Journal of Didactics of Philosophy, Vol. 6, 2022, 1-17 DOI: 10.46586/JDPh.2022.9680

SEEING THE WORLD MORE CLEARLY. STRATEGIES FOR UNLEASHING THE FULL MORAL POTENTIAL OF THOUGHT EXPERIMENTS IN THE PHILOSOPHY CLASSROOM

Dominik Balg
Tübingen University
dominik.balg@uni-tuebingen.de

Received: 28 February 2022
Accepted: 18 April 2022

Abstract

In this paper, I discuss the effects of using thought experiments for the purpose of conceptual clarification of students' hermeneutical abilities. On the one hand, by providing opportunities to explore the scope of normatively loaded concepts, thought experiments can effectively help students to interpret their social and moral reality more adequately, which in some cases might even help to reduce existing hermeneutical injustices. On the other hand, given their notorious susceptibility to distorting factors that are philosophically irrelevant, they can also push students into accepting idiosyncratic intuitions that they don't really share and thereby further impair their hermeneutical abilities. After setting out this dilemma in more detail, I will propose various strategies for facilitating the safe use of thought experiments that instructors can use to effectively exploit the empowering potential of thought experiments.

Keywords: thought experiments, intuitions, method of cases, hermeneutical abilities

Introduction

In the introduction to Plato's *Republic*, Socrates, Cephalus, Polemarchus, and Thrasymachus discuss the best definition of justice. In response to Cephalus's definition of justice as speaking the truth and paying one's debts, Socrates comes up with the following imaginary situation:

> Suppose that a friend when in his right mind has deposited arms with me and he asks for them when he is not in his right mind, ought I to give them back to him? No one would say that I ought or that I should be right in doing so, any more than they would say that I ought always to speak the truth to one who is in his condition. (Plato 1892: 331)

Although philosophy has changed in many ways since Plato wrote this passage, the method that Socrates employs here in order to criticize Cephalus's definition of justice is still well known: Philosophers present specific situations and ask their audience whether the people or objects described in these situations instantiate some philosophically important property or relation.

Although the exact philosophical nature and purpose of this so-called 'method of cases' is subject to controversy (see e.g. Machery 2017, ch. 1), its didactical potential has been widely acknowledged (see e.g. Förg 2019, Matthews 1979). In fact, given how short, accessible, and illustrative many philosophical thought experiments are, their popularity in philosophy classrooms shouldn't come as a surprise – cases like Descartes's *Evil Demon Problem*, Singer's *Drowning Child Argument*, Foot's *Trolley Case*, and Rawls's *Veil of Ignorance* are among the all-time classics in philosophy textbooks.

At the same time, several worries about the didactical implications of using thought experiments in philosophy classes have been discussed. For example, it has been argued that using thought experiments in philosophy classes conveys a problematic view of philosophy as a quasi-scientific enterprise (Martena 2018: 401), and that thought experiments are too artificial to offer students any useful guidance in real-world situations (ibid.: 393), and that they reinforce problematic social stereotypes and assumptions (Lanphier/McKiernan 2020). In this paper, I would like to open up a new perspective on the didactical risks and merits of thought experiments within the context of philosophy teaching that has so far been neglected. More specifically, I would like to discuss the effects of using thought experiments in educational settings on students' hermeneutical abilities.[1]

This discussion relies on a specific, but a widely shared and comparatively uncontroversial conception of the philosophical method outlined above. According to this conception, thought experiments can effectively be employed as philosophical arguments, because they generate specific philosophical intuitions. Joachim Horvath and Steffen Koch characterize this conception as follows:

> The method of cases is widely regarded as a key philosophical method, with clear instances already seen in Plato's early dialogs. Even if there is no wholly uncontroversial characterization of this method, the basic idea can be put as follows: intuitive verdicts about particular cases […] often play a decisive role in supporting or undermining philosophical theories, depending on how well those theories accommodate the intuitive verdicts in question. (Horvath/Koch 2021: 2)

This basic methodological conception already allows for a better understanding of *what thought experiments are*: In what follows, I will presuppose a rather loose understanding of the term "thought experiment" that characterizes thought experiments via their function within philosophical research. According to this understanding, thought experiments are simply the kind of thing that is used in the method of cases – for example, in his *Stanford Encyclopedia of Philosophy* entry on thought experiments, James Robert Brown and Fehige Yiftach write:

> [Here is one] of the most common features of what it means to engage in the conduct of thought experiments: we visualize some situation that we have set up in the imagination; we

[1] I would like to thank the participants of the Göttingen Colloquium for Didactics of Philosophy and two anonymous reviewers for very helpful comments on earlier versions of this paper.

> let it run or we carry out an operation; we see what happens; finally, we draw a conclusion. (Brown/Fehige 2022)

One advantage of this functional characterization of thought experiments is that it allows us to stay neutral concerning several philosophical questions that are – although interesting and important – ultimately irrelevant to our didactical discussion, like the question of whether thought experiments have to be counterfactual, whether they have to be physically unrealizable or how they are best individuated. Against the background of these conceptual and methodological clarifications, the specific research question that I would like to discuss in this paper is the following: *How can thought experiments be used in educational settings in a way that effectively improves students' hermeneutical abilities?*

This paper has three sections. In section 1, I will take a closer look at how philosophical thought experiments are used in educational settings and how they have the potential to effectively improve students' ability to adequately interpret their social and moral reality. I will argue that, by helping to explore the scope of normatively loaded concepts, thought experiments can enable students to make sense of specific social experiences they were previously unable to conceptualize adequately. In cases where this prior inability results from hermeneutical marginalization, this effect even leads to a desirable decrease in hermeneutical injustice. In section 2, I will turn to the risks of using thought experiments for the purpose of conceptual clarification. Here I will argue that because of their susceptibility to distorting factors, thought experiments don't just have the potential to clarify, but also to further obscure our conceptual schemes. Depending on the plausibility of some controversial empirical assumptions about the structural profile of some of these distorting factors, this might even lead to a problematic increase in hermeneutical injustice. These considerations accentuate the urgent need for effective didactical strategies that facilitate the safe use of thought experiments in philosophy classes. In section 3, I will introduce three specific measures that philosophy teachers can take to effectively avoid the risks and dangers that have been delineated in section 2. Once these measures are included, using philosophical thought experiments for the purpose of conceptual clarification will be a promising way to improve students' ability to adequately interpret their social and moral reality.

1. The good news

In educational settings, philosophical thought experiments do not necessarily serve the same purposes that they do in their original academic context. Most notably, thought experiments in philosophy classes often serve a variety of *pedagogical* purposes that are completely irrelevant within the context of academic philosophy – for example, when they are used to help students develop certain emotional and social skills like empathy or self-awareness (Engels 2017: 193). Nevertheless, there are of course also some genuinely *philosophical* purposes that thought experiments serve in the philosophy classroom that directly mirror their use in academic

philosophy.[2] Many of these purposes have to do with the *clarification of concepts*: just like professional philosophers, students use thought experiments to support, reject or modify definitions of philosophical concepts. In this context, thought experiments often serve as classical counterexamples. They present either situations where a certain concept intuitively applies that are excluded by a given definition of this concept, or situations that are respected by a given definition even though the definiendum intuitively doesn't apply (Wieland/Endt 2017).

By providing opportunities to test philosophical definitions, thought experiments help students to explore the scope of normatively loaded concepts like *justice*, *moral wrongness*, *consent*, *identity,* and *knowledge*. Given this, it becomes clear why thought experiments have a lot of empowering potential. They allow students to critically reflect on the conceptual tools they have to describe their moral and social reality, and thereby help them to adequately interpret their respective experiences. For example, take a case where a group of students is confronted with Frankfurt-style counterexamples to definitions of moral responsibility that respect the initially plausible *principle of alternate possibilities (PAP)*, according to which a person is morally responsible for her actions only if she could have done otherwise. For those students who share the relevant intuition, these counterexamples might have an empowering effect on them by helping them to make sense of their self-experience as morally responsible agents living in a world that seems to be completely deterministic.

While in this specific example the empowering effect of the thought experiment is at least partly explained by the fact that it directly contradicts a widely shared philosophical preconception and thus enables students who already felt a vague sense of unease at this preconception to adequately *conceptualize* and *express* their concerns, thought experiments can also have empowering effects without being genuinely subversive in such a way. For example, take a case where a group of students is confronted with thought experiments that are designed to establish the moral wrongness of factory farming, such as Alastair Norcross' chocolate-lover case (Norcross 2004). In such a case, it wouldn't seem unrealistic to assume that most students already explicitly believe that factory farming is in fact morally problematic *before* being confronted with the relevant thought experiments. However, performing and discussing these thought experiments could still have a significant empowering effect by enabling the students to understand *why* factory farming is morally problematic and thus *justify* and *defend* their pre-existing beliefs. Given this, it seems that the empowering potential of thought experiments is not so much grounded in their specific content, but rather in their broader structural purpose: By providing opportunities to test the applicability of certain conceptual tools, they help students to critically engage with their moral and social reality.

This empowering potential of thought experiments is all the more important in cases where the students' inability to interpret or explain their moral and social experiences constitutes an instance of *hermeneutical injustice*. Hermeneutical injustice is a special form of epistemic injustice that puts specific groups of people at an unfair disadvantage with respect to their ability

[2] For a helpful classification of different philosophical functions that thought experiments can fulfill in didactical contexts see e.g. Klaiber 2018. Although his taxonomy is ultimately meant as a classification of different forms of *example cases*, it can also easily be applied to different forms of *thought experiments*.

to make sense of their social experiences. More specifically, hermeneutical injustice is defined as "the injustice of having some significant area of one's social experience obscured from collective understanding owing to persistent and wide-ranging hermeneutical marginalization" (Fricker 2007: 154). For example, in the above case where students are confronted with Frankfurt-style counterexamples to the principle of alternate possibilities, their prior inability to adequately interpret their self-experiences as morally responsible agents does not amount to such forms of injustice. While their experiences might indeed be obscured from collective understanding – as already said, it seems plausible that the collective hermeneutical resource operates in accordance with the principle of alternate possibilities, and thus excludes persons who could not have done otherwise from the circle of morally responsible agents (Robb 2020, section 2.2) – the reason for this is clearly not that persons who could not have done otherwise are hermeneutically marginalized based on some problematic distribution of social power. If anything, the above example is a case of what Fricker calls *epistemic bad luck*.[3]

However, there are plausibly also a lot of realistic scenarios where using thought experiments for the purpose of conceptual clarification will help students to make sense of social experiences that are obscured from collective understanding due to genuine hermeneutical marginalization. For example, take a case where students who discuss philosophical questions of consent are confronted with counterexamples to narrow conceptions of the *Performative View* of Consent. According to these conceptions, consent occurs when an agent behaves in a way that conventionally counts as an act token of consent, no matter whether she believes she is consenting or not (see e.g. Wertheimer 2003: 144; Healey 2015: 354). Discussing some of the widely accepted counterexamples to these conceptions (Schnüriger 2018) might help some (especially female) students who have had negative sexual experiences to adequately interpret these experiences by conceptualizing them as non-consensual. What's more, such students' prior inability to conceptualize their experiences as non-consensual is plausibly not the result of sheer epistemic bad luck. In fact, the very definition of hermeneutical injustice was originally developed against the background of various examples of women's inability to adequately conceptualize experiences of sexual harassment due to persistent and wide-ranging hermeneutical marginalization (Fricker 2007, ch. 7). In light of this, it seems safe to say that providing female students with the conceptual resources to adequately interpret negative sexual experiences as non-consensual is not just empowering, but a genuine act of epistemic justice.

[3] Fricker uses several examples to illustrate the difference between cases of *epistemic bad luck* and genuine instances of *hermeneutical injustice* (for the following, see Fricker 2007, ch. 7). As an example of the former, she describes a situation where a person suffers from a unique medical condition that impacts her social behavior and that is not yet medically explored. As a result, the condition remains undiagnosed, so that the person is unable to adequately conceptualize her respective experiences. Fricker contrasts this situation with a case where a woman experiences sexual harassment at a time where the concept of sexual harassment is still not publicly available. In this case, the woman's inability to adequately conceptualize her experiences is not just a form of epistemic bad luck, but a genuine instance of hermeneutical injustice. According to Fricker, the crucial difference between these two cases is that in the latter case, women's inability to adequately conceptualize experiences of sexual harassment is grounded in unequal hermeneutical participation: Due to sexist distributions of social power, women were prevented from participating on equal terms with men in those practices by which collective social meanings are generated, such as academia, law and journalism – and as a direct result of this marginalization, the concept of sexual harassment wasn't part of the collective hermeneutical resources which people can use to describe and interpret their social experiences.

Another example would be a case where students are confronted with counterexamples to traditional definitions of *woman* and *man*, according to which these words are to be defined in terms of biological sex – i.e. *woman* as „adult female human" and *man* as „adult male human" (Bogardus 2020). Discussing some prominent counterexamples to these definitions (see e.g. Corvino 2000: 174; Bettcher 2009: 103) might help students who don't (fully) identify with their biological sex to adequately interpret their experiences by conceptualizing them against the background of a distinction between *sex* and *gender*. And again, such students' prior inability to adequately interpret their experiences is plausibly the result of unfair hermeneutical marginalization: Given the amount of widespread and deeply rooted transphobia we are still facing today, it seems plausible to assume that people who don't identify with their biological sex have been systematically excluded from participating on equal terms in practices by which collective social meanings are generated. Given this, providing such people with the conceptual tools that they need to adequately interpret their social experiences will again be a genuine act of epistemic justice.

2. The bad news

Given the above considerations, it seems that using thought experiments for the purpose of conceptual clarification in philosophy classes is not just philosophically enlightening, but also pedagogically and morally beneficial. By providing opportunities to explore the scope of normatively loaded concepts, thought experiments enable students to interpret their social and moral reality more adequately, which in some cases might even help to reduce existing hermeneutical injustices. But at the same time, using thought experiments in philosophy classes also comes with a certain risk: In order to adequately interpret their social and moral reality, students need conceptual tools that are suited to adequately represent their personal experiences. And while thought experiments can be of great help in developing such conceptual tools, they also come with some notorious flaws and limitations. To see why it will be helpful to start by examining the role that thought experiments play within the context of conceptual clarification in a little more detail.

As already mentioned above, thought experiments are often used as counterexamples to candidate definitions of philosophical concepts. Counterexamples work by generating specific intuitions about the applicability of a given concept in a given situation, which then serve as a simple test for any definition of this concept. That is, a successful definition is a definition that picks out all and only those cases in which the defined concept intuitively applies. However, one well-known problem in that context is that our intuitive verdicts about hypothetical scenarios seem to be sensitive to various kinds of contextual factors that are philosophically irrelevant, such as order of presentation (see e.g. Liao et al. 2012), affective content (see e.g. Nichols/Knobe 2007) or incidental emotions (see e.g. Cameron et al. 2013). Furthermore, this is not just a problem for lay people – there is robust empirical evidence that professional philosophers are no more resistant to these irrelevant factors than ordinary people (Schwitzgebel/Cushmann 2012, 2015; Löhr 2019; Wiegmann/Horvath/Meyer 2020). In light of such findings, it looks like thought experiments don't just have the potential to clarify, but also to further obscure our conceptual schemes. More specifically, the worry is that students might

easily be pushed into accepting idiosyncratic intuitions that they don't really share and thereby adopt a conceptual scheme that doesn't map onto their own conceptual intuitions and that is therefore inapt to adequately represent their personal experiences. To successfully avoid such further obscuration, students need to acquire skills and competences that are necessary to apply the method of cases in a critical and reflective way in order to freely discover which intuitions they really have.

While the acquirement of such skills and competences is already desirable on purely pedagogical grounds and should thus be an important didactical desideratum in its own right, it could also have important political implications. This becomes clear once we realize that our conceptual intuitions might not only be affected by contextual factors, but also by systematic factors. And in fact, there is a growing body of empirical literature suggesting that intuitive verdicts about hypothetical scenarios are systematically affected by personality traits like extraversion and introversion (Feltz/Cokely 2019; Schulz/Cokely/Feltz 2011). Furthermore, there are studies suggesting significant differences in intuitions between participants with higher and lower socioeconomic statuses, and between participants from different cultural backgrounds (Weinberg et al. 2001; Nichols et al. 2003), as well as gender differences (Starmans/Friedman 2009) in relation to famous epistemological thought experiments. Similar studies suggest cultural differences (Abarbanell/Hauser 2010; Curtin et al. 2020) and gender differences (for an overview see Buckwalter/Stich 2013) in intuitions on ethical thought experiments, as well as cultural differences in intuitions about thought experiments concerning the compatibility of determinism and moral responsibility (Hannikainen et al. 2019) or free will (Berniūnas et al. 2021), and about certain thought experiments that have been developed within the philosophy of language (Machery et al. 2004; Machery/Olivola/de Blanc 2009; Machery/Sytsma/Deutsch 2015).

To be clear, these studies are subject to the ongoing controversy. Especially some of the findings on cultural differences and many of the findings on gender differences have not replicated (see e.g. Kim/Yuan 2015; Seyedsayamdost 2015; Adleberg/Thompson/Nahmias 2015). Furthermore, there are also studies that indicate high levels of cross-cultural uniformity in intuitive judgements (Machery et al. 2015). Given this, it would be clearly premature to simply presuppose that a person's gender or cultural and socioeconomic background significantly influences her intuitive judgements. However, against the background of our discussion of hermeneutical injustice in section 1, it becomes clear that *if* it turned out that such distorting influences do in fact exist, then this would directly point to an additional problem that has hitherto been neglected. To see this, one just has to envision the fact that the overwhelming majority of thought experiments that are standardly used in philosophy classes have been developed by philosophers who are WEIRD – i.e. by people from Western, Educated, Industrialized, Rich, and Democratic societies. Moreover, most canonical philosophers are male.

While the prevalence of these demographic features within the philosophical tradition and the philosophical community will be problematic on many different levels, it might also have profound implications for the widespread use of thought experiments in educational contexts. More specifically, the worry is that if these demographic features turned out to have distorting

effects on our intuitive verdicts about hypothetical scenarios, then using thought experiments for the purpose of conceptual clarification in philosophy classes – while theoretically having the potential to mitigate hermeneutical injustices – might actually cause and reinforce certain forms of hermeneutical injustice: Students who are constantly confronted with thought experiments that are specifically designed to pump conceptual intuitions that they don't share will likely start to distrust their own intuitions and adopt a conceptual scheme that isn't suited to adequately represent their personal experiences, which will significantly impair their ability to make sense of their social reality. What's more, in the case of female and non-WEIRD students, this impairment would not simply be the result of epistemic bad luck. To see why one only needs to consider the reasons behind the remarkable prevalence of WEIRD male authors within academic philosophy and within school curricula. In fact, this prevalence is plausibly caused by unfair power relations that have prevented women and non-WEIRD people from participating on equal terms with WEIRD men in academic activities (Fricker 2007: 152). So, it seems that if this prevalence really led to a structural impairment of female and non-WEIRD students' ability to adequately interpret their social reality, then this would be a direct result of persistent and wide-ranging hermeneutical marginalization of female and non-WEIRD philosophers.

In light of the above considerations, it seems that the widespread use of thought experiments in philosophy classes is not just easily affected by contextual factors that are philosophically irrelevant, but might even be an additional source of hermeneutical injustice. Given the significance of its political and moral implications, teachers need to be aware of this possibility. However, given the provisional and controversial nature of the underlying empirical literature, I will not rely on this assumption in what follows. For as we have seen, even if the distorting influences on our intuitive verdicts aren't structural, they are still philosophically irrelevant. Given this, we have good pedagogical reasons to enable students to apply the method of cases in a critical and reflective way and to provide them with opportunities to freely discover their personal conceptual intuitions. In order to do this, we need specific didactical strategies that help us to reap the pedagogical benefits of thought experiments while at the same time avoiding the risks and dangers that have been delineated in this section.

3. How to safely use thought experiments in the philosophy classroom

By providing opportunities to explore the scope of normatively loaded concepts, thought experiments can effectively help students to interpret their social and moral reality more adequately. Herein lies their empowering potential. At the same time, given their notorious susceptibility to distorting factors that are philosophically irrelevant, thought experiments also have the potential to impair students' hermeneutical abilities by pushing them into accepting idiosyncratic intuitions that they don't really share. Herein lies their destructive potential.

In light of these results from section 1 and 2, philosophy teachers need to be provided with concrete didactical strategies to exploit the empowering potential of thought experiments while at the same time avoiding the risks and dangers associated with their usage. At this point, one obvious suggestion is that teachers should simply make their students *aware* of all the different ways in which their intuitive responses to thought experiments can be influenced and distorted

by philosophically irrelevant influences. And while this might indeed be an important first step, it seems that we still have good reasons to come up with further supplementary strategies. Most importantly, psychologically distorting influences are well-known for their recalcitrance: Making people aware of implicit and unconscious factors that impact their judgements doesn't always reduce their distorting effects – in fact, it sometimes even exacerbates them (Balg 2021: 16). Given this, it seems advisable to look for alternatives. In what follows, I would like to introduce three different measures to facilitate the safe use of thought experiments in the philosophy classroom. This list is in no way meant to be exhaustive; in fact, given the tentativeness of the empirical literature discussed in the last section and given how little attention this literature has received within didactical contexts, it seems that we would first need a clearer picture of the structure and scope of the underlying problem in order to come up with more sophisticated and better-directed strategies. However, in the absence of such empirical details, the following measures are an important step towards a just and fruitful use of thought experiments in the philosophy classroom.

3.1 Bracketing the author's performance

What are the different ways in which students might be pushed into accepting idiosyncratic intuitions that they don't really share? In order to answer this question, it will be helpful to start by examining some structural details of how thought experiments can be used in philosophy classes. Some authors have argued for a distinction between a *narrow* and a *broad* meaning of the term 'thought experiment' in didactical contexts (see e.g. Engels 2017: 189). According to its narrow meaning, the term 'thought experiment' only refers to the *performance* of a thought experiment, i.e. the realization and verbalization of intuitions on the basis of a given description of a specific situation. According to its broad meaning, the term 'thought experiment' also refers to the *set-up* of a thought experiment, i.e. the description of the situation and the instructions which specify the conceptual target of the thought experiment.

To illustrate this distinction, consider Judith Jarvis Thomson's famous *Violinist Case*. In her paper 'A Defense of Abortion', Thomson starts by describing a specific situation:

> You wake up in the morning and find yourself back to back in bed with an unconscious violinist. A famous unconscious violinist. He has been found to have a fatal kidney ailment, and the Society of Music Lovers has canvassed all the available medical records and found that you alone have the right blood type to help. They have therefore kidnapped you, and last night the violinist's circulatory system was plugged into yours, so that your kidneys can be used to extract poisons from his blood as well as your own. The director of the hospital now tells you, 'Look, we're sorry the Society of Music Lovers did this to you – we would never have permitted it if we had known. But still, they did it, and the violinist now is plugged into you. To unplug you would be to kill him. But never mind, it's only for nine months. By then he will have recovered from his ailment, and can safely be unplugged from you. (Thomson 1971: 48)

Having described this situation, Thomson goes on to specify the conceptual target of her thought experiment:

> Is it morally incumbent on you to accede to this situation? No doubt it would be very nice of you if you did, a great kindness. But do you *have* to accede to it? (ibid.: 49)

These two passages constitute the *set-up* of Thomson's thought experiment – in the broad sense of 'thought experiment', this set-up already *is* the whole thought experiment. Having set the stage in this way, Thomson continues by presenting her intuitions about the described situation:

> What if it were not nine months, but nine years? Or longer still? What if the director of the hospital says, 'Tough luck, I agree, but you've now got to stay in bed, with the violinist plugged into you, for the rest of your life. Because remember this. All persons have a right to life, and violinists are persons. Granted you have a right to decide what happens in and to your body, but a person's right to life outweighs your right to decide what happens in and to your body. So you cannot ever be unplugged from him.' I imagine you would regard this as outrageous, which suggests that something really is wrong with that plausible-sounding argument I mentioned a moment ago. (ibid.: 49)

In this passage, Thomson *performs* her experiment. In the narrow sense of the term, it is only this performance that properly deserves to be called a 'thought experiment'. However, the important point for our purpose is that Thomson performs her thought experiment in a specific way: While it should be clear that Thomson is only stating her own intuitions, she simultaneously insinuates that all of her readers will share these intuitions. Instead of writing "I regard this as outrageous", she explicitly addresses her readers and makes clear that she expects them to assess the described situation in the same way that she does.

This way of performing a thought experiment is fairly common among canonical authors. Philosophers often more or less explicitly assume that their audience will share their personal intuitions and that their performances of thought experiments will be generally accepted.[4] In light of the empirical findings discussed in the previous section, this is a rather problematic thing to do. Instead of presenting their personal conceptual intuitions as universally valid insights, philosophers should invite their readers to perform the relevant thought experiments by themselves. Otherwise, they risk imposing on them conceptual schemes that do not match their readers' personal intuitions, which would likely impair their readers' hermeneutical abilities. Accordingly, teachers of philosophy should always encourage their students to perform thought experiments on their own. Moreover, they should give them the opportunity

[4] Note that this is not necessarily a psychological assumption about any conscious motives and intentions of particular philosophers, but rather a structural assumption about the dialectical mechanics of philosophical discourse. Given how philosophical debates traditionally take place, authors are expected to present and defend their philosophical positions as universally valid insights that should be generally accepted. Against the background of such a dynamic, thought experiments can only provide successful argumentative support if they generate intuitions that are shared by the majority of their readers. Given this, it becomes clear why traditional philosophical discourse already operates under the implicit assumption that conceptual intuitions can – and should be – universally shared.

to do so *before* they are confronted with the author's own performance. In fact, in light of the above considerations, this additional requirement is crucial: While many of the pedagogical benefits of letting students perform thought experiments on their own are already widely appreciated, the distinction between set-up and performance is usually neglected in this context. In practice, students are often confronted with text passages that contain the *complete* thought experiment, i.e. its set-up *and* its performance by the author.

However, before being confronted with the author's own performance, students should have the chance to explore their personal conceptual intuitions as independently as possible. To make this possible, teachers of philosophy need to carefully separate a thought experiment's set-up from its performance. In some cases, it won't even be necessary to discuss the author's own performance of the thought experiment - starting with the original set-up, students can simply rely on their personal intuitions as a basis for further discussion. On the other hand, presenting and discussing the author's own performance will often help students to better understand the dialectical purpose of a thought experiment and thus put them in a better position to critically deal with it. Accordingly, the idea behind the proposed strategy is not to simply *exclude* author performances from didactical contexts, but rather to carefully *separate* them from the mere set-ups. Distinguishing between thought experiments in the broader sense and thought experiments in the narrow sense in this way gives students the freedom to rely on their own intuitions when developing the conceptual tools that they need to adequately conceptualize their social experiences.

3.2 Removing other manipulating factors

Performing their own thought experiments is not the only way in which philosophers can – intentionally or unintentionally - push their audience into accepting their conceptual intuitions. For example, many authors explicitly state the dialectical purpose of their thought experiments *before* setting up the actual experiment. And often enough, they do this with a considerable amount of confidence. For instance, just before setting up his famous counterexamples to the traditional JTB account of knowledge, Edmund Gettier writes:

> I shall now present two cases in which the conditions stated in [… the JTB account] are true for some proposition, though it is at the same time false that the person in question knows that proposition. (Gettier 1963: 121)

By making this remark, Gettier already primes his audience to accept his subsequent performance of the relevant thought experiments. What's more, even in cases where the author doesn't make such an introductory remark, the reader is still often in a position to anticipate the dialectical purpose of the thought experiments that are presented to him. The reason for this is that thought experiments are always developed by specific authors in a specific philosophical context. For example, take a reader who knows Peter Singer to be a notorious proponent of utilitarianism and a committed supporter of the effective altruism movement. If this reader encounters Singer's *Drowning Child Case* for the first time and she knows that this thought experiment has been developed by Peter Singer, she will – even without having access to

Singer's actual performance of the experiment – already have a pretty good understanding of what this thought experiment is supposed to show.

Given this, philosophy instructors should try to reduce as many manipulating factors as possible when teaching thought experiments. Ideally, when performing thought experiments, students should have as little information as possible about what the relevant thought experiments are supposed to show. Only if students don't know which intuitions they are supposed to have, are they in a position to freely discover which intuitions they really have. But what can philosophy instructors do to reduce manipulating factors? First of all, when teaching thought experiments, they should start by presenting the bare set-up of a thought experiment, leaving out not only the author's own performance of the experiment but also any other introductory or interpretative remarks that could push the students in a certain direction. Again, this does not mean that there is no place for the authors' personal interpretations of their own thought experiments in philosophy classes. However, these interpretations should only be discussed *after* the students have first had the chance to explore their own intuitions.

Furthermore, it also seems advisable not to reveal the authors' identities before the students' performance of their thought experiments. In some cases, it will actually be best to let the students perform certain thought experiments before they even know *anything* about the philosophical context of these experiments. In fact, one promising strategy would be to start a given unit by first presenting all the thought experiments that will be discussed in this unit at the outset, and let the students perform these thought experiments before they even know what the topic of the unit will be. Later on, when particular thought experiments are examined in more detail, the students can then recall their initial intuitions and use these as a basis for further discussion. Finally, it is also important to keep in mind that the students' performance of a given thought experiment will not only be influenced by the author's own performance, but also by the performances of their classmates. To mitigate this influence, it might be a good idea to let students perform thought experiments anonymously – for example, by using digital tools that enable anonymized classroom polling.

Obviously, identifying and properly handling all these different factors at the same time is a challenging and complex task. Accordingly, teachers shouldn't have to cope with it on their own. To support them in their didactical efforts, didactics experts and textbook publishers need to provide appropriate concepts and teaching materials that effectively take the above considerations into account. Given this, the proposed strategy shouldn't primarily – or at least exclusively – be regarded as a desideratum for pedagogical practice, but also for didactical research.

3.3 Designing variations

The basic idea behind the above considerations is to create an environment in which students can perform thought experiments without constantly being influenced by external factors that might push them to accept conceptual intuitions that they don't really share. However, it might turn out that the problem runs even deeper. To manipulate their readers' intuitions, authors can do much more than just present and perform their thought experiments in specific ways that favor their own intuitions. In fact, it seems that one of the most effective strategies for

establishing certain conceptual intuitions with thought experiments is to insert philosophically irrelevant factors *directly* into the set-up of the thought experiment. For example, consider one of Peter Unger's infamous variations of Judith Jarvis Thomson's *Fat Man*-version of the Trolley Case:

> By sheer accident, an empty trolley, nobody aboard, is starting to roll down a certain track. Now, if you *do nothing about* the situation, your *first* option, then, in a couple of minutes, it will run over and kill six innocents who, through no fault of their own, are trapped down the line. (So, on your first option, you'll let the six die.) Regarding their plight, you have *three other* options: On your *second option*, if you push a remote control button, you'll change the position of a switch-track, switch A, and, before it gets to the six, the trolley will go onto another line, on the left-hand side of switch A's fork. On that line, three other innocents are trapped and, if you change switch A, the trolley will roll over them. (So, on your second option, you'll save six lives and you'll take three.) On your *third option*, you'll flip a remote control toggle and change the position of another switch, switch B. Then, a very light trolley that's rolling along another track, the Feed Track, will shift onto B's lower fork. As two pretty heavy people are trapped in this light trolley, after going down this lower fork the vehicle won't only collide with the onrushing empty trolley, but, owing to the combined weight of its unwilling passengers, the collision will derail the first trolley and both trolleys will go into an uninhabited area. Still, the two trapped passengers will die in the collision. On the other hand, if you don't change switch B, the lightweight trolley will go along B's upper fork and, then, it will bypass the empty trolley, and its two passengers won't die soon. (So, on your third option, you'll save six lives and you'll take two.) Finally, you have a *fourth option*: Further up the track, near where the trolley's starting to move, there's a path crossing the main track and, on it, there's a very heavy man on roller skates. If you turn a remote control dial, you'll start up the skates, you'll send him in front of the trolley, and he'll be a trolley-stopper. But, the man will be crushed to death by the trolley he then stops. (So, on your fourth option, you'll save six lives and you'll take one.) On reflection, you choose this fourth option and, in consequence, the six are prevented from dying. (Unger 1996: 90)

Obviously, the only differences between the above case and Johnson's original case are that in the above case (i) the agent has more options to choose from, and (ii) the agent kills the heavy man in a way that is less direct – instead of pushing him off a bridge, she manipulates his roller skates. Most people will find these differences *philosophically irrelevant*. However, while in the original case most people have the intuition that it would be morally wrong to kill the heavy man in order to save six innocent people, most people respond to the above case by saying that it is permissible to do so (ibid.). Given this, it seems that authors can already design the details of their thought experiments' set-ups in specific ways so as to make sure that their readers accept the 'right' kind of intuition (for a concrete illustration of this strategy, see e.g. Gendler/Hawthorne 2005). This dynamic is a further threat to students' hermeneutical abilities.

What can philosophy instructors do to mitigate this risk? Clearly, the above strategies are of little help in this context. The problem is not that the students' performances of thought experiments are influenced by external factors, but rather that many philosophical thought experiments are already set up in ways that suggest specific performances. To tackle this problem, instructors need to enable students to critically reflect on and challenge the specific set-ups of canonical thought experiments. One obvious way of doing this is to present different variations of the original setup that suggest different performances. For example, when discussing Thomson's version of the Trolley Case, it will be helpful to complement the original thought experiment with one of Unger's variations. By comparing their respective performances of the different variations, students will become increasingly aware of the manipulating effects of philosophically irrelevant details, and thus develop a more critical attitude towards the specific ways in which authors set up their thought experiments.

At the same time, the variations that are required for this approach are often not readily available. Given this, students should learn to independently develop new variations of thought experiments by themselves. Designing and performing their own variations will effectively enable them to test the validity of their performances, and to expose philosophically irrelevant factors that tacitly influence these performances. Although many students will initially find it rather challenging to systematically design their own philosophical thought experiments, instructors can rely here on well-established support strategies that have been developed in the didactical literature (see e.g. Engels 2017: 192ff.; Wieland/Endt 2017). In fact, letting students design their own thought experiments is already a popular didactical tool in philosophy classes, with well-known merits that a lot of teachers appreciate. However, in light of the above considerations, it seems that this tool is not just didactically, but also morally beneficial. It doesn't just promote students' creativity and imagination but also improves their ability to adequately interpret their social and moral reality, which in some cases even constitutes an effective measure against serious forms of educational injustice.

Conclusion

With respect to the important educational goal of improving students' hermeneutical abilities, using thought experiments for the purpose of conceptual clarification is both a blessing and a curse. On the one hand, by providing opportunities to explore the scope of normatively loaded concepts, thought experiments can effectively help students to interpret their social and moral reality more adequately, which in some cases might even help to reduce existing hermeneutical injustices. On the other hand, given their notorious susceptibility to distorting factors that are philosophically irrelevant, they can also push students into accepting idiosyncratic intuitions that they don't really share and thereby further impair their hermeneutical abilities.

In this paper, I have proposed three different strategies for effectively exploiting the empowering potential of thought experiments. These strategies enable students to independently explore their personal conceptual intuitions and to critically challenge the specific and often tendentious ways in which authors present and perform their thought experiments. A successful implementation of these strategies will be an important first step towards a safe and fruitful use of thought experiments in the philosophy classroom.

References

Abarbanell, Linda/Hauser, Marc D. (2010), Mayan morality: An exploration of permissible harms, *Cognition* 115 (2), 207–224.

Adleberg, Toni/Thompson, Morgan/Nahmias, Eddy (2015), Do men and women have different philosophical intuitions? Further data, *Philosophical Psychology* 28 (5), 615–641.

Balg, Dominik (2021), Who Is Who? Testimonial Injustice and Digital Learning in the Philosophy Classroom, *Teaching Philosophy* 45 (1), 1–21.

Berniūnas, Renatas/Beinorius, Audrius/ Dranseika, Vilius/Silius, Vytis/Rimkevičius, Paulius (2021), The weirdness of belief in free will, *Consciousness and Cognition* 87, 103054.

Bettcher, Talia Mae (2009), Trans identities and first-person authority, in Laurie Shrage (ed.), *You've changed. Sex reassignment and personal identity,* Oxford: OUP. 98–120.

Bogardus, Thomas (2020), Evaluating Arguments for the Sex/Gender Distinction, *Philosophia* 48 (3), 873–892.

Brown, James Robert/Fehige, Yiftach (2022), Thought Experiments, in: Edward N. Zalta (ed.), The Stanford Encyclopedia of Philosophy (Spring 2022 Edition), URL: https://plato.stanford.edu/archives/spr2022/entries/thought-experiment/. Last access: 24.05.2022.

Buckwalter, Wesley/Stich, Stephen (2013), Gender and Philosophical Intuition, in Joshua Knobe/Shaun Nichols (eds.), *Experimental Philosophy, Vol.2*, Oxford: OUP, 307–346.

Cameron, C. Daryl, B. Keith Payne, & John M. Doris (2013), "Morality in High Definition: Emotion Differentiation Calibrates the Influence of Incidental Disgust on Moral Judgments", *Journal of Experimental Social Psychology* 49 (4), 719–725.

Corvino, John (2000), Analyzing gender, *Southwest Philosophy Review* 17 (1), 173–180.

Curtin, Cameron M./Barrett, H. Clark/Bolyanatz, Alexander/Crittenden, Alyssa N./Fessler, Daniel/Fitzpatrick, Simon/Gurven, Michael/Kanovsky, Martin/Laurence, Stephen/Pisor, Anne/Scelza, Brooke/Stich, Stephen/von Rueden, Chris/Henrich, Joseph (2020), Kinship intensity and the use of mental states in moral judgment across societies, *Evolution and Human Behavior* 41 (5), 415–429.

Engels, Helmut (2017), Gedankenexperimente, in: Julian Nida-Rümelin/Irina Spiegel/Markus Tiedemann (eds.), Handbuch Philosophie und Ethik. Band I – Didaktik und Methodik. 2., durchgesehene Auflage, Paderborn: Brill, 187-196.

Feltz, Adam/Cokely, Edward (2019), Extraversion and compatibilist intuitions, A ten-year retrospective and meta-analyses, *Philosophical Psychology* 32 (3), 388–340.

Förg, Melanie (2019), The Sense of Wonder. How To Inspire Children to (Continue to) Ask Philosophical Questions, *Questions – Philosophy for Young People* 19, 25–27.

Fricker, Miranda (2007), Epistemic Injustice. Power & the Ethics of Knowing, Oxford: OUP.

Gendler, Tamar Szabó/Hawthorne, John (2005), The Real Guide to Fake Barns. A Catalogue of Gifts for Your Epistemic Enemies, *Philosophical Studies* 124 (3), 331–352.

Gettier, Edmund (1963), Is Justified True Belief Knowledge?, *Analysis* 23 (6), 121–123.

Hannikainen, Ivar R./Machery, Edouard/Rose, David/Stich, Stephen/Olivola, Christopher Y./Sousa, Paulo/Cova, Florian/Buchtel, Emma E./Alai, Mario/Angelucci, Adriano/Berniûnas, Renatas/Chatterjee, Amita/Cheon, Hyundeuk/Cho, In-Rae/Cohnitz,

Daniel/Rasika, Vilius/Lagos, Ángeles Eraña/Ghadakpour, Laleh/Grinberg, Maurice/Hashimoto, Takaaki/Horowitz, Amir/Hristova, Evgeniya/Jraissati, Yasmina/Kadreva, Veselina/Karasawa, Kaori/Kim, Hackjin/Kim, Yeonjeong/Lee, Minwoo/Mauro, Carlos/Mizumoto, Masaharu/Moruzzi, Sebastiano/Ornelas, Jorge/Osimani, Barbara/Romero, Carlos/López, Alejandro Rosas/Sangoi, Massimo/Sereni, Andrea/Songhorian, Sarah/Struchiner, Noel/Tripodi, Vera/Usui, Naoki/del Mercado, Alejandro Vázquez/Vosgerichian, Hrag A./Zhang, Xueyi/Zhu, Jing (2019), For Whom Does Determinism Undermine Moral Responsibility? Surveying the Conditions for Free Will Across Cultures, *Frontiers in Psychology* 10, 2428.

Healey, Richard (2015), The Ontology of Consent. A Reply to Alexander, *Analytic Philosophy* 56 (4), 354–363.

Horvath, Joachim/Koch, Steffen (2021), Experimental philosophy and the method of cases, Philosophy Compass 16 (1), 1–13.

Kim, Minsun/Yuan, Yuan (2015), No Cross-Cultural Differences in the Gettier Car Case Intuition. A Replication Study of Weinberg et al. 2001, *Episteme* 12 (03), 355–361.

Klaiber, Tilo (2018), Die Macht des Beispiels beim Philosophieren, *Zeitschrift für Didaktik der Philosophie und Ethik* 4, 80–94.

Lanphier, Elizabeth/McKiernan, Amy (2020), Thinking about Thought Experiments in Ethics, *Teaching Ethics* 19 (1), 17–34.

Liao, S. Matthew/Wiegmann, Alex/Alexander, Joshua/Vong, Gerard (2012), Putting the Trolley in Order. Experimental Philosophy and the Loop Case, *Philosophical Psychology* 25 (5), 661–671.

Löhr, Guido (2019), The experience machine and the expertise defense, *Philosophical Psychology* 32 (2), 257–273.

Machery, Edouard (2017), *Philosophy Within Its Proper Bounds*, Oxford: OUP.

Machery, Edouard/Stich, Stephen/Rose, David/Chatterjee, Amita/Karasawa, Kaori/Struchiner, Noel/Sirker, Smita/Usui, Naoki/Hashimoto, Takaaki (2015), Gettier Across Cultures, *Noûs*, 51 (3), 645–664.

Machery, Edouard/Sytsma, Justin/Deutsch, Max (2015), Speaker's Reference and Cross-Cultural Semantics, in: Andrea Bianchi (ed.), *On Reference*, Oxford: OUP, 62–76.

Machery, Edouard/Olivola, Christopher Y./de Blanc, Molly (2009), Linguistic and metalinguistic intuitions in the philosophy of language, *Analysis* 69 (4), 689–694.

Machery, Edouard/Mallon, Ron/Nichols, Shaun/Stich, Stephen (2004), Semantics, cross-cultural style, *Cognition* 92 (3), 1–12.

Martena, Laura (2018), Thinking Inside the Box, *Teaching Philosophy* 41 (4), 381–406.

Matthews, Gareth (1979), Thinking in Stories, *Thinking: The Journal of Philosophy for Children* 1 (2), 2-3.

Nichols, Shaun/Knobe, Joshua (2007), Moral responsibility and determinism. The cognitive science of folk intuitions, *Noûs* 41 (4), 663–685.

Nichols, Shaun/Stich, Stephen/Weinberg, Jonathan (2003), Meta-skepticism. Meditations on ethnoepistemology, in: Steven Luper (ed.), *The skeptics. Contemporary Essays,* Ashgate: Routledge, 227–247.

Norcross, Alastair (2004), Puppies, pigs, and people. Eating meat and marginal cases, Philosophical Perspectives 18 (1), 229–245.

Plato (1892), The Republic, in: Raphael Demos (ed.), *The dialogues of Plato. Vol. I*,translated by Benjamin Jowett, New York City: Random House.

Robb, David (2020), Moral Responsibility and the Principle of Alternative Possibilities, in: Edward N. Zalta (ed.), The Stanford Encyclopedia of Philosophy (Fall 2020 Edition), URL: https://plato.stanford.edu/archives/fall2020/entries/alternative-possibilities/. Last access: 24.05.2022.

Schnüriger, Hubert (2018), What is Consent?, in: Andreas Müller/Peter Schaber (eds.), *The Routledge Handbook of the Ethics of Consent*, New York: Routledge, 21–31.

Schulz, Eric/Cokely, Edward/Feltz, Adam (2011), Persistent bias in expert judgments about free will and moral responsibility. A test of the expertise defense, *Consciousness and Cognition* 20 (4), 1722–1731.

Schwitzgebel, Eric/Cushman, Fiery (2012), Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers, *Mind&Language* 27 (2), 135–153.

Schwitzgebel, Eric/Cushman, Fiery (2015), Philosophers' biased judgments persist despite training, expertise and reflection, *Cognition* 141, 127–137.

Seyedsayamdost, Hamid (2015), On Gender and Philosophical Intuition. Failure of Replication and Other Negative Results, *Philosophical Psychology* 28 (5), 642–673.

Starmans, Christina/Friedman, Ori (2009), Is knowledge subjective? A sex difference in adults' epistemic intuitions, Poster presented at the *Biennial Meeting of the Cognitive Development Society*, October 16–17, 2009, San Antonio, TX.

Thomson, Judith Jarvis (1971), A defense of Abortion, *Philosophy and Public Affairs* 1 (1), 47–66.

Unger, Peter (1996), *Living High & Letting Die. Our Illusion of Innocence,* Oxford: OUP.

Weinberg, Jonathan/Nichols, Shaun/Stich, Stephen (2001), Normativity and epistemic intuitions, *Philosophical Topics* 29 (1-2), 429–460.

Wertheimer, Alan (2003), *Consent to Sexual Relations*, Cambridge: CUP.

Wiegmann, Alexander/Horvath, Joachim/Meyer, Karina (2020), Intuitive expertise and irrelevant options, in: Tania Lombrozo/Joshua Knobe/Shaun Nichols (eds.), *Oxford Studies in Experimental Philosophy 3*, Oxford: OUP, 275–310.

Wieland, Jan Willem/Endt, Mathijs (2017), Analysing Thought Experiments, *Teaching Philosophy* 40 (3), 367–383.

How to cite this article

Balg, Dominik (2022): Seeing the World More Clearly. Strategies for unleashing the full moral potential of thought experiments in the Philosophy Classroom, *Journal of Didactics of Philosophy 6*, 1-17. DOI: 10.46586/JDPh.2022.9680.