Croatian Journal of Philosophy
Vol. XXV, No. 73, 2025
https://doi.org/10.52685/cjp.25.73.3
Received: May 10, 2023
Accepted: October 14, 2024

Thought Experiments, Fictions, and Irrelevant Details

BOJAN BORSTNER and TADEJ TODOROVIĆ
University of Maribor, Maribor, Slovenia

The article explores the problem of the cognitive value of thought experiments (TEs) and fictions. Specifically, it deals with the claim that fictions have cognitive value in virtue of being (elaborate) thought experiments. First, a short overview of the cognitive value of TEs is presented, followed by the recent findings from experimental philosophy, which cast doubt on the value of TEs. This is followed by an examination and rejection of the claim that fictions are TEs (as presented by Elgin) for two reasons. First, the analogy between scientific and thought experiments and fictions ultimately fails, as fictions contain the very variables that must be absent for performing successful scientific and thought experiments; second, because of this and based on the research in experimental philosophy, fictions should bias the reader to a greater degree than TE—this is shown to be collaborated by text comprehension research. This claim is further substantiated by analysing two examples of fictions, Le Guin's The Matter of Seggri and her satirical piece A Modest Proposal: Vegempathy. Finally, a more modest claim is considered, namely that fictions contain TEs, which must be properly extrapolated and analysed, yet this leads to issues that are similar to the value of TEs debate. The article thus concludes that using TEs is not advisable for securing the cognitive value of fiction.

Keywords: Thought experiments; fiction; experimental philosophy; cognitive value; text comprehension.

1. *Introduction*

We read fiction because we believe this is of some value to us. The values differ from person to person: some read fiction because it is a fun activity for them, some do it for aesthetic value, and others for altogether different reasons. However, some, if not most, people also believe that fiction provides an additional cognitive value. There are several

kinds of cognitive values in the philosophical discussion on the value of fiction: some argue that fiction can provide propositional knowledge, moral knowledge, conceptual knowledge, or psychological truths. Some cognitivists also argue that fiction could provide practical knowledge or even phenomenal knowledge (Kroon and Voltolini 2024). Nünning for example argues that reading fiction could "broaden our emotional horizon" (Nünning 2018: 49) by expanding our emotional repertoire. Yet there are also several concerns when it comes to the cognitive value of fiction: e.g., fiction could persuade people to believe in nonsense or change their beliefs (Nünning 2015: 43).

This article will focus on a very narrow kind of cognitive value, namely philosophical knowledge similar to knowledge gained by engaging in thought experiments (TEs). Two issues immediately arise: what kind of cognitive value (if any) do TEs provide and how do we reliably learn something new from something that is, by definition, fictional? The reason why we will focus on this specific aspect of cognitive value is that some authors argue that literary fiction has (cognitive) value just because of the similarities between fiction and TEs—either by claiming that (some) works of fiction are TEs (Elgin 2014) or by arguing for a more modest claim that fiction can at the very least be read as TEs (Sorensen 1999: 223). It is thus problematic for the cognitive value of TEs that they are fictional, yet the fact that fictions are TEs is, on the other hand, used to secure the cognitive value of works of fiction. The issue is compounded by the fact that there remains some controversy on the actual value of TEs in general: Klampfer for example argues that because of the shortcoming of TEs, we should use "more sound alternatives to thought-experimentation in moral and political philosophy" (Klampfer 2017: 346). This fascinating interplay between literary fictions and TEs is elegantly summarised by Davies: "[…] rather paradoxically, that TE's are fictions has been taken (by some) to call into question the very thing that is supposed to be established (for others) by the fact that fictions are TE's!" (Davies 2010: 53).

In this article, we would like to argue that justifying the cognitive value of fictions by claiming that fictions are literally TEs is not a good strategy. We do this by presenting two arguments against this claim, namely that the analogy between TEs and fictions is unsuccessful, and that, considering the recent literature in experimental philosophy on TEs and research in text comprehension, it seems plausible that problems for TEs are exacerbated for fictions. We begin by offering a short overview of the cognitive value of TEs and the recent findings on TEs from experimental philosophy, focusing on morally irrelevant factors in TEs that nonetheless impact our intuitions in TEs. We continue by analysing and critically examining the argument in favour of understanding fictions as TEs, ultimately arguing that the argument fails. Based on this argument, the research in experimental philosophy, and the research in text comprehension, we further argue that the reasons

why TEs are problematic in experimental philosophy are much more prominent in fictions, which is why understanding literary fictions as TEs is not advisable. We demonstrate this with two examples of literary fictions, Le Guin's novelette *The Matter of Seggri* and her satirical piece *A Modest Proposal: Vegempathy*, highlighting how the additional factors in fiction exacerbate the problem of biasing the reader.

In the final section, we argue that a more modest claim, namely that fictions contain TEs or that, at the very least, philosophers can extrapolate TEs from fictions, could be sufficient for securing some sort of cognitive value for fictions. We show this by reconstructing the TEs from the analysed works, demonstrating that, in terms of thought experimentation, such extrapolation offers a better starting point for the cognitive value of fictions by eliminating potential confounding factors. Despite this possibility, we remain pessimistic about the idea that we could justify this kind of cognitive value of fiction using TEs.

2. *Cognitive value of TEs*

As is often the case in philosophical problems, there seem to be two 'extreme' positions or camps for a particular problem, with other positions being placed somewhere in between. The same seems to hold true for the cognitive value of TEs. On the one side we have Norton's reductionist stance, according to which fictional scenarios in TEs are merely ornamental, important only in a heuristic or illustrative sense. TEs are "merely picturesque arguments and in no way remarkable epistemologically" (Norton 1996: 334). For Norton, TEs can be reduced, becoming nothing more than arguments "disguised in a vivid pictorial or narrative form" (Norton 2004: 45). The other side is represented by Brown (Brown 1992, 2011), according to whom TEs, and especially a subspecies of them, "platonic TEs," are an autonomous source of knowledge, with the help of which we can "grasp an abstract pattern" through "intellectual perception;" TEs are "telescopes into the abstract realm" (Brown 2004: 1131). The middle ground is represented by proponents of mental modelling (Gendler 2004; Miščević 1992, 2022; Nersessian 1993), who argue that TEs draw upon tacit cognitive resources and build mental models that enable the production of new data via the manipulation of old data. TEs, as mental models, manipulate our cognitive resources in such a way that paraphrasing them as arguments would result in an epistemic loss.

At the very least then, TEs have cognitive value as arguments, and at the most, they are "telescopes into the abstract realm." For the purposes of analysing TEs in fiction, we will presuppose the weakest claim, i.e., TEs are at the very least valuable as arguments (in line with Norton) but will remain open to stronger claims (like mental modelling). If TEs amount to anything more than arguments (mental models or platonic TEs), so much better for TEs, although we do remain sceptical of

this (because of the recent work in experimental philosophy described in the next section). Naturally, the position taken here affects the cognitive value of fictions (as TEs): if TEs have cognitive value as arguments and fictions are TEs, then fictions are valuable as arguments; if TEs have cognitive value as mental models and fictions are TEs, then fictions are valuable as mental models. However, as we will be arguing against the claim that fictions are TEs, this is not that important.

3. *TEs in Experimental Philosophy*

Recent years have shown an upward trend in research doubting the intuitions generated by philosophical TEs (especially in ethics). For example, Uhlmann and colleagues found that in moral scenarios where people are sacrificed for the greater good, the ethnicity and nationality of the sacrificed persons play a role (Uhlmann et al. 2009). Gino and colleagues find that people judge behaviour as more unethical in cases where the victims are identifiable than in cases where they are not (Gino et al. 2010), Greene argues that in the trolley case, different intuitions are triggered by factors like personal force, i.e., when we have to push the person off the bridge, and intention to kill, i.e., killing the person by pushing them off the bridge instead of switching the lever and them dying as a side-effect (Greene et al. 2009), and there also seems to be a lot of evidence that TEs are vulnerable to framing effects (Sinnott-Armstrong 2008) and order effects (Schwitzgebel and Cushman 2012). As Königs writes,

> […] people's case-specific moral intuitions are sensitive to factors that lack intrinsic moral significance. We respond differently to moral scenarios due to the presence of (what seem to be) morally irrelevant factors, such as personal force, distance, ethnicity or nationality. (Königs 2020: 2606)

A further problem is that this does not seem to be the case only for laypeople, but also for professional philosophers. Even philosophers with relevant expertise familiar with moral dilemmas are not immune to framing effects (Schwitzgebel and Cushman 2015). If we take this research at face value, i.e., that morally *irrelevant* factors affect our moral intuitions (specifically in ethical TEs), and that even professional philosophers fall prey to such factors, then the influence of such (morally irrelevant) factors casts serious doubts on the cognitive values of judgements produced in this way. Two responses are possible: either we give up on the project of producing intuitions in this manner and get rid of thought experimentation in general or, being epistemologically informed by the results from experimental philosophy, we either minimize the factors that might affect our intuitions or address them in a different way.

The first option seems quite radical: TEs seem to be almost indispensable as a philosophical tool, and thus, they, at the very least, deserve the benefit of the doubt. The second option seems more promising—somehow, we must address the issue. A reasonable response is to

minimize the confounding factors and hope that this solves the problem. However, the results are also relevant for a related discussion of TEs and fictions—whereas we might be able to minimize confounding factors in TEs, this is obviously not possible for fictions. Moreover, if morally irrelevant factors bias TEs, then shouldn't they bias fictions to a much greater degree, thus making the claim that fictions are TEs that much less plausible? Such experimental results should be applied to this debate; at least prima facie, such problems are compounded in fictions—for every seemingly morally irrelevant factor in a philosophical TE, there are probably orders of magnitudes more in literary TEs. This can serve as an independent argument against understanding fictions as TEs. Before returning to this, however, a positive case for fictions as TEs must be presented.

TEs and Fiction

It should come as no surprise that TEs have inspired works of fiction; one only has to remember Descartes's Evil Genius (Descartes 1984) or Putnam's Brains in a Vat (Putnam 1981), which are so often used in comparison to *The Matrix* (Wachowski and Wachowski 1999). However, the opposite is also the case: indisputably, works of fiction also inspired TEs. Jackson, in his famous Knowledge Argument (Jackson 1982), first presents the now less popular TE of Fred, who sees two different colours in cases where the rest of us see only one, e.g., Fred sees red_1 and red_2 while we only see red (or only red_1), to make a similar point that he makes with the now almost infamous Mary in a black and white room. In the example of Fred, however, he compares the idea that Fred can see one extra colour to H. G. Wells's "The Country of the Blind," where the protagonist, a sighted person in the land of the blind, never managed to convince the population of the existence of an extra sense (Wells 2007). A conclusion that TEs inspire works of fiction and vice-versa, and that many works of fiction could potentially be used by philosophers in constructing new, ingenious TEs, should thus not be controversial. Nevertheless, *The Matrix* is much more than a humble Brains in a Vat TE, and "The Country of the Blind" surely offers more than just the supposed conclusion that something is wrong with physicalism in the mind-body problem. The main problem is therefore as follows: does the fact that fiction is so much more than a TE, change its cognitive value (provided that TEs have cognitive value in the first place)? Or are fictions just (more) elaborate TEs, retaining the cognitive value of TEs?

Let us explore the argument in favour of the claim that fictions are TEs.[1] Elgin presents an argument comprised of two stages. The first by establishing the cognitive value of TEs by analogy with scientific exper-

[1] We will be using Elgin's argument, as she has explicitly defended the view that fictions are TEs, while others have argued for somewhat weaker versions of the argument (Carroll 2002; Sorensen 1999).

iments, and the second by establishing the cognitive value of fiction by analogy with TEs. Starting with a description of scientific experiments, Elgin (rightly) argues that experiments are not just "a mere matter of bringing nature indoors" (Elgin 2014: 222), but require us to isolate the studied phenomena from the "welter of complexities" and fix any variables that might change the outcome of the experiment in order to eliminate possible confounding factors—this enables us to determine the cause of the observed phenomena with a higher degree of certainty. This is analogous to TEs:

> […] a thought experiment fixes certain parameters (e.g., about the relevant laws of nature and the supposed initial conditions), provides a description of the experimental situation that sets out all and only the features considered relevant, and works out the consequences. (Elgin 2014: 230)

Just like scientific experiments, TEs also require interpretation and allow multiple interpretations, which can change over time, as new evidence or arguments come to light. The interpretation can also change because of the change in our background beliefs and tacit assumptions, which are two additional commonalities that both scientific and thought experiments share. The key commonality between scientific and thought experiments, however, is that both exemplify, i.e., scientific and TEs offer epistemic access to examined phenomena via exemplification, which Elgin defines as "the relation of a sample, example, or other exemplar to whatever it is a sample of" (Elgin 2014: 224);[2] e.g., fabric swatches exemplify available colours or patterns. On the other hand, the biggest difference between scientific experiments and TEs is, of course, the fact that the former are actual, whereas the latter are fictional. Yet this is a problem for TEs in general, whereas we are ultimately interested in the cognitive value of fictions. If it turns out that the fact that TEs are fictional means that they have no cognitive value, then fictions have no value as well.[3]

The further analogy between TEs and fiction seems to be more problematic, though. Elgin understands fiction as elaborate TEs, which would endow fiction with the same cognitive value as TEs:

> If an austere thought experiment can afford epistemic access to a range of properties and can do so in a context that is not tightly beholden to a particular theory, there seems to be no reason to deny that a more extensive thought experiment can do the same. […] Like an experiment, a work of fiction selects and isolates, contriving situations and manipulating circumstances so that patterns and properties stand out. It may frame or isolate mundane features of experience so that their significance is evident. (Elgin 2014: 232)

However, we believe the analogy seems to break down at this point. Remember that one of the key similarities between scientific experiments and TEs is that they both exemplify by isolating the studied

[2] See also Elgin (1999) and Goodman (1968).

[3] Nevertheless, even if this is the case, it would be hard to argue against the very weak claim that TEs have at least some cognitive values as arguments.

phenomena from the "welter of complexities," thus removing possible confounding factors. Yet comparing TEs and fiction, that does not seem to be the case. Fiction is much more than an "austere thought experiment;" it is rich and full of emotional language, details, and symbolism that TEs obviously lack. In fact, if scientific experiments are isolating aspects of nature, then, by analogy, it would seem more apt to say that TEs isolate aspects of fiction—the argumentative parts (or the mental modelling parts). The emotional, symbolic, and other fictional aspects are thus, by analogy, the confounding factors that influence the outcome of the experiments, which seems to imply that they should be removed for epistemic purposes. So, while a skilled reader (and especially a philosopher) might be able to extract TEs from fiction or recognize a potential TE in the making, there is an important difference between the two; fiction is by no means austere, but rich and vibrant, evoking emotions, filled with symbolism, details, subplots, etc. This seems to be a crucial difference because in comparing natural phenomena and phenomena of scientific experiments, it is the austerity of the conditions of scientific experiments that enables us to recognize the proper causes of the studied phenomena.

Egan (2016) argues in a similar fashion. He grants that fiction can be used as a source or inspiration for TEs—*applicability claim*, but denies that the analogies between fiction and TEs are strong enough to justify the cognitive value of fiction by using arguments for the cognitive value of TEs—*cognitivism claim*, and he further (and consequently) denies that literary fictions are TEs—*identity claim* (Egan 2016). He also argues that there is an important difference between allegorical and literary reading:

> Thought-experimental readings, then, are naive allegories—allegories whose every concrete element has an allegorical analogue at the abstract level—that contribute to an argument [whereas] literary reading—the sort of reading that seeks to maximize aesthetic pleasure—draws meaning from the connections between elements at the concrete level rather than finding meaning only at the allegorical level. These concrete particularities, then, cannot be straightforwardly reduced to abstract ideas. (Egan 2016: 44)

As mentioned, we do not argue that an astute reader is not capable of extracting a TE from fiction or recognizing a TE in fiction, just like an excellent scientist, due to her extensive knowledge or insight, might be able to recognize or isolate a phenomenon or a relevant cause in nature before repeating the experiment in the laboratory. Yet for it to count as "real" science, it must be repeated in such a setting; similarly, while good philosophers could recognize potential TEs in fictions, they should nevertheless test them in a TE setting before determining its cognitive value. Perhaps the additional fictional values confounded the integrated TE, just like variables in nature confound the observed phenomenon of the scientist. Considering the research on morally irrelevant factors influencing our intuitions in TEs, we should proceed with caution, because such experimental data seems to be even more relevant for the

value of fiction. In the following section, we would thus like to show, on two examples of fiction and some text comprehension research, how the problem of additional fictional elements makes matters much worse for the cognitive value of fictions as TEs.

4. *A Tale of Two TEs in Le Guin's Work*

In this section, we will focus on two works by Le Guin, a novelette *The Matter of Seggri* (Le Guin 2016) and a short satirical work *A Modest Proposal: Vegempathy* (Le Guin and Fowler 2017) which we believe serve to illustrate our perspective—the first being an example of a TE reaching the "correct" conclusion, and the second arriving at the "wrong" conclusion.

The Matter of Seggri

The Matter of Seggri is a novelette containing reports and memoirs from different people (human-like aliens and human-like planet residents) and their experiences on the planet Se-ri, all related to the people of Seggri. The Seggri is a women-dominated society, created by advanced (human-like) aliens, the Hainish, who, many years ago, colonised various planets in the galaxy, occasionally genetically modifying the worlds (it is implied that this was done for experimental purposes). Ultimately, the Hainish civilisation collapsed, and the colonised worlds forgot about their ancestors. Nevertheless, in the future, the Ekumen, a coalition of advanced planets, starts to explore the galaxy, encountering the genetically modified worlds that have developed in isolation, but with different starting conditions. The Seggri differ from normal human population in that there are sixteen women for every man, and the story explores how such a society would function and organize itself:

> There are sixteen adult women for every adult man. One conception in six or so is male, but a lot of nonviable male fetuses and defective male births bring it down to one in sixteen by puberty. My ancestors must have really had fun playing with these people's chromosomes. I feel guilty, even if it was a million years ago. I have to learn to do without shame but had better not forget the one good use of guilt. (Le Guin 2002: 29)

The setup of the story sounds almost exactly like a TE: imagine a world, populated by super powerful aliens, where there are sixteen women born for every man—how would such a society be organised? How would it be different? In Le Guin's world, this produced an almost segregated society, where women lived in villages, married to (sometimes multiple) women, and men lived in castles, practicing competitive sports and martial arts every day. The interaction between men and women took place at monthly games, where men competed against men from other castles for prestige and privilege, and women used this as a source of entertainment. The winners had the privilege of going

to "pleasure houses," where they served as prostitutes—the women picked and paid for the man they wanted, the champions, either just for sex or for conceiving a child. On the other hand, men were not particularly smart. Education, technology, and knowledge in general was the domain of women; it was a society where "men have all the privilege and the women have all the power" (Le Guin 2002: 31). However, what is especially fascinating about the story is how Le Guin uses this setup to illustrate the justification for gender-based injustices on this planet and the gender-based injustices in our society. Because women are the ones that hold all the power, they use the same kinds of rationalisations that were (or are) used by men in our society to argue that women are not suited for education (or positions of power):

> They [men] aren't allowed into the colleges to gain any kind of freedom of mind. I asked Skodr why an intelligent man couldn't at least come study in the college, and she told me that learning was very bad for men: it weakens a man's sense of honor, makes his muscles flabby, and leaves him impotent. "'What goes to the brain takes from the testicles,'" she said. "Men have to be sheltered from education for their own good." (Le Guin 2002: 32)

The purpose of the story is thus (among other things) to illustrate how effortlessly we create rationalisations that serve our narratives for preserving the status quo, and to show how absurd the same kind of reasoning (e.g., women should be sheltered from education for their own good) sounds when the situation is reversed, i.e., when men are the ones with no access to education, yet we have the appropriate background knowledge that education is not detrimental to men (from the actual world). The point is further illustrated later on in the story, when, after the revolution, men are allowed to leave the castles, study, and work. Here is an excerpt from a memoir of a man that escaped the planet and was educated by the Ekumen:

> My sister Pado broached the possibility of an apprenticeship in the clayworks, and I leaped at the chance; but the managers of the Pottery, after long discussion, were unable to agree to accept men as employees. Their hormones would make male workers unreliable, and female workers would be uncomfortable, and so on. The holonews was full of such proposals and discussions, of course, and orations about the unforeseen consequences of the Open Gate Law, the proper place of men, male capacities and limitations, gender as destiny. Feeling against the Open Gate policy ran very strong, and it seemed that every time I watched the holo there was a woman talking grimly about the inherent violence and irresponsibility of the male, his biological unfitness to participate in social and political decision-making. (Le Guin 2002: 61–62)

Again, Le Guin uses the very same rationalisations that were used against women throughout actual history: hormones make female workers unreliable, male workers would be uncomfortable, the workplace is not the proper place for women, women have limitations, etc. Combined with a story about a man that was abused and raped, yet eventually managed to escape such a world, the reader is easily con-

vinced of the wrongness of the system and justifications supporting it, and, with a smidge of self-reflection, can recognize the same injustices and fallacies used to justify them in the actual world. However, the additional value of the story does not seem to be relevant in an argumentative sense; empathising with the protagonist makes us more prone to judge in favour of him, but this effect is rhetorical, not argumentative. In fact, if we would construct the story as a TE and omit the emotional aspects, the conclusion would still stand, but would be rhetorically less effective. It is hard to imagine how additional elements that the narrative contains would affect our *intellectual* processing in anything but a negative (i.e., biased) way.[4] Considering the fact that factors like spatial distance in the drowning child TE (Musen & Greene, n.d.), personal force and intention in trolley cases (Greene et al. 2009), ethnicity and nationality (Uhlmann et al. 2009) and identifiability of victims (Gino et al. 2010) affect decision making in relatively austere TEs,[5] what chance do we have to produce reliable intuitions in much richer and more complex fictions, which is rife with just such factors (nationality, ethnicity, identifiable victims, motivations/intentions etc.). Considering this, it seems reasonable to predict that fictions would bias the reader to a much greater degree than TEs. And that seems to be exactly the case.

Even though there are not many studies on testing TE intuitions produced by fictions, there are plenty of studies concerning how critically readers scrutinize the presented information in fictions. And the results are less than promising:

> […] the evidence indicates that for some kinds of information, readers are at least as likely, if not more likely, to believe what they read in fiction than in non-fiction, because they fail to scrutinize the information. (Friend 2014: 227)

For example, in some studies, participants did significantly better or worse on exams based on the peripheral true of false statements in fictional stories compared to the control group (they agreed with claims that were consistent with the fictional stories and disagreed with claims that were not) (Marsh 2003; Marsh and Fazio 2006).[6] In a different study (Butler et al. 2012) participants did worse (compared to the control group) on questions regarding general knowledge if stories contained false information and vice versa (better if they contained true information). A particularly alarming finding comes from Prentice and Bailis (1995, reported in Prentice and Gerrig 1999), where two groups read the same story; however, one group was told the story was fictional, the other that it was not. The fiction group was significantly more persuaded by false statements, agreeing with statements like "Mental

[4] This kind of approach might nevertheless be useful for convincing people with severe cognitive dissonance, but again, the value seems to be purely rhetorical (or pragmatic).

[5] Not to mention experiments that suggest that even cleanliness affects the generation of intuitions, even in professional philosophers (Tobia et al. 2013).

[6] See Friend (2014) for a comprehensive overview of the literature.

illness is contagious" if such a statement was contained in the story. On the other hand, the fact group was not as persuaded by the same story, rejecting the false information!

The conclusions of the studies appear so strong that psychologists studying text comprehension speculate that the more immersive a narrative is, the more likely it is that the reader will be influenced by it. As Friend summarises the empirical findings: "fiction presents hostile conditions for acquiring empirical knowledge; but rather than increase our scrutiny, we may even reduce it, and this makes it more likely that we will accept what we read whether or not it is true" (Friend 2014: 237).

So not only would it be unadvisable to claim that fictions have cognitive value in virtue of being TEs, but this might also be straight up dangerous. Considering the research on biased judgments in TEs and the worrying conclusions from studies on text comprehension, we should be instead doubly wary of anything we might learn from fiction. Even though examples like Le Guin's *The Matter of Seggri* arrive at the correct conclusion, the steps in arriving at the conclusion cannot be philosophically justified. This is especially important considering that even great writers (like Le Guin) sometimes just get it completely wrong. Consider the next example.

The Bad – A Modest Proposal: Vegempathy

A Modest Proposal: Vegempathy, is a satirical piece by Le Guin (Le Guin and Fowler 2017), where she seems to be arguing that vegetarianism and veganism are absurd positions. We choose this example not to attack Le Guin, but to highlight that even great, insightful writers have severe blind spots and biases when it comes to defending the status quo (e.g., Aristotle and slavery). Le Guin's satirical proposal is that we should no longer be omnivores, vegetarians, or vegans, and should instead adopt oganism—"ingesting only the unsullied purity of the O [oxygen]" (Le Guin and Fowler 2017: 130). The argument presented in favour of such a view is a familiar one, namely that plants have feelings, or a weaker claim that we do not know that plants do not have feelings. Suffering of living beings is thus unavoidable, so it makes no sense to prefer killing plants over animals, thus the vegan position is pointless, as they are just as hypocritical as omnivores. This can be presented as a TE: imagine that it turns out that plants are sentient (and thus, for the sake of the argument, of equal moral worth as animals)—if that is the case, is it morally permissible to continue eating animals? If we take this position seriously, believing that the life of an animal is equal to the life of a plant, then because of the second law of thermodynamics and general energy loss,[7] it would of course still make

[7] The research of this seems to be clear now—we require vastly more resources and land for raising animals than plants, e.g., one of the biggest analysis of data on the environmental impact of nearly 40,000 farms shows that meat and dairy use

more sense to eat the plants directly, as we would have to kill much more plants to feed the animals (which we would also have to kill). Unless the argument is that we should inflict the maximum amount of death and suffering (i.e., kill as many plants and animals as possible), then the question of plant sentience is trivial—if we care about causing the least amount of harm to sentient beings, we should still be vegan regardless of whether plants are sentient or not.

Yet the described TE (simply imagine that plants are sentient and apply the situation to the ethical question of killing animals for food) differs from Le Guin's satire. Here's an excerpt:

> Consider, for one moment, what plants undergo at our hands. We breed them with ruthless selectivity, harass, torment, and poison them, crowd them into vast monocultures, caring for their well-being only as it affects our desires, raising many merely for their byproducts such as seed, flower, or fruit. And we slaughter them without a thought of their suffering when "harvested," uprooted, torn living from their earth or branch, slashed, chopped, mown, ripped to pieces—or when "cooked," dropped to die in boiling water or oil or an oven—or, worst of all, eaten raw, stuffed into a human mouth and masticated by human teeth and swallowed, often while alive. (Le Guin and Fowler 2017: 128–129)

We can immediately notice emotional language, such as tormenting, breeding, and slaughtering the plants,[8] followed by the final graphic description of plants getting "masticated by human teeth and swallowed, often while alive." These kinds of emotions influence our interpretation of the TE (or an argument in general), and we know this because this is exactly the kind of language that some argue vegan activists should not be using in their advocacy for animals, e.g., tormenting, slaughtering, murdering, raping, etc. Considering the empirical findings regarding TEs in general, such as framing effects and order effects even on *professional philosophers* familiar with the arguments and TEs (Schwitzgebel and Cushman 2015), and all the above mentioned research on text comprehension, we simply cannot ignore the much higher prevalence of such language and other emotionally charged content (e.g., empathising with the protagonist, or with plants in this case) in cases of fiction. This is especially evident in the discussed example because it is a relatively short work (about 3 pages), with practically no (human) protagonists, yet it still manages to easily bias the reader in a way that would be "illegal" in a philosophical setting. Remember that one of the key steps for TE, according to proponents of mental modelling, is that the "contemplation of the scenario takes place with a specific purpose: the

83% of farmland while only providing 18% of calories and 37% of protein—moving to a vegan diet would enable us to reduce more than 75% of farmland (area equivalent to the size of USA, China, EU, and Australia) and also reduce arable land by 19% (Poore and Nemecek 2018).

[8] Ironically, when Le Guin describes cows, this kind of language is absent, even though it would be perfectly applicable to the dairy industry: "We can't ask the cow's opinion on being milked, although we can hypothesize that if her udder was full she might feel relief" (Le Guin and Fowler 2017: 130).

confirmation or disconfirmation of some hypothesis or theory" (Gendler 2004: 1155). TEs are supposed to create an environment where we are to come to a conclusion via critically examining and reflecting on a particular scenario, reaching reflective equilibrium, whereas fictions, empirically speaking, seem to create an environment hostile to critical thinking.

So *even if* we were to understand fiction as actual TEs, then we would almost definitely have to admit that because of the inherent emotional content of fictional TEs, they bias the reader to the degree that the ascription of reliable cognitive value of such TEs is impossible, not to mention the different "mode" of reading fiction and TEs—as the literature suggests, reading fiction seems to create hostile conditions for generating knowledge, whereas reading and thinking about TEs is supposed to stimulate (critical) contemplation of a scenario with the purpose of confirming or disconfirming a theory. Overall, whatever the shortcomings of TEs we mention, we must be certain that fictional TEs will suffer the same shortcomings, but to a much greater degree, and then some. We believe that Currie rightly states that:

> […] the epistemically exemplary thought experiments we find in science and philosophy have certain features on which their reliability depends, and those features are generally lacking or much attenuated in the kinds of fictions this book is concerned with. What I am questioning is whether the fictions […] have even the modest reliability we can attribute to thought experiments in the sciences and in philosophy. (Currie 2020: 138)

So much for the strong claim that fictions are TEs. Instead, a weaker claim, namely that there are TEs in fiction, but only after we properly extrapolate them, should be adopted.

5. *Extrapolating TEs from fictions*

We believe that Elgin's (Elgin 2014) original analogy, that between nature and scientific experiments, is, in a modified form, a perfect fit for the connection between fictions and TEs. Namely, scientific experiments are valuable because they exemplify, and they exemplify by isolating the studied phenomena, thus removing as many confounding factors as possible, something that is not possible in nature. And just like a scientific experiment is an aspect of nature brought indoors, a TE could be understood as an aspect of fiction. The analogy is much stronger, as fictions, just like nature, contain a plurality of confounding factors that influence/bias the (scientific and thought) experiment. The only difference that remains is that scientific experiments and nature reside in the realm of the actual, whereas fictions and TEs reside in the realm of the possible. This can still accommodate some of Elgin's claims. For example, when she argues that *Oedipus Rex* can be read as TE in favour of Aristotle's hypothesis that we should not call any man happy until they are dead (Elgin 2019), she elaborates the claim by (in a sense) extrapolating such a TE from the play. To understand

the point (in the context of Aristotle's hypothesis) one does not have to read the actual play; in fact, many would overlook the TE in the play because of all other fictional elements. There is nothing to lose and a lot to gain by extrapolation—it is not only useful in a pragmatic sense (reading the TE instead of the entire novel or play), but also offers, due to philosophical formulation and "guidelines" concerning TEs, more credibility from charges of biasing the reader and thus at the very least addresses the bias effect that fiction appears to have on us.

The examples that we discussed, Le Guin's *The Matter of Seggri* and *A Modest Proposal: Vegempathy*, both contain elements that are, at the very least, problematic for TEs, but by extrapolation, this is easily rectified. For *The Matter of Seggri*, we can simply claim that it can be read as the following TE:[9]

> Imagine a possible world where, due to genetic factors, there are much more women than men, which leads to a society where women have all the power and scientific knowledge, whereas men serve only for procreation and entertainment for women, usually in the form of playing games (Gladiator style). When we visit this world, we ask the women why they believe men should not have access to education, work, and positions of power. They reply that this is not the proper place for men, that the male hormonal profile is not suited for such work, that they would distract the women already working and studying, etc. Such justification is obviously erroneous, as we know from experiences from the actual world. So why does it seem (or hopefully only was) socially acceptable to use the same kind of reasoning for arguing against education, work, and positions of power for women?

This removes the unnecessary details pertinent to the thought experimentation and avoids additional fictional elements that might bias the reader (e.g., the protagonist in the story was raped, he fell in love, was betrayed by the people he loved, etc.—all factors that distract from the TE at play). For *A Modest Proposal: Vegempathy*, the TE is even simpler, as it is already present in some forms in public discourse regarding animal ethics:

> Imagine that plants feel pain and that they are as sentient as animals. If that were the case, would it be futile to stop killing animals for dietary purposes, because we would still be killing plants?

This removes the emotional imagery presented by Le Guin, while still retaining the main point that she was trying to make. Accordingly, the reader is much less likely to be biased by such emotional pleas in forming their judgment.

Such extrapolation could avoid the pitfalls of the worrying research on text comprehension (albeit the worries regarding TEs, unfortunately, remain), yet still offer some cognitive value for fiction, but only after the claims have been suitably processed and analysed. However, formulating such a view might encounter issues that are practically the same as the more general issues with TEs. How should the extrapola-

[9] This also does not exclude the possibility that we can extrapolate distinct and even mutually incompatible TEs from the same work of fiction.

tion work, what process should we use? Which details should we deem as relevant and which as irrelevant? As highlighted by an anonymous reviewer:

> If the reader can determine which nonmoral factors should play a role in the process of extrapolation and which should be disregarded, then the reader no longer needs to do the extrapolation to get the point of the story. In other words, knowing how to properly extrapolate the morally relevant aspects of the story shows that the reader already has moral knowledge (that the extrapolated TE should provide)—so why bother?

Of course, one could answer that perhaps philosophers are especially equipped for such a task, but this is exactly the same problematic claim that arises in the debate on TEs in general! Miščević for example argues that TEs work as mental models with specific stages, the first of which is the construction of the TE itself (Miščević 2017). An important part of the construction of the TE is to decide which aspects of the situation are relevant and which are not:

> Importance and coverage seem to allow for trade-offs: if the centrally important variables are correctly represented in TE, the construction can survive without extensive coverage of all details. On the other hand, detailed coverage guarantees that all central variables will be taken into account. (Miščević 2013: 521)

But this is problematic, as explained by e.g. Gartner. Think of the following analogy with cookies: the task is to write a successful recipe, and we have two options. We can either write the generally important steps without the minute details or we can write a very specific and exact recipe. Deciding which ingredients are the centrally important ones (i.e. the relevant features of TEs or the extrapolated TE in our instance) will vary from person to person, which is problematic. On the other hand, a very detailed recipe will appear extremely guided, like the above instances of Le Guin's works—both stories guide the reader to a very specific conclusion, and there is little room for other interpretations. As Gartner concludes (for ethical TEs, ETEs for short):

> […] without details, ETEs are not useful, because every added feature could change the judgement or change the relevance of existing features […] and, consequently, change the judgement about the case, or it would be so purely constructed that it would be very far away from the actual world; and (ii) with all of the details that the thought experimenter could imagine, the purpose of the ETE would be that an agent (the reader) would confirm the constructor's claims and not test it. (Gartner 2017: 159)

If we transfer the analogy to fictions and extrapolating TEs from fictions, the problem just seems to be exacerbated—the philosopher (or just a general astute reader) will have already decided on what the result of their extrapolated TE should be. In other words, the TE that they would detect would be the one confirming their prior beliefs, which would make such extrapolation vulnerable to confirmation bias, not to mention the same exact problems that were listed in the third part of this paper.

Therefore, both options are problematic: if works of fiction are just elaborate TEs, the problems of TEs are multiplied for works of fiction, so the argumentation that the cognitive value of fiction is secured via TEs is not advisable; and if we adopt a very modest claim that we can extrapolate TEs from works of fiction, we run into issues that plague TEs in general: which aspects are relevant and which are not and who should decide? Overall, the prospects for using TEs as a mechanism for securing the cognitive value of fiction are rather grim: we believe a better alternative is to focus on other virtues and other kinds of cognitive values, e.g. affective value (Nünning 2018), and leave TEs aside, at least until some central issues regarding the value of TEs are resolved, and reevaluate the idea at that time.

References

Brown, J. R. 1992. "Why Empiricism Won't Work." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1992*, 271–279.

Brown, J. R. 2004. "Peeking into Plato's Heaven." *Philosophy of Science* 71 (5): 1126–1138. https://doi.org/10.1086/425940

Brown, J. R. 2011. *The laboratory of the mind: Thought experiments in the natural sciences* (2nd ed). Abingdon: Routledge.

Butler, A. C., Dennis, N. A. and Marsh, E. J. 2012. "Inferring facts from fiction: Reading correct and incorrect information affects memory for related information." *Memory* 20 (5): 487–498.

Carroll, N. 2002. "The Wheel of Virtue: Art, Literature, and Moral Knowledge." *The Journal of Aesthetics and Art Criticism* 60 (1): 3–26.

Currie, G. 2020. *Imagining and Knowing: The Shape of Fiction* (1st ed.). Oxford: Oxford University Press.

Davies, D. 2010. "Learning Through Fictional Narratives in Art and Science." In R. Frigg and M. Hunter (eds.). *Beyond Mimesis and Convention: Representation in Art and Science*. Springer Netherlands, 51–69.

Descartes, R. 1984. *The Philosophical Writings of Descartes*. Cambridge: Cambridge University Press.

Egan, D. 2016. "Literature and Thought Experiments." *The Journal of Aesthetics and Art Criticism* 74 (2): 139–150.

Elgin, C. Z. 2014. "Fiction as Thought Experiment." *Perspectives on Science* 22 (2): 221–241.

Elgin, C. Z. 2019. "Imaginative Investigations: Thought Experiments in Science, Philosophy and Literature." In F. Bornmüller, J. Franzen and M. Lessau (eds.), *Literature as Thought Experiment?*. Leiden: Brill, 1–16.

Friend, S. 2014. "Believing in Stories." In G. Currie, M. Kieran, A. Meskin and J. Robson (eds.). *Aesthetics and the Sciences of Mind*. Oxford: Oxford University Press, 227–248.

Gartner, S. 2017. "Did a Particularist Kill the Thought Experiments?" In B. Borstner and S. Gartner (eds.). *Thought Experiments between Nature and Society: A Festschrift for Nenad Miščević*. Cambridge: Cambridge Scholars Publishing, 154–165.

Gendler, T. S. 2004. "Thought Experiments Rethought—And Reperceived." *Philosophy of Science* 71 (5): 1152–1163.

Gino, F., Shu, L. L. and Bazerman, M. H. 2010. "Nameless+harmless=blameless: When seemingly irrelevant factors influence judgment of (un)ethical behavior." *Organizational Behavior and Human Decision Processes* 111 (2): 93–101.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E. and Cohen, J. D. 2009. "Pushing moral buttons: The interaction between personal force and intention in moral judgment." *Cognition* 111 (3): 364–371.

Jackson, F. 1982. "Epiphenomenal Qualia." *The Philosophical Quarterly (1950-)* 32 (127): 127–136.

Klampfer, F. 2017. "The False Promise of Thought-Experimentation." In B. Borstner and S. Gartner, *Thought Experiments between Nature and Society: A Festschrift for Nenad Miščević*. Cambridge: Cambridge Scholars Publishing, 328-348.

Königs, P. 2020. "Experimental ethics, intuitions, and morally irrelevant factors." *Philosophical Studies* 177 (9): 2605–2623.

Kroon, F. and Voltolini, A. 2024. "Fiction." In E. N. Zalta and U. Nodelman (eds.). *The Stanford Encyclopedia of Philosophy* (Summer 2024). Metaphysics Research Lab: Stanford University. https://plato.stanford.edu/archives/sum2024/entries/fiction/

Le Guin, U. K. 2002. *The Birthday of the World and Other Stories*. Perfect-Bound.

Le Guin, U. K. 2016. *The unreal and the real: The selected short stories of Ursula K. Le Guin* (First Saga edition). Saga Press.

Le Guin, U. K. and Fowler, K. J. 2017. *No time to spare: Thinking about what matters*. Houghton Mifflin Harcourt.

Marsh, E. 2003. "Learning facts from fiction." *Journal of Memory and Language* 49 (4): 519–536.

Marsh, E. J. and Fazio, L. K. 2006. "Learning errors from fiction: Difficulties in reducing reliance on fictional stories." *Memory & Cognition* 34 (5): 1140–1149.

Miščević, N. 1992. "Mental models and thought experiments." *International Studies in the Philosophy of Science* 6 (3): 215–226.

Miščević, N. 2013. "In Search of the Reason and the Right—Rousseau's Social Contract as a Thought Experiment." *Acta Analytica* 28 (4): 509–526.

Miščević, N. 2017. "Accounting for Thought Experiments – 25 Years Later." In B. Borstner and S. Gartner (eds.). *Thought Experiments between Nature and Society: A Festschrift for Nenad Miščević*. Cambridge: Cambridge Scholars Publishing, 11–31.

Miščević, N. 2022. *Thought Experiments*. New York: Springer International Publishing.

Musen, J. D. and Greene, J. D. (n.d.). *Mere Spatial Distance Weakens Perceived Moral Obligation to Help Those in Desperate Need. (Unpublished Manuscript)*.

Nersessian, N. 1993. "In the Theoretician's Laboratory: Thought Experimenting as Mental Modeling." In D. Hull and M. Forbes (eds.). *PSA 1992: Proceedings of the Biennial Meeting of the Philosophy of Science Association* (Vol. 2). Chicago: The University of Chicago Press, 291–301.

Norton, J. D. 1996. "Are Thought Experiments Just What You Thought?" *Canadian Journal of Philosophy* 26 (3): 333–366.

Norton, J. D. 2004. "Why Thought Experiments Do Not Transcend Empiricism." In C. Hitchcock (ed.). *Contemporary Debates in Philosophy of Science*. Hoboken: Wiley-Blackwell, 44–66.

Nünning, V. 2015. "Narrative Fiction and Cognition." *Forum for World Literature Studies* 7 (1): 41–61.

Nünning, V. 2018. "The Affective Value of Fiction Presenting and Evoking Emotions." In I. Jandl, S. Knaller, S. Schönfellner and G. Tockner (eds.). *The Affective Value of Fiction Presenting and Evoking Emotions*. Bielefeld: transcript Verlag, 29–54.

Poore, J. and Nemecek, T. 2018. "Reducing food's environmental impacts through producers and consumers." *Science* 360 (6392): 987–992.

Prentice, D. A. and Gerrig, R. J. 1999. "Exploring the boundary between fiction and reality." In S. Chaiken and Y. Trope (eds.). *Dual-process theories in social psychology*. New York: The Guilford Press, 529–546.

Putnam, H. 1981. "Brains in a Vat." In H. Putnam. *Reason, Truth, and History* Cambridge: Cambridge University Press, 1–21.

Schwitzgebel, E. and Cushman, F. 2012. "Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers." *Mind & Language* 27 (2): 135–153.

Schwitzgebel, E. and Cushman, F. 2015. "Philosophers' biased judgments persist despite training, expertise and reflection." *Cognition* 141: 127–137.

Sinnott-Armstrong, W. 2008. "Framing moral intuitions." In W. Sinnott-Armstrong (ed.). *Moral psychology, Vol 2: The cognitive science of morality: Intuition and diversity*. Cambridge: MIT Press, 47–76.

Sorensen, R. A. 1999. *Thought Experiments*. Oxford: Oxford University Press.

Tobia, K., Chapman, G. and Stich, S. 2013. "Cleanliness is next to morality, even for philosophers." *Journal of Consciousness Studies* 20 (11–12): 195–204.

Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D. and Ditto, P. H. 2009. "The motivated use of moral principles." *Judgment and Decision Making* 4: 476–491.

Wachowski, L. and Wachowski, L. (Directors). 1999. *The Matrix*. Warner Bros.

Wells, H. G. 2007. *The country of the blind and other selected stories*, In A. Sawyer and P. Parrinder (eds.). *This selection*. London: Penguin classics.