DOI: 10.1111/meta.12724

ORIGINAL ARTICLE

Danger signals for untrustworthy thought experiments

Henri Tuohimaa

Department of Philosophy, Tampere University, Finland

Correspondence
Henri Tuohimaa, Department of Philosophy, Faculty of Social Sciences, Tampere University, 33014 Tampere, Finland.
Email: henri.tuohimaa@tuni.fi

Abstract
A key question in contemporary metaphilosophy of thought experiments is the "wheat from chaff" problem: How can we separate the good and trustworthy thought experiments from the untrustworthy ones? This article examines this problem by viewing thought experimentation as a form of mental simulation. It argues that we should approach the limitations of thought experiments in light of the general shortcomings of our capacity to run mental simulations. Furthermore, the article proposes an answer to the wheat from chaff problem by presenting three danger signals for untrustworthy thought experiments. These are (1) high counterfactuality of the imagined scenario, (2) complexity of the imagined scenario, and (3) a large psychological distance to the imagined scenario.

KEYWORDS
danger signals, mental simulation, thought experiments

1 | INTRODUCTION

Thought experiments are one of the core methods in contemporary analytic philosophy.[1] Typically, in philosophical thought experiments we are asked to imagine hypothetical scenarios and form judgments about them, and these judgments are then used as a part of philosophical argumentation, for example to corroborate or refute philosophical theories. Famous examples of such argumentative use of thought experiments include Gettier cases (Gettier 1963), Trolley problems (Foot 1967; Thomson 1985), and Thomson's (1971) Violinist case. Thought experiments and their use have been extensively discussed in metaphilosophy, especially after the rise of so-called experimental philosophy (Weinberg, Nichols, and Stich 2001).

What has been called the "negative program" of experimental philosophy has challenged the use and reliability of philosophical thought experiments. Critics claim that the judgments we make about thought experiments often seem to be unreliable and influenced by supposedly

[1]Thought experiments are also known as the method of cases; the terms may be used interchangeably.

irrelevant demographic factors (such as age, ethnicity, and sex) and presentational factors (such as choice of words, order of cases, and smell of paper) (e.g., Machery [2017, chap. 2]). One of the central questions arising from these studies is the "wheat from chaff" problem: How can we separate the good and trustworthy thought experiments from the bad and untrustworthy ones?[2]

Some critics from the negative program (e.g., Machery [2011, 2017]) have tried to answer this problem by explicating "danger signals": that is, features that influence the trustworthiness of our judgments about thought experiments. For example, Edouard Machery (2011, 2017) argues that we use our ordinary reasoning and judgments (whatever they are) in everyday thought experiments, and the problem with many or most philosophical thought experiments is simply that they are out of the proper bounds of this ordinary reasoning, since philosophical thought experiments often depict very atypical and unusual situations. In contrast, a key proponent of thought experiments, Timothy Williamson (2016a), has argued that the danger signals proposed so far by critics haven't yet been clear and useful enough: we still need more precise danger signals and one hopes also an explanation for why these danger signals make a thought experiment untrustworthy.

There is an interesting tension between Williamson's and Machery's views. Both men seem to agree that thought experimentation is based on ordinary reasoning and judgment. Williamson (2016b) takes this as a good reason to trust our judgments about thought experiments, at least in most cases. Machery agrees that our ordinary judgments are most of the time trustworthy but argues that philosophical thought experiments are a special case where these ordinary judgments break down. This leads us to a pretty simple empirical question: How reliable is ordinary reasoning in general, and how reliable when it is applied to philosophical thought experiments?

In this article, I'm proposing an answer to the wheat from the chaff problem from an angle slightly different compared to that of experimental philosophers. I aim to evaluate thought experimentation in light of earlier literature on cognitive psychology and decision-making to understand the ordinary reasoning involved in thought experiments. I contend that thought experimentation is based on our more general cognitive capacity of mental modeling and mental simulation, and hence we can understand the limits of thought experimentation by looking at the general limitations of our mental simulations. Resting on these general limitations, I formulate three new danger signals of an untrustworthy thought experiment: (1) high counterfactuality of the imagined scenario, (2) complexity of the imagined scenario, and (3) a large psychological distance to the imagined scenario.

The plan is as follows. In the first section, I summarize the mental modeling/simulation view of thought experimentation. The second section looks at the key factors influencing the trustworthiness of mental simulations. I start the third section by discussing some desirable features of danger signals for untrustworthy thought experiments, and then propose the danger signals (1) high counterfactuality, (2) complexity, and (3) psychological distance, respectively. Finally, in the fourth section, I draw some conclusions.

2 | THOUGHT EXPERIMENTS AND MENTAL SIMULATIONS

To understand the limits and failures of thought experimentation, we should start by looking at the reasoning involved. A popular view here is the so-called ordinary judgment proposal, according to which "the judgments elicited by thought experiments are underwritten by the psychological capacities that also underlie the judgments we make about everyday situations" (Machery 2011, 194). On this view, some critics of thought experiments (like Machery [2011,

[2]See also Weinberg's (2016) characterization of what he calls the "wheat from chaff project" in experimental philosophy.

2017]) and some proponents of thought experiments (like Williamson [2016a, 2016b]) seem to agree (see also Hawke and Schoonen [2020]). But what exactly are these ordinary psychological capacities used in thought experiments?

One of the most promising answers to this question is the mental modeling view (MMV), which claims that thought experimentation is based on active simulation and manipulation of mental models in our imagination (Nersessian 1992, 2008, 2018; see also Miscevic 1992, 2021; Alexander 2016).[3] In short, one of the main proponents of MMV, Nancy Nersessian contends that thought experimentation uses and manipulates mental representations that come from real-life experiences. When we form these kinds of simulations in our imagination, we can use our explicit and tacit knowledge about the processes, entities, causal structures, and so forth involved (Nersessian 2018, 319–22.)

In this line of thinking, thought experimentation is taken to be based on a more general capacity of mental modeling or mental simulation, which is widely used in different kinds of situations, from decision-making, reasoning, and problem-solving to retrospection and prospection.[4] In our everyday life, if we are thinking about major decisions like having a child, choosing a spouse, or changing jobs, we often use our ability of mental simulation: we imagine ourselves in these different situations so we can preview and "prefeel" our happiness or satisfaction in the situation of interest (Gilbert and Wilson 2007, 1352). The mental-modeling view thinks that we make judgments about thought experiments in the same way. We simulate the case in our imagination (such as a runaway trolley racing toward five railroad workers), and based on this simulation we make a judgment about the case (for example, it would or wouldn't be morally permissible to force the trolley onto a second track, where there is only one railroad worker).

It is interesting to note that Williamson (in 2016b) seems to agree with this kind of view when he claims that humans routinely get knowledge through their imagination—for example, a person who doesn't know if they can jump over a certain river often imagines jumping over the river to check if they could make it. Williamson (2016b, 122) then claims that thought experiments are not a distinctive method of philosophy but rather just a more "elaborate and eye-catching" use of exercising our imagination. Even though Williamson writes about "exercising imagination" and sometimes (e.g., in Williamson 2007) about "rational imagination," it seems to me that he is pretty much referring to mental modeling and mental simulations (see also Alexander 2016).

For simplicity's sake, I will be framing the argument in this paper in terms of the limitations of mental simulations, but it should be said that the limitations I discuss and hence the danger signals that I propose apply also to intuitive judgment in general. This is because the factors that make a mental simulation accurate and trustworthy are the same factors that influence the trustworthiness of our intuitions (see Kahneman and Klein 2009). Hence, we should expect the judgments we form about thought experiments that exhibit danger signals to be untrustworthy, even if we sometimes get an instant gut-reaction intuition about them without actively using our imagination.

Assuming that all this is right, and we actively exercise our imagination when we face and judge philosophical thought experiments, the key question that follows is this: When and why should we trust our imagination, and when should we expect our imagination to lead us astray? Next, I propose a precursory answer to this question by summarizing some key findings in psychological research on mental simulation and its limitations.

[3]While Nersessian's work mostly focuses on thought experiments in science, Alexander (2016), for example, has applied the MMV-framework more specifically to philosophical thought experiments.

[4]For a general discussion of this psychological capacity, see, e.g., Klein 1998; Gilbert and Wilson 2007; Kahneman and Klein 2009; Markman, Klein, and Suhr 2009.

3 | THE LIMITS OF MENTAL SIMULATIONS

Mental simulations, like any other type of human reasoning, work quite well in everyday situations—for example, when we need to simulate the best way to carry a sofa in a tight stairway or imagine fun activities for a birthday party. Psychological research on mental simulations suggests, however, that in many situations our imaginative capacities aren't even close to as good as we would like to think, as will be shown in the following pages. The main takeaway is that the ability to form accurate mental simulations of particular scenarios depends crucially on our experience, knowledge, and understanding of similar situations. In short, the more we have experience and knowledge of something (for example, birthday parties), the better we can simulate them in our mind, and vice versa. In this section I present four major factors that are relevant for the making or breaking of an accurate mental simulation: (i) earlier experience with similar situations, (ii) relevant background knowledge, (iii) simplicity of the imagined situations, and (iv) practice and feedback.

3.1 | Earlier experience with similar situations

The first factor that influences the accuracy of a mental simulation is our memories and earlier experiences of similar situations.[5] When we form and run mental simulations, we use our memories and experiences (both implicitly and explicitly) as a building block for the simulation, and hence our simulations can only be accurate if the memories/experiences used in the simulation are themselves accurate, relevant, and representative (see, e.g., Gilbert and Wilson 2007, 1352–53). This is also why the complete lack of experience is a bad sign for the accuracy of our simulations.

Klein (1998, chap. 10) illustrates this point with a case of parents taking their children to kindergarten for the first time: when the child clings and doesn't want to leave their parents, the parents typically will feel devastated and fear that the rest of the year will go like that. A skilled kindergarten teacher who sees the same situation isn't worried at all, since they have seen the same situation hundreds of times and know that the same child will be running in with joy by the end of the week. Here the teacher has the necessary earlier experience that enables them to easily mentally simulate how the rest of the year will progress, whereas the parents have seen only the first day and tend to misimagine the future on the basis of this limited first-day experience.

Another example of the consequences of experience (and the lack of it) can be seen when healthy people try to imagine the effects that chronic illnesses and disabilities will have on their lives. They often expect that serious illnesses or disabilities will drastically lower their quality of life, whereas people who actually have serious chronic illnesses or disabilities report having a "paradoxically" high quality of life and happiness (Ubel et al. 2005). For example, healthy people imagine that becoming paraplegic would result in a long-term decrease in their happiness, even though there are only small differences in the happiness of paraplegics and control groups, and people seem to return to their prior happiness levels fairly quickly after a major accident (Brickman, Coates, and Janoff-Bulman 1978; Schkade and Kahneman 1998). Interestingly, Schkade and Kahneman (1998, 340) point out that just knowing a paraplegic person significantly changes the predicted happiness, and they claim that "[t]he message is clear: The less you know about paraplegics, the worse off you think they are."

As a final caveat, one should note that experience isn't always enough. For example, in the case of affective forecasting people notoriously overestimate their emotional reactions even in cases where they do have experience, such as failing a driving test for a second time (Ayton,

[5] Klein 1998; Gilbert and Wilson 2007; Kahneman and Klein 2009; Nersessian 2018.

Pott, and Elwakili 2007), visiting a dentist, or going to a theater (Wilson and Gilbert 2003; Gilbert and Wilson 2007).

3.2 | Relevant background knowledge

When we form and execute mental simulations, we use our knowledge of the relevant objects and phenomena (both implicit and explicit) in the same way we use our earlier experience. What follows is that our simulations will be accurate representations of reality only when the background knowledge and theories used are accurate. And if it happens that we believe some wildly false theory on some subject, our imagination on that subject is probably going to be aligned with the wildly false theory, and hence prone to lead us astray (McCloskey 1983; Wilson and Gilbert 2003, 360–61).

The impact of background knowledge and beliefs can be seen when people run mental simulations about the trajectories of objects. For example, when subjects are asked to imagine and draw the path of a metal ball (or a bomb) that is dropped from an airplane flying at a constant speed, they often get the path totally wrong (see McCloskey, Caramazza, and Green 1980; McCloskey 1983; Stanovich 2003, 81–95). Training in physics expectedly improves these kinds of judgments, and professional physicists are often very good at solving similar problems. Interestingly, for example, McCloskey (1983, 321) argues that untrained subjects seem to employ "naïve theories of motion" besides using analogs to isolated facts and experiences, like memories of throwing a rock with a sling.[6] Noteworthy here is that the background knowledge and understanding of the expert physicists allow them to accurately simulate a large variety of cases. In a similar manner, lay people make use of their background theories and beliefs, but since these naive theories are often imperfect or false, they often lead them to errors.

It should be noted that it might not always be easy to discern the background knowledge and theories from our experiences and that, in most cases, these might merge into a mess of understanding or knowledge. If we again take the example of a kindergarten teacher, it is quite hard to tell how much (and what parts) of their understanding comes from their experience of the work and how much comes from, let's say, education and theoretical knowledge. I take the distinction between knowledge and experience to be a useful one, however, because it seems that in some cases these might substitute for each other—that is, a wide base of real-life experience might substitute for a lack of explicit theoretical knowledge, and vice versa.

3.3 | Simplicity of the imagined scenario

The third factor needed for successful mental simulation is the simplicity of the scenario. For example, Klein (1998, chap. 5) contends that limitations of our (working) memory pose some limits for our mental simulations. Klein estimates that people usually use no more than three variables and six transition steps. If the scenario is much more complicated than this—that is, there are too many moving parts and too many interactions to take into account—our working memory can't grasp the scenario. To combat these memory limits, people often radically simplify the imagined scenarios by leaving out "nonessential parts," but this simplification can lead to other troubles, since simplifying needs a good amount of experience, knowledge, and expertise (Klein 1998; Gilbert and Wilson 2007).

[6]These theories interestingly resemble pre-Galileo medieval theories and concepts of mechanics.

To borrow an example from Klein (1998), a professional firefighter who sees the site of a car accident can pretty easily use their imagination to come up with the best rescue plan. Because of their earlier experiences, background knowledge, and understanding they can run mental simulations of different strategies, notice and account for the crucial details of this particular case, and so on. Car accidents of course are actually quite complicated phenomena, and there are quite a lot of moving parts, but since our firefighter is an expert on this subject, they can leave irrelevant details aside and form a simplified mental simulation of the situation. All this is something that most people without training can't do properly. In the same situation, a random person on the street might try to simulate a haphazard rescue plan, but this simulation would probably be unduly complex and chaotic, or simplified in all the wrong ways, such that it wouldn't factor in, for example, the necessity of supporting the necks of injured people. In addition to these shortcomings (depending on the number of action movies recently viewed) the mental simulation might even contain some outright false beliefs about exploding cars and so forth, which might lead to further problems.

The rescue example above is simple enough for an expert, but more complex scenarios and phenomena are often out of reach even for experts in the field. Typical examples include long-term forecasts in finance, economics, and politics (Kahneman and Klein 2009, 520; Tetlock 2005). Large-scale political or economic events and developments are just too complicated to reasonably simulate in one's mind, since there are too many actors and interactions, and too much uncertainty.

3.4 | Practice and feedback

To summarize earlier points, accurate mental simulations need earlier experience, relevant background knowledge and understanding, simplicity, and the skill to simplify the scenarios. In general, the lesson is the following: Experts in different fields who have experience, a wide base of knowledge, and the capacity to simplify scenarios are often able to forecast and solve problems in their area of expertise. Vice versa, people with little experience and knowledge about a certain situation are not.[7] To put this last point simply, experience and background knowledge aren't always enough. Using our imagination and running mental simulations are cognitive skills, and hence one must practice and get feedback to improve these skills (see Pliske, McCloskey, and Klein 2001; Ross, Shafer, and Klein 2006, 412–13). The practice part is probably self-explanatory, but it is the feedback that often poses problems because not all feedback is good feedback.

The research on human expertise and expert judgments has shown that experts in many fields, ranging from stockbrokers and court judges to psychiatric diagnosticians and clinical psychologists, quite often make poor judgments and forecasts (see, e.g., Shanteau 1992 for a review). The key feature common in these fields is either a lack of feedback or bad, that is, sparse, delayed, and ambiguous, feedback (Kahneman and Klein 2009, 523). This is to say that we can improve our judgments and make them more accurate only if we can know within a reasonable time frame what we got right and what we got wrong. The same goes more specifically for our mental simulations: we can imagine stuff (like becoming a famous millionaire) as much as we want, but if we don't get good and accurate feedback on our imaginings, the simulations probably won't get any more accurate.

So, to summarize, I have presented four key factors influencing the trustworthiness of our mental simulations: (i) earlier experience, (ii) background knowledge, (iii) simplicity of the scenario, and (iv) practice and feedback. Of course, much more could be said about the factors behind a successful mental simulation, and this list is not supposed to be exhaustive. Still, I

[7]See, e.g., Klein 1998; Gilbert and Wilson 2007; Kahneman and Klein 2009.

hope that the elements discussed above are enough to provide a starting point for the ensuing metaphilosophical discussion of thought experimentation and its limitations.

4 | THE DANGER SIGNALS

To get started with this section, it is worthwhile to briefly discuss some general goals of the danger signals. From Williamson's (2016a) response to and critique of the danger-signal candidates proposed earlier (e.g., Machery 2011; Alexander and Weinberg 2014), we can derive some desirable features of the danger signals, on which (one hopes) the critics and defenders of thought experiments can agree.

First, the danger signals should be compatible with the baseline accuracy of our judgments in everyday contexts. For example: arguing that our capacity to make knowledge attributions is unreliable and that the problem with Gettier cases is that we use this unreliable capacity isn't a very attractive danger signal, since the baseline accuracy of our knowledge attributions seems to be quite high (see, e.g., Boyd and Nagel 2014; Alexander and Weinberg 2014). The danger signals we pose should be compatible and account for this kind of baseline accuracy, at least to the extent that we don't have a good reason to suspect the accuracy of some types of judgments altogether.

Second, the danger signals shouldn't overgeneralize skepticism. For example, Machery (2011, 2017) has argued that one problem with philosophical thought experiments is that “thought experiments typically pull apart the features that go together in everyday life: for instance, using physical violence and doing more harm than good are pulled apart in the footbridge case” (Machery 2011, 203). Williamson (2016a, chap. 2.3.) claims, however, that this proposed danger signal wildly overgeneralizes skepticism: it is not rare at all that real-life cases pull apart features that typically go together, be it “physical violence and doing harm,” “justified true belief,” or “proper name and correct description.” If just pulling apart features that often go together would be enough to warrant skepticism, then in Williamson's view it would follow that we should be skeptical of our judgments when we judge, for example, that a person was right to have kicked their assaulter to defend themselves, but this is probably a conclusion that we don't want to make.

Third, the danger signals should be useful in the sense that they are easy to recognize and track. Williamson (2016a) contends that, for example, “atypicality,” “marginality,” and “subtlety” are not very clear attributes of a case and hence, at least without good further definitions, not very useful. This is mainly because many real-life cases at minimum seem “atypical,” “marginal,” and/or “subtle,” but we wouldn't want to say that our judgments are warrantless in those cases. Relating to the easy-to-track clause, one could propose the presence of random presentational or demographic influences in some thought experiments as a danger signal (see Machery 2017, 118–20). This, however, might not be a very useful (even if true) danger signal, since one cannot actually know from one's armchair if these kinds of influences are present or how strong they are in some particular thought experiment.

All the danger signals I'm going to formulate here are not on-off criteria but fit on a scale: from actual-world cases to highly counterfactual, from simple to complex, and from psychologically proximate to psychologically distant. The limits on how dangerous the thought experiments we may use depend on how much unreliability we are ready to tolerate in our methods. I don't address this question in this paper but instead just propose a list of the attributes we should be looking at when trying to determine the trustworthiness of a thought experiment.

A final caveat before introducing the danger signals: the presence of the proposed danger signals and hence the absence of the key factors needed for accurate mental simulations doesn't mean that we can't run mental simulations of such cases. The danger signals affect only the accuracy of the mental simulation, since if we are, for example, missing relevant experience and knowledge of a certain situation, our imagination relies on semi-random and probably irrelevant pieces of knowledge or/and the inaccurate or false background beliefs we have. Again, I can certainly imagine what my life would be like if I were to be a famous millionaire rockstar, but since I don't have experience or know anything about being famous, or a millionaire, or a rockstar, my imagined rockstar life probably won't resemble anything like the reality.

All this said, let's now move on to my proposed danger signals for untrustworthy thought experiments: (1) high counterfactuality, (2) complexity, and (3) psychological distance.

4.1 | High counterfactuality

The first danger signal I'm proposing is a high counterfactuality of the scenario depicted in a thought experiment.[8] The philosophical literature is filled with wild counterfactual thought experiments that ask us to imagine brain swapping (Williams 1970), experience machines (Nozick 1974), twin earths (Putnam 1975), teletransportation (Parfit 1984), and so forth. In earlier literature on thought experiments and their trustworthiness, the degree of counterfactuality has perhaps been the most discussed factor supposedly influencing the trustworthiness of our judgments.[9] I stick with the term "high counterfactuality" here because in my opinion it's the most precise, but there are of course other options, like "sci-fi-cases," "exotic cases," "unusual cases," "outlandish cases," and so on, which have all been used in earlier discussions. Let's start with an example.

Lippert-Rasmussen (2008) argues against the substantive self-ownership thesis (that is, people have substantive, fact-insensitive, and freestanding moral ownership of their bodies and minds) by using thought experiments like the following:

> Let us consider an eye redistribution scheme in which half the population is born with two pairs of eyes and the other half with no eyes. In sighted individuals, one pair of eyes is located normally and fulfills the usual function. The other pair is located inside the human body, say, in the shoulder. Although this latter pair would enable those who have them to see if they were surgically moved to the eye sockets, they play no role where they are. Indeed they cannot perform any visual or other bodily function without being moved. Suppose further that the body of a person born with two pairs of eyes will expel the spare pair when that person reaches twenty years of age. The pair can then easily be reabsorbed into the shoulder of its owner, or the owner can transfer his spare eyes to a blind person. Typically people will feel alienated from their spare pair of eyes in the same way that the main character in Sartre's *Nausea*, Antoine Roquentin, feels alienated from his hand when he perceives it as a crab lying on its back moving its legs and claws. As it happens, the state implements a compulsory redistributive scheme in which whenever a person's body expels a nonoperative pair of eyes, these are appropriated and transplanted into a blind person, giving him normal sight. (Lippert-Rasmussen 2008, 98).

[8]We should separate the counterfactuality that is relevant to the case from the superficial counterfactuality. It is of course possible to raise the degree of counterfactuality just by changing random details that are not relevant to the case—for example, the number of arms and eyes Jim has in a Gettier-type clock case. In contrast, the counterfactuality in teletransportation cases is relevant because we cannot simply cut out the sci-fi elements or substitute the teletransporter with something more ordinary because the teletransporter has no counterparts in our world.

[9]One can find this kind of argument in one form or another in Wilkes 1988; Brendel 2004; Gendler 2010; Elster 2011; Machery 2011, 2017; Zarnitzyn 2015, 64; Pesonen 2019, 303; Hawke and Schoonen 2020.

Typically, the argument against highly counterfactual thought experiments goes something along the following line of reasoning. In ordinary life, we are used to making judgments (about right or wrong, knowledge attribution, reference, and the like) about our world and real-life cases. In highly counterfactual thought experiments, though, we are asked to imagine scenarios that differ wildly from our world. In these thought experiments, we really have no experience of such highly counterfactual scenarios and no way to know how they would work and what facts would hold. Since our capacities to make judgments (about right or wrong, knowledge, reference, and so forth) have developed to work in ordinary real-life cases, we shouldn't expect them to be reliable in wildly counterfactual scenarios where we don't have and by definition even cannot have experience (see, e.g., Elster 2011, sec. 6; Machery 2017, 113–15). I take this line of reasoning to be on the right track, but I think we can make this point even more persuasive by looking at the general limitations of mental simulations.

If we grant that accurate and reliable mental simulations need (i) earlier experience with similar situations, (ii) relevant background knowledge, (iii) simplicity of the scenario, and (iv) practice and feedback from our simulations and judgments, we can start to see the problem with highly counterfactual thought experiments. They simply do not seem to fulfill any of these criteria.[10]

Let's start with the earlier experience. The earlier-experience criterion shouldn't warrant much discussion, since we, by definition, can't have it in highly counterfactual cases. After all, if we had or could have experience with, for example, teletransporters or experience machines, these cases would not be highly counterfactual. The further we go from our world, the less we have and can have experience, and hence the more we should be suspicious about the accuracy and trustworthiness of our imagination.

When it comes to relevant background knowledge, highly counterfactual thought experiments don't fare much better. For example, Wilkes (1988), Elster (2011), and Machery (2017) have all argued that the crucial problem with highly counterfactual thought experiments is precisely that we do not have relevant background knowledge to imagine or judge them. The point here is that the knowledge we have about our actual world and, for example, the physical, social, biological, or chemical facts involved becomes less applicable and useful the further we go from the actual world. For example, if we try to judge the acceptability of the redistribution policy in the four-eyes case above, we really have no idea how the peculiar eye distribution affects other areas relevant to our judgment, like social order, psychological features of the people, or the rights that people have over their bodies (see Elster 2011 for further discussion of this case). These kinds of interrelated and unknown implications of counterfactual suppositions also raise the complexity of thought experiments: changing the fact about the prevalence of eyes would have a ton of implications for other areas of life, but trying to imagine and track these will quickly make our simulation extremely complicated.

It is reasonable to ask what the relevant background knowledge even could be in highly counterfactual thought experiments. The most plausible answer would probably be our philosophical theories themselves. But if the background knowledge used in philosophical thought experiments is our philosophical theories, this leads to problems. First, inaccurate or false theories lead to inaccurate simulations, as with lay people's mental simulations about trajectories of objects that obey naive folk theories of motion. But since these folk theories are often wrong, people end up making errors (see McCloskey, Caramazza, and Green 1980; McCloskey 1983). This is especially a problem for philosophy, since we rarely know if our theories are correct or which one of the competing theories is the right one. The simulations and judgments I make about, for example, brain-swapping cases (as in Williams 1970) might correspond to the theory of personal identity I believe, but if other as-good competing theories lead to different conclusions, my judgments become questionable. Second, this lack of independent background knowledge might lead to a problem of circularity. Judgments made about thought experiments

[10]Perhaps sometimes excluding the simplicity criterion.

are often used to support and refute philosophical theories, but if we form a particular judgment because we already believe the matching philosophical theory, it seems quite problematic to then use this judgment again to support this theory.

As a side note, it should be said that in other areas of study, the high counterfactuality of a thought experiment might not be as big of a problem. For example, physics seems to employ "highly counterfactual" thought experiments (see Schindler and Saint-Germier 2020), but I would argue that in physics the status of background knowledge and theories is crucially different. In physics, unlike in philosophy, we have strong and mostly uncontested theories like classical mechanics and the theory of relativity, which seem to provide a good basis of background knowledge and understanding that can guide us through even the weirder counterfactual cases. In philosophy, however, high counterfactuality is an especially vicious problem precisely because we don't have these kinds of strong and uncontested theories that could fill the same role.

Third, there is the problem of feedback. Again, given the nature of highly counterfactual scenarios, we cannot realize such cases and hence have no possibility of getting feedback regarding the accuracy of our simulations. No matter how much we try to imagine the world of the four-eyes case, if we never get useful feedback on our imaginings and errors, we won't be able to improve our judgments. We could even say that highly counterfactual thought experiments are "hopeless" (see Weinberg 2007), since there is no possibility of error detection and error correction. Here one could argue, like, for example, Williamson (2011), that philosophers tend to get a lot of feedback on their judgments from other philosophers. But while I do think that the feedback from other philosophers is crucial for the general development of philosophical understanding and competence, I don't think it can save highly counterfactual thought experiments. This is simply because just any random feedback doesn't enhance our judgments, only good feedback does (see, e.g., Kahneman and Klein 2009). In the case of highly counterfactual thought experiments, we have fair reason to think that the feedback coming from other philosophers isn't very good, since no one has superior epistemic access to crazy sci-fi worlds. The discussions between philosophers might, of course, align their judgments with one another, but this conformity is as likely to lead them away from the truth as it is likely to lead them to the truth. After all, a group of political scientists who try to forecast the state of the world in the 2500s can discuss and give feedback to each other, but do we really believe that this feedback leads them to accurate forecasts?

4.2 | Complexity

If we grant (as claimed in, e.g., Klein 1998 and Kahneman and Klein 2009) that our capacity to run mental simulations works only in relatively simple cases, one clear danger signal seems to follow: namely, the high complexity of the imagined scenario. Klein (1998, chap. 5) claims that typically successful mental simulations use about three variables. If this is even remotely right, we can start to see the problem in philosophical and political thought experiments where we try to imagine, for example, humankind in the "state of nature" or whole societies with some tweaks.

To give a more precise definition, the complexity of a thought experiment depends on the number of factors/objects in the simulation, the number of interactions between these factors/objects, and sometimes on the number of sequences or transitions in the simulation; but this might not very often be relevant to philosophy. So, a simple thought experiment might consist of me (or some random Jim) looking at a clock and forming a belief that "it is 10 A.M." Here we must consider only one person (me/Jim), one object (the clock), one interaction (between me/Jim and the clock), and finally the mental state of a belief (mine/Jim's). In contrast, if we try to imagine whether the Second World War would have been avoided if Hitler had died in the First World War, we have to imagine, just to get started, the rise and development of the Nazi Party, the goals of most European governments in that time frame, the alliances and other connections between different countries, and the central causes of the Second World War.

The impact that the complexity of a thought experiment has on the accuracy of our judgments has been widely discussed in the context of political and historical thought experiments (see, e.g., Tetlock and Belkin 1996). In philosophy, the factor of complexity has for now received less attention, where, for example, Mayo-Wilson and Zollman (2021) are notable exceptions.

Mayo-Wilson and Zollman (2021) argue that the thought experiments used in philosophy are often far too complex to imagine and simulate accurately. They suggest (2021, 3650) as an example Kant's thought experiments in the *Groundwork of Metaphysics of Morals*: Kant tries to imagine a world where everyone breaks promises when it suits them, and he then concludes that in such a world no one would believe promises and this in turn would destroy the act of promising. Later Mayo-Wilson and Zollman (2021, 3658) compare the complex thought experiments in philosophy to the successful thought experiments in physics, noting that the thought experiment used by Kant is much more complex than, for example, the thought experiments used by Galileo: “Kant asks us to imagine a society filled of people breaking promises, whereas Galileo asks us to imagine a single ball and plane.” Relevantly for philosophy, Mayo-Wilson and Zollman add that social systems are by default more complex than mechanical/physical ones, since the number of variables and the number of interactions in them is often much greater. According to Mayo-Wilson and Zollman (2021, 3659), all this is a good reason to doubt our judgments in such cases, and this conclusion seems fairly plausible if we accept the limitations of mental simulations posed by, for example, Klein and Kahneman.

To summarize, a crucial problem in complex thought experiments is that we simply can't account for and keep track of all the relevant factors involved. There is, however, another problem with imagining complex phenomena that relates to the possibility of background knowledge: even if we could overcome some of our memory limitations, we rarely know all or even most of the relevant factors and their interactions in complex scenarios and environments to begin with. This is one reason why simulating social dynamics, societies, large-scale historical events, and economies is so hard or even impossible.

As a final caveat, it should be said here that I think that in philosophical thought experiments the danger signal of complexity rarely comes alone. Most of the time, the highly complex thought experiments are also highly counterfactual, which means that there are at least two different features disturbing our mental simulations.

4.3 | Psychological distance

The last danger signal I am proposing here is psychological distance. I use this term as it is defined in cognitive psychology: psychological distance means egocentrically determined distance a person has from some object, situation, or phenomenon, where the reference point is that person in that time, place, and situation (see, e.g., Trope and Liberman 2010). The basic claim I want to make here is that our imagination isn't untrustworthy only in highly counterfactual cases and unduly complex scenarios; even in simple realistic/real-world cases, our mental simulations are not always accurate (see, e.g., Wilson and Gilbert 2003; Gilbert and Wilson 2007; and for philosophical discussion Maibom 2016; Arpaly 2020). Take some examples from psychological literature: people often imagine different experiences to be much better or much worse than they really are (such as winning a lottery, going to the dentist, living in California (Schkade and Kahneman 1998; Wilson and Gilbert 2003; Gilbert and Wilson 2007). Most women imagine that they would be angry and confront the perpetrator if they were asked sexually harassing questions in job interviews, whereas harassed interviewees are very rarely confrontational and report a dominant feeling of fear instead of anger (Woodzicka and LaFrance 2001). And, as pointed out in section 3 above, healthy people misimagine and overestimate the impact chronic illnesses or disabilities will have on their life (Ubel et al. 2005).

It seems uncontroversial that in the highly counterfactual scenarios nobody has earlier experience of, knowledge of, or expertise about the scenarios. But with realistic/real-life cases there is a crucial difference: people have different degrees of experience, knowledge, and so forth in different cases. In other words, while the highly counterfactual scenarios are equally distant for all of us, the real-life cases are distant for some and familiar to others.

For example, nobody can accurately imagine what it would be like to be a vampire, or whether they should turn into one if possible, or how this transformation into a vampire would affect morals because no one (at least one hopes) has relevant earlier experience of, relevant knowledge of, and expertise about vampirehood.[11] In contrast, if we are considering having a child, we should expect the accuracy of our mental simulations to vary depending on our relevant earlier experience. A person who already has several children can probably imagine pretty accurately what it would be like to have another child. Someone who has no children, never had younger siblings, and hasn't even seen a baby since they were in kindergarten themselves probably won't be accurate at all in their simulations of what having a child would be like. And finally, a person who has no children themselves but has much younger siblings, some young nieces or nephews, and has done a lot of babysitting lies somewhere between these two persons in their capability to run mental simulations and make decisions on the subject of becoming a parent.

Firsthand experience isn't, of course, always needed to have a good grasp of a situation. For example, I would expect that a sociologist studying drug addicts would have a good understanding of drug addiction and related activities, even though they have never abused drugs themselves. Hence, our sociologist might be quite good at imagining how it would feel to have an addiction or what they would and what they should do in the types of harsh situations they study. I would claim, however, that this doesn't mean that our sociologist can accurately imagine situations that are psychologically distant from them but instead means that the academic and clinical background knowledge they have makes the aforementioned scenarios psychologically closer to them, even though they don't have actual firsthand experience of them.

Turning to the philosophical thought experiments, my hypothesis is the following. The larger the psychological distance some person has from some scenario, the more skeptical we should be about that person's judgments about that particular case. If we again remind ourselves about the factors influencing the accuracy of mental simulations, this conclusion seems very plausible. Large psychological distance implies that we do not have experience or other background knowledge about the scenario, though such experience or knowledge is precisely what is needed to accurately simulate the scenario. Relatedly, I suppose that large psychological distance most of the time also means little practice in simulating the scenario and little if any feedback on judgments (excluding some random cases where, for example, I try to imagine being a parent and some person who is a parent gives me feedback on my imaginings and corrects the errors I make). Lastly, it is worth noting that often even simple-looking thought experiments are not so simple after all. To continue with the parenthood example, it seems that we only need to imagine one thing—namely, having children—being different. But actually, as pointed out by Laurie Paul (2014), becoming a parent changes an unprecedented number of things, ranging from hormonal and physiological changes to values, house tidiness, and amount of sleep. We simply cannot even know, much less account for, all the relevant factors in our mental simulations, and for this very reason we are prone to misimagine such situations.

The suspicious use of psychologically distant thought experiments is perhaps most common in the field of ethics. For example, Frances Kamm's *The Moral Target: Aiming at Right Conduct in War and Other Conflicts* (2012) is filled with realistic (and sometimes even

[11] This amusing example is borrowed from Paul (2014).

actual) cases that relate (as the topic of the book suggests) to moral conduct in war and other tragic conflict situations. For instance, Kamm uses the following case when discussing the permissibility of violent resistance: “Consider a hypothetical case called Locked in the House: Somebody will keep you (an innocent person) locked up in a house permanently. (He thinks God commands this as the appropriate way to treat you.) He is not going to kill you or physically harm you so long as you stay in the house, and you know that this is so. In one variant of this case, if you leave, he will attack you and even kill you. It seems to me that if the only way you could escape is by killing the perpetrator, it would be permissible to kill him” (Kamm 2012, 219).

Take a second example. When discussing the permissibility of collaboration with an enemy or oppressor, Kamm uses the case of “Judenräte”: “On the other hand, we have Jews in the Judenräte (the Jewish community's governing bodies during World War II) whose members were assigned by the Nazis to select people to be taken either to work or to concentration camps, where they would eventually die or be killed. It is said (suppose correctly) that at least some members of the Judenrat selected these Jews for the Nazis in order to make it the case that there was less suffering and damage overall than the Nazis would have caused if they had done the task themselves” (Kamm 2012, 116).

These cases aren't far-fetched science fiction or highly complex but are instead seemingly simple and depict (tragically) realistic situations from our actual world.[12] Still, I would contend that a well-off philosopher from a peaceful and safe country shouldn't be very confident in their imagination and judgments about these thought experiments.[13] This is because these situations, even though real, are probably psychologically extremely distant for many academic philosophers. And as I argued earlier, our mental simulations of such psychologically distant situations are prone to go wrong if we do not have experience or knowledge of anything like the cases above.

Since there hasn't so far been much discussion of the role of psychological distance in philosophical thought experiments, I want to address a few crucial questions for the future. First, since psychologically distant scenarios are often real or at least resemble real scenarios, we can start thinking about the ways we could overcome the problems psychological distance poses. Should we prioritize the (for example, ethical) judgments made by people who have actually experienced situations like those depicted in thought experiments? If we want to know when it is acceptable to sacrifice lives to save more lives, maybe we should study the people who have actually made such decisions instead of thought experimenting in an armchair. And for those who are not interested in leaving the comfortable armchair, maybe there might be other possible ways, like literature, to increase our understanding and close the psychological gap to unfamiliar situations. Second, the fact that there are people who have had (often probably traumatic) experiences with many types of scenarios depicted in philosophical thought experiments raises some quite obvious ethical concerns about the use of these kinds of psychologically distant cases. But I will leave these discussions for later.

5 | CONCLUSION

To sum up, then, philosophical thought experiments and their trustworthiness have been one of the main disputes in the recent metaphilosophical literature. A crucial question in these

[12]This becomes obvious, for example, with the “locked in the house” case, when Kamm later goes on to discuss the house arrest of Afghani women under the Taliban regime.

[13]To be clear, I'm of course not suggesting or assuming anything about Kamm's background or proximity to these cases specifically.

debates concerns demarcating good and trustworthy thought experiments from untrustworthy ones. Critics from the tradition of experimental philosophy, like Machery (2011, 2017), have tried answering this question by explicating disturbing factors or danger signals that affect the trustworthiness of a thought experiment. Proponents like Williamson (2016a) have not yet been convinced and have called for more precise and useful danger signals. In this article, I have proposed an answer to this problem.

I argued that we can start to understand the limitations of thought experiments by looking at the psychological reasoning processes underlying thought experimentation. I claimed (in the vein of Nersessian 1992 and Miscevic 1992) that philosophical thought experiments and thought experimentation rely on our more general cognitive capacity of mental simulation, and I summarized some key factors that influence the reliability of our mental simulations: (i) earlier experience, (ii) background knowledge, (iii) simplicity of the scenario, and (iv) practice and feedback. Resting on these general limitations, I proposed a new list of danger signals of an untrustworthy philosophical thought experiment: (1) high counterfactuality of the scenario, (2) complexity of the scenario, and (3) high psychological distance to the imagined scenario. I hope that these danger signals can provide a point of agreement for the naturalistically oriented proponents and critics of philosophical thought experiments.

REFERENCES

Alexander, Joshua. 2016. "Thought Experiments, Mental Modeling, and Experimental Philosophy." In *Advances in Experimental Philosophy and Philosophical Methodology*, edited by Jennifer Nado, 53–68. New York: Bloomsbury Academic.

Alexander, Joshua, and Jonathan M. Weinberg. 2014. "The 'Unreliability' of Epistemic Intuitions." In *Current Controversies in Experimental Philosophy*, edited by Edouard Machery and Elizabeth O'Neill, 128–45. New York: Routledge.

Arpaly, Nomy. 2020. "What Is It Like to Have a Crappy Imagination?" In *Becoming Someone New: Essays on Transformative Experience, Choice, and Change*, edited by John Schwenkler and Enoch Lambert, 122–33. Oxford: Oxford University Press.

Ayton, Peter, Alice Pott, and Najat Elwakili. 2007. "Affective Forecasting: Why Can't People Predict Their Emotions?" *Thinking and Reasoning* 13, no. 1: 62–80.

Boyd, Kenneth, and Jennifer Nagel. 2014. "The Reliability of Epistemic Intuitions." In *Current Controversies in Experimental Philosophy*, edited by Edouard Machery and Elizabeth O'Neill, 109–27. New York: Routledge.

Brendel, Elke. 2004. "Intuition Pumps and the Proper Use of Thought Experiments." *Dialectica* 58, no. 1: 89–108.

Brickman, Philip, Dan Coates, and Ronnie Janoff-Bulman. 1978. "Lottery Winners and Accident Victims: Is Happiness Relative?" *Journal of Personality and Social Psychology* 36 no. 8: 917–27.

Elster, Jakob. 2011. "How Outlandish Can Imaginary Cases Be?" *Journal of Applied Philosophy* 28, no. 3: 241–58.

Foot, Philippa. 1967. "The Problem of Abortion and the Doctrine of the Double Effect." *Oxford Review* 5: 5–15.

Gendler, Tamar Szabó. 2010. *Intuition, Imagination, and Philosophical Methodology*. Oxford: Oxford University Press.

Gettier, Edmund L. 1963. "Is Justified True Belief Knowledge?" *Analysis* 23, no. 6: 121–23.

Gilbert, Daniel T., and Timothy D. Wilson. 2007. "Prospection: Experiencing the Future." *Science* 317, no. 5843: 1351–54.

Hawke, Peter, and Tom Schoonen. 2020. "Are Gettier Cases Disturbing?" *Philosophical Studies* 178, no. 5: 1503–27.

Kahneman, Daniel, and Gary Klein. 2009. "Conditions for Intuitive Expertise: A Failure to Disagree." *American Psychologist* 64, no. 6: 515–26.

Kamm, F. M. 2012. *The Moral Target: Aiming at Right Conduct in War and Other Conflicts*. New York: Oxford University Press USA.

Klein, Gary. 1998. *Sources of Power: How People Make Decisions*. Cambridge, Mass.: MIT Press.

Lippert-Rasmussen, Kasper. 2008. "Against Self-Ownership: There Are No Fact-Insensitive Ownership Rights over One's Body." *Philosophy and Public Affairs* 36, no. 1: 86–118.

Machery, Edouard. 2011. "Thought Experiments and Philosophical Knowledge." *Metaphilosophy* 42, no. 3: 191–214.

Machery, Edouard. 2017. *Philosophy Within Its Proper Bounds*. Oxford: Oxford University Press.

Maibom, Heidi L. 2016. "Knowing Me, Knowing You: Failure to Forecast and the Empathic Imagination." In *Knowledge Through Imagination*, edited by Amy Kind and Peter Kung, 185–206. Oxford: Oxford University Press.

Markman, Keith D., William P. Klein, and Julie A. Suhr. 2009. *Handbook of Imagination and Mental Simulation*. New York: Psychology Press.

Mayo-Wilson, Conor, and Kevin J. S. Zollman. 2021. "The Computational Philosophy: Simulation as a Core Philosophical Method." *Synthese* 199, nos. 1–2: 3647–73.
McCloskey, Michael. 1983. "Naïve Theories of Motion." In *Mental Models*, edited by Dedre Gentner and Albert L. Stevens, 299–324. New York: Psychology Press.
McCloskey, Michael, Alfonso Caramazza, and Bert Green. 1980. "Curvilinear Motion in the Absence of External Forces: Naïve Beliefs About the Motion of Objects." *Science* 210, no. 4474: 1139–41.
Miscevic, Nenad. 1992. "Mental Models and Thought Experiments." *International Studies in the Philosophy of Science* 6, no. 3: 215–26.
Miscevic, Nenad. 2021. *Thought Experiments*. Cham: Springer.
Nersessian, Nancy J. 1992. "In the Theoretician's Laboratory: Thought Experimenting as Mental Modeling." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1992: 291–301.
Nersessian, Nancy J. 2008. *Creating Scientific Concepts*. Cambridge, Mass.: MIT Press.
Nersessian, Nancy J. 2018. "Cognitive Science, Mental Modeling, and Thought Experiments." In *The Routledge Companion to Thought Experiments*, edited by Michael T. Stuart, Yiftach Fehige, and James R. Brown, 309–26. New York: Routledge.
Nozick, Robert. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.
Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.
Paul, Laurie. 2014. *Transformative Experience*. Oxford: Oxford University Press.
Pesonen, Renne. 2019. "Casual Reasoning: A Social Ecological Look at Human Cognition and Common Sense." Tampere University dissertations no. 96. http://urn.fi/URN:ISBN:978-952-03-1172-8
Pliske, Rebecca M., Michael J. McCloskey, and Gary Klein. 2001. "Decision Skills Training: Facilitating Learning from Experience." In *Linking Expertise and Naturalistic Decision Making*, edited by Eduardo Salas and Gary Klein, 37–53. Mahwah: Lawrence Erlbaum.
Putnam, Hilary. 1975. "The Meaning of 'Meaning.'" *Minnesota Studies in the Philosophy of Science* 7: 131–93.
Ross, Karol G., Jennifer L. Shafer, and Gary Klein. 2006. "Professional Judgments and 'Naturalistic Decision Making.'" In *The Cambridge Handbook of Expertise and Expert Performance*, edited by K. Anders Ericsson, Neil Charness, Paul J. Feltovich, and Robert R. Hoffman, 403–19. New York: Cambridge University Press.
Schindler, Samuel, and Pierre Saint-Germier. 2020. "Are Thought Experiments 'Disturbing'? The Case of Armchair Physics." *Philosophical Studies* 177, no. 9: 2671–95.
Schkade, David A., and Daniel Kahneman. 1998. "Does Living in California Make People Happy? A Focusing Illusion in Judgments of Life Satisfaction." *Psychological Science* 9, no. 5: 340–46.
Shanteau, James. 1992. "Competence in Experts: The Role of Task Characteristics." *Organizational Behavior and Human Decision Processes* 53, no. 2: 252–66.
Stanovich, Keith E. 2003. *How to Think Straight About Psychology*, 7th edition. Boston: Allyn and Bacon.
Tetlock, Philip E. 2005. *Expert Political Judgement: How Good Is It? How Can We Know?* Princeton: Princeton University Press.
Tetlock, Philip E., and Aaron Belkin. 1996. *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*. Princeton: Princeton University Press.
Thomson, Judith J. 1971. "A Defense of Abortion." *Philosophy and Public Affairs* 1, no. 1: 47–66.
Thomson, Judith J. 1985. "The Trolley Problem." *Yale Law Journal* 94, no. 6: 1395–1415.
Trope, Yaacov, and Nira Liberman. 2010. "Construal-Level Theory of Psychological Distance." *Psychological Review* 117, no. 2: 440–63.
Ubel, Peter A., George Loewenstein, Norbert Schwarz, and Dylan Smith. 2005. "Misimagining the Unimaginable: The Disability Paradox and Health Care Decision Making." *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association* 24, no. 4S: S57–S62.
Weinberg, Jonathan M. 2007. "How to Challenge Intuitions Empirically Without Risking Skepticism." *Midwest Studies in Philosophy* 31, no. 1: 318–43.
Weinberg, Jonathan M. 2016. "Going Positive by Going Negative." In *Blackwell Companion to Experimental Philosophy*, edited by Wesley Buckwalter and Justin Sytsma, 71–86. Malden, Mass.: Blackwell.
Weinberg, Jonathan M., Shaun Nichols, and Stephen Stich. 2001. "Normativity and Epistemic Intuitions." *Philosophical Topics* 29, nos. 1–2: 429–60.
Wilkes, Kathleen V. 1988. *Real People: Personal Identity Without Thought Experiments*. Oxford: Oxford University Press.
Williams, Bernard. 1970. "The Self and the Future." *Philosophical Review* 79, no. 2: 161–80.
Williamson, Timothy. 2007. *The Philosophy of Philosophy*. Malden, Mass.: Wiley-Blackwell.
Williamson, Timothy. 2011. "Philosophical Expertise and the Burden of Proof." *Metaphilosophy* 42, no. 3: 215–29.
Williamson, Timothy. 2016a. "Philosophical Criticisms of Experimental Philosophy." In *Blackwell Companion to Experimental Philosophy*, edited by Wesley Buckwalter and Justin Sytsma, 22–36. Malden, Mass.: Blackwell.
Williamson, Timothy. 2016b. "Knowing by Imagining." In *Knowledge Through Imagination*, edited by Amy Kind and Peter Kung, 113–23. Oxford: Oxford University Press.
Wilson, Timothy D., and Daniel T. Gilbert. 2003. "Affective Forecasting." In *Advances in Experimental Social Psychology*, volume 35, edited by Mark P. Zanna, 345–411. San Diego: Elsevier Academic Press.

Woodzicka, Julie A., and Marianne LaFrance. 2001. "Real Versus Imagined Gender Harassment." *Journal of Social Issues* 57, no. 1: 15–30.
Zarnitsyn, Aleks. 2015. "The Cognitive Value of Fiction in Thought Experiments in Personal Identity." *Journal of Aesthetic Education* 49, no. 2: 62–81.

AUTHOR BIOGRAPHY

Henri Tuohimaa is a doctoral researcher at Tampere University, Finland. Their main interests include philosophical methodology, metaphilosophy, and research on thinking and reasoning.

How to cite this article: Tuohimaa, Henri. 2025. "Danger Signals for Untrustworthy Thought Experiments." *Metaphilosophy* 56(2): 209–224. https://doi.org/10.1111/meta.12724.