "thin" VOI. That is, one is asked to imagine a lack of knowledge about (some) morally irrelevant personal characteristics, but not about one's society. (5) The rVOI is supposed to prompt third-person thinking. Those who engage in the thought experiment think about the choice that would be made by rational representatives of free and equal persons. The iVOI, in contrast, usually asks which normative judgments we ourselves would endorse.

A particularly influential example of this epistemic use of the iVOI is Joseph Carens' discussion of the ethics of migration:

> I gladly concede that I am using the original position in a way that Rawls himself does not intend, but I think that this extension is warranted by the nature of the questions I am addressing and the virtues of Rawls's approach as a general method of moral reasoning. […] Those in the original position would be prevented by the "veil of ignorance" from knowing their place of birth or whether they were members of one particular society rather than another.[27]

So far only few researchers have explicitly set out to empirically investigate versions of the iVOI (though some of the explicitly Rawls-inspired investigations that we cited above may actually be understood in such a way). One notable recent exception is a study by Karen Huang, Joshua Greene and Max Bazerman, who say that they "depart from the conventional use of the veil of ignorance as a device for thinking about the general organization of society. Instead, we apply veil-of-ignorance reasoning to a set of more specific moral and social dilemmas."[28] This study investigates how test persons respond to moral dilemmas from behind versions of the iVOI, i.e., it employs the iVOI to run thought experiments such as *Trolley* that have been developed to justify or illustrate very specific normative claims. However, we are not aware of any empirical investigation of the iVOI that was designed to test whether or not political philosophers or ethicists were right to assume that behind their iVOI people would agree to the particular normative claim they defend. For example, would people behind Carens' iVOI really opt for open borders?[29]

Empirical research on the iVOI can have important normative implications. Of course, as the iVOI is significantly different from the rVOI, these implications will not directly pertain to Rawls scholarship. They can at best yield very indirect evidence as to his theory of justice. However, empirical studies on versions of the iVOI can help to assess the normative arguments of those who have proposed these versions. After all, like Rawls with the rVOI, these political philosophers and ethicists have assumed that behind their iVOI people would agree to a certain normative claim. So if studies show

[27] Carens (1987): 257.
[28] Huang, Greene, Bazerman (2019).
[29] We will come back to the study by Huang, Greene, Bazerman (2019) later.

that there is indeed wide agreement of this kind this would support their argument; if there isn't such agreement, this would undermine these scholars' reliance on the argument from iVOI (without necessarily debunking their normative claim, because this claim could be justified in ways other than by using the iVOI as well).

To illustrate, let's look at an influential example of the use of the iVOI from bioethics, namely the debate about John Harris' critique[30] of the idea of quality adjusted life years (QALYs) between Harris[31] and Peter Singer, John McKie, Helga Kuhse and Jeff Richardson[32]. This debate is centrally about whether or not the idea of QALYs would be chosen behind the VOI. Here is how Singer, McKie, Kuhse and Richardson use the idea of the VOI; it is worth quoting at length:

> So, in this case, we imagine people choosing a basis for allocating health care without knowing whether, at some point in their lives, they will be in need of health care to prolong their lives; we imagine also that they do not know whether, if this happens, they will be among those whose interest in continued life is low, or among those whose interest is high. How would two *rational egoists* choose if they were faced with a situation in which they each needed life-saving treatment, and each had an interest in continued life, but there was enough lifesaving treatment for only one? Obviously each would choose the treatment for herself, if he or she could; but suppose they had to make the choice behind a veil of ignorance, in which *they knew the details of the two patients' conditions, but did not know which patient they were*? […] To maximise the satisfaction of their own interests, rational egoists would have to choose a system that gives preference to saving life when it is most in the interests of the person whose life is saved. This means that if QALYs were an accurate way of measuring when life is most in one's interests, then rational egoists would choose to allocate in accordance with QALYs.[33]

As this quote makes clear, the authors merely set two conditions for the people who are supposed to decide behind the (very "thin") iVOI: (1) these people are imagined to be rational egoists and (2) they are imagined to know "the details of the two patients' conditions, but … not … which patient they" would be.

Empirical studies could show how people actually decide behind the iVOI about the particular question under consideration (be it the QALYs proposal or any other question). This could confirm or disconfirm the philosophers' assumption that people would agree with them, and hence support or undermine their normative arguments. For example, if most subjects of a study indicated that in the above health care allocating case (where they know the details of the two patients' conditions, but do not know which patient they are) they would indeed choose an allocation system in accordance with the idea of QALYs, this would undermine Harris' particular argument.[34]

[30] Harris (1987).
[31] Harris (1995; 1996).
[32] Singer, McKie, Kuhse et al. (1995); McKie, Kuhse, Richardson et al. (1996b; 1996a).
[33] Singer, McKie, Kuhse et al. (1995): 148, italics ours.
[34] For a recent example of the use of the iVOI in the context of healthcare see Fritz, Cox (2019); for the use of rational choice procedures similar (but not quite identical) to the VOI see Żuradzki (2014).

Other than direct empirical investigations of the "thick" rVOI, direct investigations of very "thin" versions of the iVOI are much more likely to really hit their normative target. So this limitation of the philosophical significance is much less severe here. Take, again, the debate about QALYs. Since these bioethicists rely on the much simpler iVOI and not on the rVOI, it is easier to design experiments that approximate their imagined choice situation behind the VOI. The conditions mentioned in iVOI scenarios are often far less demanding than those imagined by Rawls.

Process-Focused Investigations

In the last section we showed how empirical research on the VOI may be directly relevant to justifying (and illustrating) normative claims. In what follows we will propose an argument that, to our knowledge, has so far hardly been made at all.[35] We will attempt to show that investigations of the VOI may also be normatively relevant by shedding light on the VOI as a method of justification.

Several philosophers have discussed whether and in what formulations the VOI is a proper way of justifying normative claims. For example, John Harsanyi has criticized Rawls' design of the original position for its decision theoretical assumptions.[36] There is reason to believe that empirical investigations can contribute to such methodological debates about the VOI. In particular, these investigations may help assessing both to what extent the VOI is *effective* (in achieving its purported aim) and to what extent it generates normative intuitions that are *robust*.

Effectiveness

One important criterion for thought experiments is that they are effective, i.e., that they fulfill the function that they were supposed to fulfill by those who developed and use them. Above we argued that one of the potential functions of the rVOI as well as the dominant function of the iVOI has been to justify normative claims by means of approximating an impartial standpoint. To what extent, then, does the VOI succeed in promoting impartiality? Several kinds of empirical studies and results could contribute to answering this question.

People are generally rather partial in making moral judgments.[37] Suppose, then, that we conduct an empirical study that involves the following two conditions: a condition in which subjects arrive at a normative judgment in an ordinary way (non-veiled), and a condition in which they are first asked to disregard their gender, race, income, etc. (veiled). If it turned out that judgments in the veiled condition do not statistically significantly differ from judgments in the non-veiled condition (which, supposedly, reflect a fair amount of self-interest) this would suggest that the VOI is ineffective as a method of justification. All else equal, this would in turn suggest that the thought experiment likely fails to lead people towards a more impartial point of view.

[35] Except, mostly in little detail, in the context of some empirical work, e.g., by Aguiar, Becker, Miller (2013).
[36] Harsanyi (1955; 1975).
[37] See, e.g., Bocian, Wojciszke (2014); DeScioli, Massenkoff, Shaw et al. (2014).

From the armchair, we consider it unlikely that making people engage in the VOI thought experiment will not have any effect at all. There is also initial empirical evidence which lends support to this suspicion. In their above-mentioned study Huang et al. found that iVOI reasoning does change the content of subjects' judgments about moral dilemmas. It made these judgments more consequentialist than they otherwise would have been.[38]

Suppose it was true that the VOI affects the content of people's normative judgments. The next question to ask in assessing its effectiveness would then be to what extent this effect is due to increasing impartiality (as opposed to unintended influences). There are several ways to approach this question experimentally. Perhaps most naturally and simplest, one could test for correlations of subjects' normative judgments with certain demographic characteristics such as their gender, race, income, etc. The lower these correlations turn out in the veiled condition, compared to the unveiled condition, the more effective the VOI can be considered to be (in the absence of plausible alternative explanations for the lower correlations).

Imagine, for example, that subjects are asked to select one of several principles of distribution. It is plausible that in the non-veiled condition subjects with low income will tend to choose a principle that favors recipients with low income (e.g., a strictly egalitarian principle), while subjects with high income will tend to choose a principle that favors recipients with high income (e.g., a libertarian principle). The study would show the VOI to be effective with regard to income to the extent to which this correlation between income level and choice of principle weakens in the veiled condition, with there being no plausible non-impartiality explanations for this weakening.

If the VOI did not show any effect at all or if its effect were not in any way due to increasing impartiality this could have important philosophical implications. Arguments that rely on the respective versions of the VOI as a means of justification might be epistemologically challenged. After all, if people are generally rather partial,[39] if impartiality is crucial for normative justification, and if the VOI does not increase impartiality then why

[38] Huang, Greene, Bazerman (2019). Note that Huang et al. used a form of veiling that is somewhat different from the Rawlsian idea. Following Caspar Hare (2016) they did not ask participants to disregard their actual identity (as the Rawlsian veil would have it), but to imagine that they do not know who they would be among those affected by the decision under consideration. That is, they designed the VOI conditions as situations of "risk" as compared to "ambiguity". For instance, in the *Trolley* scenarios, there is always a *known* chance of 1 out of 6, 1 out of 9 and so on to be the person who gets killed for the others to be saved. This might make an important difference. Take *Footbridge*: many people might feel very tempted to say that the person should be pushed off the bridge when they know they have a 5 out of 6 chance to be saved. However, when they don't know the odds, they might fear that the odds are much worse. So this difference between ambiguity and uncertainty might have significant effects; interpreting the VOI in one way rather than another might thus change the effects of VOI reasoning.

Also, the difference between the Rawlsian understanding of the VOI without known odds and Huang et al.'s understanding with known odds can be taken to reflect the fact that Rawls' aim was to construct principles of justice for the basic structure of society, whereas Huang et al.'s aims are much more mundane. An approximation of Rawls' procedure might look like this: Instead of a 5 out 6 chance to be saved (in a society of 1000 people, say), you would have a chance of 5 out of 1000 to be saved, and a 1 out of 1000 chance to be killed – simply because dilemmatic situations are so rare.

[39] See again, e.g., Bocian, Wojciszke (2014); DeScioli, Massenkoff, Shaw et al. (2014).

ought we to believe that the thought experiment confers justification on its outcomes? All else equal, these outcomes would not seem more justified than if they arose from processes that were not meant to increase impartiality at all.

That said, from the armchair we would not only expect that the VOI causes *some* difference in normative judgments (compared to non-veiled conditions) but also that this effect is *at least partly* explained by increasing impartiality.[40] In this case the empirical research's normative implications would depend on its exact findings.

Most importantly, it is not clear *how much* the VOI would need to increase impartiality in order for it to count as effective. If the difference between the veiled and the non-veiled condition would only be, say, 0.1 on a 7 point-scale of rightness/wrongness this would clearly be insufficient. But what about a difference of 0.3 or 0.5 or 0.7? Would that be enough? In our view, questions like this cannot be answered in the abstract. They require a number of judgments that depend on the particular experimental design of the relevant empirical studies (including, most importantly, their ways of measuring impartiality), the sizes of obtained effects, the content of the prompted normative judgments and the particular version of the VOI that are at issue. It also leads to a more general question about the effectiveness that philosophers expect from tools such as impartiality thought experiments.

One way of getting a better grasp of the VOI's effectiveness might be to compare it to the effectiveness of other impartiality thought experiments, such as the *Impartial Spectator*.[41] Does the VOI increase impartiality to a (significantly) higher or lower extent than these other thought experiments? For example, if we find that subjects in a study's veiled condition make *somewhat* more impartial normative judgments than subjects in a non-veiled condition but *much* less impartial judgments than subjects in an impartial spectator condition this may again give rise to a challenge. We might want to declare the VOI to be a comparatively ineffective thought experiment, thereby removing some evidentiary weight from arguments that involve it as a (important) component. At the very least we could conclude that, with regard to the requirement of impartiality, VOI-based arguments are weaker than arguments that are based on the impartial spectator thought experiment.

So far there is only one empirical study that has tested aspects of the VOI's effectiveness relative to other impartiality thought experiments. Fernando Aguiar, Alice Becker and Luis Miller had subjects choose among different distributions of goods. They found that the iVOI, an impartial spectator method (inspired by Smith) and an involved spectator method (inspired by Scanlon) led subjects to prefer substantially different distributional outcomes. In particular, the iVOI and the involved spectator method turned out to be less effective in bringing about impartiality than the impartial spectator method.[42]

[40] See also, e.g., the study by Aguiar, Becker, Miller (2013) discussed below.

[41] On Smith's understanding of the spectator see Raphael (2007).

[42] Aguiar, Becker, Miller (2013). In a recent study Bruner and Lindauer (2018) compared the VOI to the impartial spectator too. However, they did not investigate which of these thought experiments is more effective in increasing impartiality but rather which of them people find more appropriate in terms of determining the justice of principles.

Robustness

Impartiality thought experiments should not only be effective; their results should also be sufficiently robust. That is, the normative judgments that people arrive at by engaging in these thought experiments should not be (easily or overly) influenced by irrelevant factors.

In the recent past more and more research has converged on the hypothesis that normative judgements sometimes lack in robustness. For example, several studies suggest that the way in which information is framed (e.g., as *saving* four persons' versus *killing* one person in *Trolley*) influences subjects' responses.[43] Effects have also been found for the order in which scenarios are presented (if *Trolley* is presented prior to *Footbridge*, rather than subsequently, subjects tend to respond differently),[44] and for incidental emotions such as disgust and happiness (subjects who were primed with these emotions tended to make harsher/less harsh judgments than subjects who were not primed in this way).[45]

To be sure, the effects that were found by many of these studies are rather small.[46] Some of them have also been subject to methodological worries.[47] Yet, the extant research on normative robustness certainly makes it plausible that applications of the VOI (and other impartiality thought experiments) *could* be sensitive to irrelevant factors. For example, it *could* be the case that subjects make different normative judgments depending on whether they are first told to abstract from their age and only then from their gender versus *vice versa*.

So far there has not been any empirical research on the robustness of the VOI (or other impartiality thought experiments) at all. But such research could have important normative implications. If it turned out that certain versions of the VOI considerably lack in robustness—for example, by being subject to order or framing effects—this would again decrease the evidentiary support that these versions lend to normative judgments that have been claimed to result from them. We might end up having less reason to believe Rawls' theory of justice, the QALY-proposal, etc.

Whether and to what extent such conclusions can indeed be drawn from findings of non-robustness will again depend on matters of judgment.[48] How were the respective irrelevant factors measured? How large, exactly, was their effect on subjects' normative judgments? At what point do such influences start to get epistemically problematic? Proponents of VOI arguments might also respond that at least they themselves are likely to be able to avoid being influence by irrelevant influences (especially once common instances of them have been identified and made public). After all, as philosophers they regularly engage in thought experiment, are trained in counteracting their biases, etc. This potential objection brings us to the last section of our paper.

[43] E.g., Petrinovich, O'Neill (1996).
[44] E.g., Petrinovich, O'Neill (1996); Schwitzgebel, Cushman (2012); Wiegmann, Okan, Nagel (2012).
[45] E.g., Schnall, Haidt, Clore et al. (2008); Schnall, Benton, Harvey (2008); Valdesolo, DeSteno (2006).
[46] Demaree-Cotton (2016); May (2018).
[47] Landy, Goodwin (2015); Pölzler (2018).
[48] See, e.g., Demaree-Cotton (2019); Paulo (2020).

The Expertise Objection

All extant empirical research on the VOI has involved lay people, i.e., people without any special philosophical expertise. For example, Frohlich et al.'s original study deliberately tested students that had not been exposed to Rawls or theories of distributive justice in any way;[49] and Aguiar et al.'s study[50] tested undergraduate students from different disciplines. Up to this point we have assumed that under certain conditions empirical research on lay people's intuitions about the VOI can be normatively relevant. But is this really the case?

From its very beginning one of the main objections against experimental philosophy has been that lay people's intuitions are too unreliable for them to be possibly philosophically relevant.[51] The above-mentioned normative implications of empirical investigations of the VOI might be doubted on the very same grounds. Critics may argue that the only or main reason why such investigations do not result in subjects endorsing their favored normative judgment (in the case of outcome-focused investigations) or do not show the VOI to be sufficiently effective or robust (in the case of process-focused investigations) is that subjects lacked in expertise. If they had been experts – e.g., philosophers, moral philosophers (those who do and were meant to perform VOI thought experiments)[52], or "competent judges" (as Rawls imagined them for reflective equilibrium as a method of justification)[53] – then the investigations would have yielded much more favorable results.

This objection deserves serious consideration. Yet, we are not convinced that it succeeds in undermining the normative relevance of empirical investigations of the VOI. Such a conclusion seems doubtful for at least three reasons. First, on several plausible understandings of expertise, and with regard to several philosophical issues, it was found that philosophers may in fact fail to (considerably) outperform lay people in terms of their expertise. They are similarly susceptible to order effects, framing effects, etc.[54] It is thus plausible that applications of the VOI (and other impartiality thought experiments) would yield similar robustness-results across the alleged expert/non-expert divide.

Second, once normative or philosophical expertise has been defined it may be possible to (partly) operationalize it. For example, studies may involve certain kinds of attention checks, comprehension checks, requests for verbal explanations of responses, etc. If a subject performs badly on these criteria it may then be excluded from analyses on ground of lacking in expertise. In this way studies' results would only reflect the judgments of experts (whatever population makes up the sample of the study; whether it involves lay people, philosophers or others).[55]

[49] Frohlich, Oppenheimer, Eavey (1987): 611.
[50] Aguiar, Becker, Miller et al. (2013).
[51] Kauppinen (2007); Williamson (2008).
[52] See Kauppinen (2018).
[53] See Greenspan (2015).
[54] E.g., Löhr (2019); Machery (2017); Schwitzgebel, Cushman (2012); Tobia, Buckwalter, Stich (2013) Buckwalter, and Stich 2013.
[55] See, e.g., Nadelhoffer, Feltz (2008); Sytsma, Livengood (2015); Pölzler, Wright (2019).

Finally, suppose critics were right that only (moral) philosophers' responses to thought experiments are to be trusted. Even in this case empirical studies could still make important contributions to assessing VOI-based normative arguments. After all, there is nothing that prevents researchers from investigating philosophers in the way that they did or might investigate lay people. Philosophers could be presented with the very same materials and asked the very same questions. (This response to the expertise objection has recently gained prominence with regard to other kinds of philosophical issues as well.[56]

Conclusion

Empirical researchers have recently studied lay people's responses to a number of philosophical thought experiments. One kind of thought experiment that has so far only received little attention are impartiality thought experiments, i.e., thought experiments that attempt to generate intuitions which are unaffected by personal characteristics such as age, gender or race. In this paper we focused on the most prominent impartiality thought experiment, the VOI. We showed that both in its original Rawlsian version and in more widespread generic versions empirical investigations can be normatively relevant.

There are two ways in which empirical investigations of the VOI turned out to be normatively relevant. First, on the assumption that the VOI is effective and robust, if subjects dominantly favor a certain normative judgment behind the VOI this provides evidence in favor of that judgment; if, on the other hand, they do not dominantly favor a judgment this reduces our justification for it. Second, empirical investigations can also contribute to assessing the effectiveness and robustness of the VOI in the first place, thereby supporting or undermining its applications across the board. Analogous conclusions may apply to other impartiality thought experiments as well.

As mentioned, so far only little empirical research has been done on impartiality thought experiments. It also bears re-emphasizing that some of this research has been subject to serious methodological worries; and some of it has not been (primarily) motivated by philosophical aims at all. By addressing the potential normative relevance of empirical investigations on impartiality thought experiments we hope that this article motivates philosophically-motivated investigations of these thought experiments. We also provided concrete methodological suggestions as to what such investigations might look like. If done properly, future empirical research on impartiality thought experiments could have an important impact on a large number of debates in political philosophy and ethics.

[56] See, e.g. Pölzler, Zijlstra, Dijkstra (under review); Schwitzgebel, Ellis (2017).

Acknowledgements

Both authors have contributed equally to this paper. This research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 805498).

References

Aguiar F., Becker A., Miller L. (2013), "Whose Impartiality? An Experimental Study of Veiled Stakeholders, Involved Spectators and Detached Observers," *Economics & Philosophy* 29 (2): 155–174.

Bocian K., Wojciszke B. (2014), "Self-Interest Bias in Moral Judgments of Others' Actions," *Personality and Social Psychology Bulletin* 40 (7): 898–909.

Bond D., Park. J.-C. (1991), "An Empirical Test of Rawls's Theory of Justice: A Second Approach, in Korea and the United States," *Simulation & Gaming* 22 (4): 443–462.

Brun G. (2017), "Thought Experiments in Ethics," [in:] *The Routledge Companion to Thought Experiments*, M.T. Stuart, Y. Fehige, J.R. Brown (eds.), Routledge, London: 195–210.

Bruner J.P. (2018), "Decisions Behind the Veil. An Experimental Approach," [in:] *Oxford Studies in Experimental Philosophy* (vol. 2), T. Lombrozo, J. Knobe, S. Nichols (eds.), Oxford University Press, Oxford, New York.

Bruner J.P., Lindauer M. (2018), "The Varieties of Impartiality, or, Would an Egalitarian Endorse the Veil?," *Philosophical Studies* 177 (2): 459–477.

Carens J.H. (1987), "Aliens and Citizens: The Case for Open Borders," *The Review of Politics* 49 (2): 251–273.

Cath Y. (2016), "Reflective Equilibrium," [in:] *The Oxford Handbook of Philosophical Methodology*, H. Cappelen, T.S. Gendler, J. Hawthorne (eds.), Oxford University Press, Oxford, New York: 213–230.

Celikates R. (2012), "Der Schleier des Nichtwissens," [in:] *Philosophische Gedankenexperimente*, G.W. Bertram (ed.), Reclam, Ditzingen: 229–235.

Chan H.M. (2005), "Rawls' Theory of Justice: A Naturalistic Evaluation," *Journal of Medicine and Philosophy* 30 (5): 449–465.

Cohnitz D. (2005), *Gedankenexperimente in der Philosophie*, Mentis, Paderborn.

Demaree-Cotton J. (2016), "Do Framing Effects Make Moral Intuitions Unreliable?," *Philosophical Psychology* 29 (1): 1–22.

Demaree-Cotton J. (2019), "Analyzing Debunking Arguments in Moral Psychology: Beyond the Counterfactual Analysis of Influence by Irrelevant Factors," *Behavioral and Brain Sciences* 42: e151.

DeScioli P., Massenkoff M., Shaw A. et al. (2014), "Equity or Equality? Moral Judgments Follow the Money," *Proceedings of the Royal Society B: Biological Sciences* 281 (1797): 20142112.

Foot P. (2003), *Moral Dilemmas*, Clarendon Press, Oxford, New York.

Freeman S. (2006), *Rawls*, Routledge, London, New York.

Fritz Z., Cox C. (2019), "Conflicting Demands on a Modern Healthcare Service: Can Rawlsian Justice Provide a Guiding Philosophy for the NHS and Other Socialized Health Services?," *Bioethics* 33 (5): 609–616.

Frohlich N., Oppenheimer J.A. (1993), *Choosing Justice: An Experimental Approach to Ethical Theory*, University of California Press, Berkeley.

Frohlich N., Oppenheimer J.A., Eavey C.L. (1987), "Laboratory Results on Rawls's Distributive Justice," *British Journal of Political Science* 17 (1): 1–21.

Gähde U. (2000), "Zur Funktion ethischer Gedankenexperimente," [in:] *Wirtschaftsethische Perspektiven V: Methodische Ansätze, Probleme der Steuer- und Verteilungsgerechtigkeit, Ordnungsfragen*, W. Gaertner (ed.), Duncker & Humblot, Berlin: 183–206.

Gaus G., Thrasher J. (2016), "Rational Choice and the Original Position: The (Many) Models of Rawls and Harsanyi," [in:] *The Original Position*, T. Hinton (ed.), Cambridge University Press, Cambridge: 39–58.

Gendler T.S. (2007), "Philosophical Thought Experiments, Intuitions, and Cognitive Equilibrium," *Midwest Studies in Philosophy* 31 (1): 68–89.

Greene, J.D. (2008), "The Secret Joke of Kant's Soul," [in:] *Moral Psychology*, W. Sinnott-Armstrong (ed.), vol. 3, MIT Press, Cambridge, MA: 35–80.

Greenspan P. (2015), "Confabulating the Truth: In Defense of 'Defensive' Moral Reasoning," *Journal of Ethics* 19 (2): 105–123.

Hare C. (2016), "Should We Wish Well to All?," *The Philosophical Review* 125 (4): 451–472.

Harris J. (1987), "QALYfying the Value of Life," *Journal of Medical Ethics* 13 (3): 117–123.

Harris J. (1995), "Double Jeopardy and the Veil of Ignorance – a Reply," *Journal of Medical Ethics* 21 (3): 151–157.

Harris J. (1996), "Would Aristotle Have Played Russian Roulette?," *Journal of Medical Ethics* 22 (4): 209–215.

Harsanyi J.C. (1955), "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," *Journal of Political Economy* 63 (4): 309–321.

Harsanyi J.C. (1975), "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory," *The American Political Science Review* 69 (2): 594–606.

Herne K., Mard T. (2008), "Three Versions of Impartiality: An Experimental Investigation," *Homo Oeconomicus* 25: 27–53.

Herne K., Suojanen M. (2004), "The Role of Information in Choices Over Income Distributions," *Journal of Conflict Resolution* 48 (2): 173–193.

Hinton T. (2016), "Introduction: The Original Position and The Original Position – an Overview," [in:] *The Original Position*, T. Hinton (ed.), Cambridge University Press, Cambridge, UK; New York, USA: 1–17.

Huang K., Greene J.D., Bazerman M. (2019), "Veil-of-Ignorance Reasoning Favors the Greater Good," *Proceedings of the National Academy of Sciences*, 116 (48): 201910125.

Hübner D. (2017), "Three Remarks on 'Reflective Equilibrium'," *Philosophical Inquiry* 41 (1): 11–40.

Jollimore T. (2018). "Impartiality," [in:] *The Stanford Encyclopedia of Philosophy (Summer 2020 Edition)*, E.N. Zalta (ed.), URL = https://plato.stanford.edu/archives/win2018/entries/impartiality/ [Accessed 04.03.2020].

Kauppinen A. (2007), "The Rise and Fall of Experimental Philosophy," *Philosophical Explorations* 10 (2): 95–118.

Kauppinen A. (2018), "Who's Afraid of Trolleys?," [in:] *Methodology and Moral Philosophy*, J. Suikkanen, A. Kauppinen (eds.), Routledge, New York.

Kumar V., Campbell R. (2012), "On the Normative Significance of Experimental Moral Psychology," *Philosophical Psychology* 25 (3): 311–330.

Landy J.F., Goodwin G.P. (2015), "Does Incidental Disgust Amplify Moral Judgment? A Meta-Analytic Review of Experimental Evidence," *Perspectives on Psychological Science* 10 (4): 518–536.

Lissowski G., Tyszka T., Okrasa W. (1991), "Principles of Distributive Justice: Experiments in Poland and America," *Journal of Conflict Resolution* 35 (1): 98–119.
Löhr G. (2019), "The Experience Machine and the Expertise Defense," *Philosophical Psychology* 32 (2): 257–273.
Machery E. (2017), *Philosophy within Its Proper Bounds*, Oxford University Press, Oxford, New York.
May J. (2018), *Regard for Reason in the Moral Mind*, Oxford University Press, New York.
McKie J., Kuhse H., Richardson J. et al. (1996a), "Another Peep Behind the Veil," *Journal of Medical Ethics* 22 (4): 216–221.
McKie J., Kuhse H., Richardson J. et al. (1996b), "Double Jeopardy, the Equal Value of Lives and the Veil of Ignorance: A Rejoinder to Harris," *Journal of Medical Ethics* 22 (4): 204–208.
Michelbach P.A., Scott J.T., Matland R.E. et al. (2003), "Doing Rawls Justice: An Experimental Study of Income Distribution Norms," *American Journal of Political Science* 47 (3): 523–539.
Miscevic N. (2017). "Thought Experiments in Political Philosophy," [in:] *The Routledge Companion to Thought Experiments*, M.T. Stuart, Y. Fehige, J.R. Brown (eds), Routledge, London: 153–170.
Nadelhoffer T., Feltz A. (2008), "The Actor–Observer Bias and Moral Intuitions: Adding Fuel to Sinnott-Armstrong's Fire," *Neuroethics* 1 (2): 133–144.
Paulo N. (2018), "In Search of Greene's Argument," *Utilitas* 31 (1): 38–58.
Paulo N. (2020), "Moral Intuitions between Higher-Order Evidence and Wishful Thinking," [in:] *Higher-Order Evidence and Moral Epistemology*, M. Klenk (ed.), Routledge, London.
Paulo N., Pölzler T. (under review), "Thought Experiments and Experimental Ethics".
Petrinovich L., O'Neill P. (1996), "Influence of Wording and Framing Effects on Moral Intuitions," *Ethology and Sociobiology* 17 (3): 145–171.
Pölzler T. (2018), *Moral Reality and the Empirical Sciences*, Routledge, New York.
Pölzler T., Wright J.C. (2019), "Anti-Realist Pluralism: A New Approach to Folk Metaethics," *Review of Philosophy and Psychology* 11 (1): 53–82.
Pölzler T., Zijlstra L., Dijkstra J. (under review), "Moral Progress, Knowledge, and Error: What Are the Folk's Implicit Commitments about Moral Objectivity?".
Raphael D.D. (2007), *The Impartial Spectator: Adam Smith's Moral Philosophy*, Oxford University Press, Oxford.
Rawls J. (2001). *Justice As Fairness: A Restatement*, E. Kelly (ed.), The Belknap Press, Cambridge, MA.
Rawls J. (2005), *A Theory of Justice* (original edition), The Belknap Press, Cambridge, MA.
Rini R.A. (2013), "Making Psychology Normatively Significant," *Journal of Ethics* 17 (3): 257–274.
Schnall S., Benton J., Harvey S. (2008), "With a Clean Conscience: Cleanliness Reduces the Severity of Moral Judgments," *Psychological Science* 19 (12): 1219–1222.
Schnall S., Haidt J., Clore G.L. et al. (2008), "Disgust as Embodied Moral Judgment," *Personality & Social Psychology Bulletin* 34 (8): 1096–1109.
Schwitzgebel E., Cushman F. (2012), "Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers," *Mind and Language* 27 (2): 135–153.
Schwitzgebel E., Ellis J. (2017), "Rationalization in Moral and Philosophical Thought," [in:] *Moral Inferences*, J.-F. Bonnefon, B. Trémolière (eds.), Psychology Press, London, New York.

Singer P., McKie J., Kuhse H. et al. (1995). "Double Jeopardy and the Use of QALYs in Health Care Allocation," *Journal of Medical Ethics* 21 (3): 144–150.

Singer P. (1972), "Famine, Affluence, and Morality," *Philosophy & Public Affairs* 1 (3): 229–243.

Stuart M.T. (2017), "How Thought Experiments Increase Understanding," [in:] *The Routledge Companion to Thought Experiments*, M.T. Stuart, Y. Fehige, J.R. Brown (eds), Routledge, London: 526–544.

Sytsma J., Livengood J. (2015), *The Theory and Practice of Experimental Philosophy*, Broadview Press, Peterborough.

Thomson J.J. (1976), "Killing, Letting Die, and the Trolley Problem," *The Monist* 59 (2): 204–217.

Tobia K., Buckwalter W., Stich S. (2013), "Moral Intuitions: Are Philosophers Experts?," *Philosophical Psychology* 26 (5): 629–638.

Valdesolo P., DeSteno D. (2006), "Manipulations of Emotional Context Shape Moral Judgment," *Psychological Science* 17 (6): 476–477.

Wiegmann A., Okan Y., Nagel J. (2012), "Order Effects in Moral Judgment," *Philosophical Psychology* 25 (6): 813–836.

Williamson T. (2008), *The Philosophy of Philosophy*, Wiley-Blackwell, Malden, MA.

Wolf S., Dron C. (2015), "Intergenerational Sharing of Non-Renewable Resources: An Experimental Study Using Rawls's Veil of Ignorance," Constitutional Economics Network Working Paper Series, URL = https://www.econstor.eu/bitstream/10419/109031/1/821463128.pdf [Accessed 11.5.2020].

Wolf S., Lenger A. (2014), "Utilitarianism, the Difference Principle, or Else? An Experimental Analysis of the Impact of Social Immobility on the Democratic Election of Distributive Rules," [in:] *Experimental Ethics: Toward an Empirical Moral Philosophy*, C. Luetge, H. Rusch, M. Uhl (eds.), Palgrave Macmillan, London: 94–111.

Żuradzki T. (2014), "Preimplantation Genetic Diagnosis and Rational Choice under Risk or Uncertainty," *Journal of Medical Ethics* 40 (11): 774–778.