CROATIAN

JOURNAL

OF PHILOSOPHY

Vol. XXV · No. 73 · 2025

Introduction

The Croatian Journal of Philosophy *launches its new volume with two significant changes. Following the passing of its longtime Editor-in-Chief, Nenad Miščević (1950–2024), a new editorial team is now taking the journal forward. Alongside the new editor-in-chief, Tvrtko Jolić, the editorial team includes Dunja Jutronić (advisory editor), Viktor Ivanković (assistant editor), Nino Kadić (managing editor), Mirela Fuš-Holmedal, Karolina Kudlek, and Andres Moles. We look forward to welcoming additional colleagues in the near future to join us in broadening the journal's scope across all areas of philosophy.*
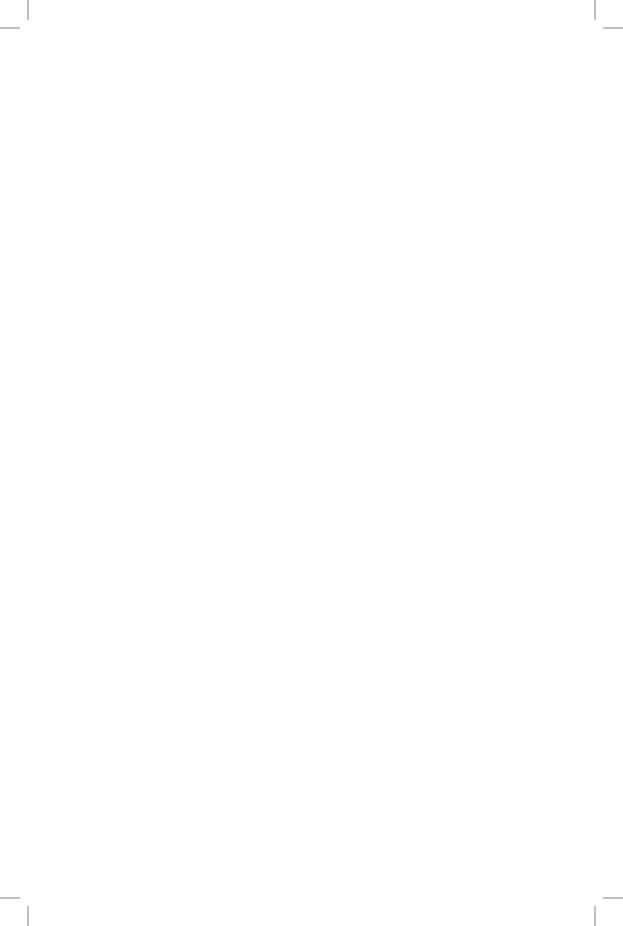
I want to take this opportunity to honour and express gratitude to Professor Miščević for his tireless efforts and unwavering dedication to the journal throughout his long editorial tenure. His dedication to fostering a collaborative environment inspired countless authors to contribute their work, and his relentless promotion of the journal elevated its visibility and academic standing. Collaborating with him was a profound privilege; his guidance was marked by generosity and wisdom, and his leadership ensured a seamless, harmonious editorial process.

I also want to thank former editorial members Stipe Kutleša, Davor Pećnjak, and Joško Žanić for their contributions.

The second significant change is that, after 24 years with the publishing house Kruzak, the journal's publisher is now the Institute of Philosophy in Zagreb. Without the commitment and perseverance of Mr. Kruno Zakarija, owner of Kruzak and one of the journal's founding figures, the Croatian Journal of Philosophy *would not exist today. Zakarija's tenacity and constructive stubbornness in publishing every new issue, despite financial hurdles and bureaucratic challenges, commands immense respect from all of us involved in the journal's quarter-century journey—whether as readers, authors, or editors. Kruno, thank you once again for everything!*

Looking ahead, the journal will remain open to submissions across all areas of contemporary philosophical research. The editorial team, alongside reviewers, will strive to maintain rigorous standards while expediting the peer-review process—a crucial step in an era defined by rapid technological advancement and the accelerated flow of information.

TVRTKO JOLIĆ

Safety and Future Dependence

BIN ZHAO
Institute of Foreign Philosophy, Peking University, Beijing, PR China

According to the safety account of knowledge, one knows that p only if one's belief in p could not easily have been false. In the literature, most objections to the safety account rely on intuition of knowledge that could be easily denied by the safety theorists. In this paper, an objection to the safety account which does not make use of such intuition is raised. It is argued that either there are instances of unsafe knowledge or the safety account has an implausible implication that one's epistemic status might depend on what happens in the future.

Keywords: Future dependence; knowledge; safety.

1. *Introduction*

In many Gettier cases, though the subject holds a true belief, her belief could easily have been false. This motivates the safety account of knowledge, according to which, *S* knows that *p* only if *S*'s belief in *p* is safe, that is, only if *S* could not easily have falsely believed *p*. A belief that satisfies the safety condition counts as knowledge unless it exhibits some non-modal shortcomings that would deprive it of the status of knowledge.

Like other accounts of knowledge in the literature, the safety account is susceptible to a variety of objections. It has been argued that, in some putative counterexamples, one could know some proposition though one could easily have falsely believed the proposition. We thus have instances of unsafe knowledge. However, the safety theorists are ready to deny their opponents' intuition of knowledge. Thus, such a debate often leads to a stalemate.

In this paper, an objection to the safety condition which does not rely on such intuition is raised. It is argued that, if the safety theorists accept the intuition of knowledge, then there are instances of unsafe knowledge. But, if the safety theorists deny the intuition of knowledge, then the safety account has an implausible implication that one's epis-

temic status might depend on what happens in the future. Therefore, no matter whether the safety theorists deny the intuition of knowledge or not, the safety account is problematic.

2. *The Safety Account of Knowledge*

Consider a scenario where one looks at a reliable clock under normal lighting conditions and thus forms a true belief that *it is now 12:00*. Since the clock is reliable and the lighting conditions are normal, the belief counts as knowledge. Consider another scenario where one looks at a clock that stops 12 hours earlier under normal lighting conditions and thus forms a true belief that *it is now 12:00*. The belief is true as a matter of luck and does not count as knowledge. After all, it is true because the dose of bad luck that the clock had stopped is canceled out by the good luck that the clock had stopped 12 hours ago.

 Why is it the case that one knows the correct time in the first scenario but not in the second scenario though in both scenarios one has a true belief about the time? The safety account of knowledge offers a simple answer to this question. According to this account, S knows that p only if S's belief in p is safe, that is, only if S could not easily have falsely believed p. To put it formally,

> *Safety*: S's belief in p, formed on belief-formation method M, is safe, if and only if, in all nearby possible worlds where S forms a belief in p on M, p is true.[1]

This makes us consider whether p is true in nearby possible worlds where S believes that p. If p is false in some of these possible worlds, then S's belief in p is not safe, and S does not know that p. If p is true in all these possible worlds, then S's belief in p is safe, and S knows that

[1] The safety condition is usually relativized to the belief-formation methods to avoid putative counterexamples such as Alfano's (2009) REDWOOD, Goldman's (1976) DACHSHUND, and Nozick's (1981) GRANDMOTHER. Because the argument here does not hinge on whether the conditions should be thus relativized, I shall leave this point aside. For discussions of the individuation of belief-formation methods, see Alfano (2009), Becker (2008, 2012), and Zhao (2022a, 2022d, 2024a, 2024b, 2025, forthcoming). Proponents of the safety condition on knowledge include Ball (2016), Beddor and Pavese (2020), Dutant (2010), Grundmann (2020), Hirvelä (2019), Luper-Foy (1984), Manley (2007), Peet and Pitcovski (2018), Pritchard (2005, 2009), Sainsbury (1997), Sosa (1999a, 1999b, 2015), and Williamson (2000).

 It has also been argued that the safety condition should be globalized to a set of propositions rather than the target proposition to account for why beliefs in necessary truths could still be true as a matter of luck. See Pritchard (2009), Sosa (2015), and Williamson (2000). Here is the globalized version of the safety condition:

> *Globalized Safety*: S's belief that p, formed on belief-formation method M, is safe, if and only if, in all nearby possible worlds where S forms a belief on method M, the belief is true.

Because the argument in this paper does not depend on which version of the safety condition we opt for, I shall not delve into the distinction here. For discussions of the globalized version of the safety condition, see Becker (2006), Hirvelä (2019), and Zhao (2022a, 2022b, 2022c).

p unless it exhibits some non-modal shortcomings that would deprive it of the status of knowledge. One knows the correct time in the first scenario because his belief is safe. In all nearby possible worlds where he believes that *it is now 12:00* via the clock, it is 12:00. One does not know the correct time in the second scenario because his belief is unsafe. There are some nearby possible worlds where he looks at the clock one minute earlier and believes that it *is now 12:00* via the clock, while the time is 11:59.

In a word, the safety condition is supposed to be a necessary condition on knowledge that helps to eliminate epistemic luck. In virtue of this, the safety account of knowledge does an excellent job handling cases where the subject's belief is true as a matter of luck, e.g., the Gettier cases.

3. *Does Knowledge Require Safety?*

In the literature, a variety of objections to the safety account of knowledge have been raised. The objections usually take the following structure: an example is constructed. It is then argued that the subject in the example knows some proposition though she could easily have falsely believed the proposition. We thus have an instance of unsafe knowledge (Baumann 2008; Bogardus 2014; Comesaña 2005; Neta and Rohrbaugh 2004; Yamada 2011). However, the safety theorists could push back and deny their proponents' intuition of knowledge (Pritchard 2009). In that case, the putative counterexamples to the safety account are discharged. Thus, the debate between them often leads to a stalemate.

In this section, I argue that the safety theorists' denial of their proponents' intuition of knowledge is not without some serious costs. To be specific, the safety theorists are faced with a dilemma here: if they choose not to deny the intuition of knowledge, then there are instances of unsafe knowledge. But, if they choose to deny the intuition of knowledge, then the safety account has an implausible implication that one's epistemic status might depend on what happens in the future.

Here is the example I shall consider,

> *Dead President.* Consider . . . the situation of a generally well-informed citizen N. N. who has not yet heard the news from the theatre where Lincoln has just been assassinated. Since Lincoln is dead, he is no longer President, so N. N. no longer knows that Lincoln is President (knowing is factive). However, N. N. is in no position to know that anything is amiss. He continues reasonably to believe that Lincoln is President . . . N. N. does not know that Lincoln is President. (Williamson 2000: 23).

Let t_0, t_1, ... , t_n be a series of times at one millisecond intervals in this case. To be specific, t_0 is the starting point when N. N. begins to believe that Lincoln is President; t_{n-1} is one millisecond before Lincoln dies; t_n is the time Lincoln dies. Let L be the proposition that Lincoln is President; B(L) the condition that N. N. believes that L; and K(L) the

condition that N. N. knows that L. It is reasonable to say that we have K(L) at t_0.

Let us consider the possible world w where Lincoln is assassinated one millisecond earlier at t_{n-1} rather than t_n. In w, N. N. falsely believes that L on the same belief-formation method as that in the actual world, i.e., his memory, at t_{n-1}. Is w a nearby possible world? I think so. There is at least no difference in the belief-formation method between the actual world and w. In both worlds, the beliefs are formed on the same belief-formation method, i.e., his memory. The most salient difference between them is that L has different truth-values: L is true at t_{n-1} in the actual world while false at t_{n-1} in w. However, this difference should not bother us. If a possible world counts as a nearby possible world only if the target proposition has the same truth-value as that in the actual world, then the safety condition would be satisfied for any true belief and thus the condition would be of no use in eliminating luckily true belief from the realm of knowledge. One might appeal to other differences between the actual world and w to exclude w from the realm of nearby possible worlds. However, for any such difference, we can tweak the example such that the actual world and w are the same or, at least, similar in these respects. For instance, if one thinks that Lincoln's being assassinated one millisecond before t_n makes w very different from the actual world where Lincoln dies at t_n, we can reformulate the case such that Lincoln was assassinated one microsecond, one nanosecond, or one picosecond before t_n in w.

To clarify, this is not to say that temporal proximity suffices for modal proximity. In cases where temporal proximity is not accompanied by violations of law or large-scale mismatch of particular facts, temporal proximity leads to modal proximity. For instance, w is a nearby possible world because the only difference between w and the actual world is that John Wilkes Booth assassinated Lincoln slightly more efficiently in w. In contrast, in cases where temporal proximity is accompanied by violations of law or large-scale mismatches of particular facts, temporal proximity does not lead to modal proximity. For instance, there will be a total Solar eclipse on September 3, 2081, at 9:07:31 (Terrestrial time, Central Europe) in the actual world. However, the possible world where the total Solar eclipse happens on September 3, 2081, at 9:07:32 (Terrestrial time, Central Europe) is not a nearby possible world because there are either violations of law or large-scale mismatches of particular facts in that possible world. In sum, we have good reasons to think that w is a nearby possible world.

Anthony Brueckner and M. Oreste Fiocco (2002) think that, in Dead President, we intuitively have K(L) at t_{n-1} in the actual world. The problem for the safety theorists is that there is a nearby possible world, i.e., w, where N. N. falsely believes that L at t_{n-1}. This is a counterexample for the safety account as the subject knows some proposition though he could easily have falsely believed the proposition (Brueckner and Fiocco 2002: 288).

In sum, nothing prevents us from counting w as a nearby possible world. As a consequence, we are forced to give up the safety account. Is there a way out for the safety theorists? The safety theorists may bite the bullet and deny K(L) at t_{n-1} in the actual world. If the only reason for accepting K(L) at t_{n-1} in the actual world is that it seems intuitive, then the safety theorists are ready to deny such intuition as this is what they usually do regarding other putative counterexamples.

This is the stance taken by Williamson. As he writes,

> [safety] need not be determined by local properties of the basis. For instance, if someone continues over some time to believe that Lincoln is President on the basis of automatic updating, without receiving further confirmation, the [safety] of the basis may depend on whether Lincoln is about to be assassinated. (Williamson 2008: 280)

That is to say, N. N.'s belief that L at t_{n-1} is not safe because there is a nearby possible world, i.e., w, where Lincoln is assassinated one millisecond earlier at t_{n-1} such that he falsely believes that L on a similar basis at t_{n-1}. In virtue of its being unsafe, N. N.'s belief that L at t_{n-1} does not count as knowledge. Contrary to the intuition of knowledge from the anti-safety theorists, we do not have K(L) at t_{n-1} in the actual world. Therefore, the putative counterexample for the safety condition is dismissed.

Once again, we seem to have a stalemate here. On the one hand, the anti-safety theorists argue that, because we have K(L) in at t_{n-1}, safety is not a necessary condition for knowledge. On the other hand, the safety theorists argue that, because safety is a necessary condition for knowledge, we do not have K(L) in at t_{n-1}. The safety theorists' argument is merely a modus tollens of the anti-safety theorists' modus ponens. If so, then it seems that Dead President is not in a better place than other putative counterexamples in the literature to resolve the debate. Nonetheless, I don't think this is a dead-end for the discussion because the safety theorists' move is not without some serious cost. To explicate this point, let us examine the anti-safety and the safety theorists' stances in turn.

If the anti-safety theorists' story is true, then N. N. knows that L until t_n. Why doesn't N. N. know that L at t_n? The anti-safety theorists' answer is as follows: because knowledge is factive, one cannot know something if it is not factive. That is to say, L's being false at t_n makes it the case that ~K(L) at t_n.

In contrast, suppose that the safety theorists' story is true, that is, N. N. does not know that L at t_{n-1}. What makes it the case that N. N. does not know that L at t_{n-1} while he knows that L at t_0? The safety theorists' answer is as follows: because knowledge is safe, one cannot know something if one's belief is not safe. His belief at t_0 is safe but it becomes unsafe at some point before t_{n-1}.

Here comes another question: why is his belief that L at t_{n-1} unsafe? The safety theorists' answer should be as follows: because there is a nearby possible world, i.e., w, where Lincoln is assassinated one mil-

lisecond earlier at t_{n-1} such that he falsely believes that L on the same belief-formation method as that in the actual world, i.e., his memory, at t_{n-1}. w's being close to the actual world is in virtue of N.N.'s falsely believing that L one millisecond later at t_n in the actual world. In addition, N.N.'s falsely believing that L one millisecond later at t_n in the actual world is in virtue of L's being false at t_n. That is to say, L's being false at t_n, at least partly, makes B(L) unsafe at t_{n-1} and thus ~K(L) at t_{n-1}.

If the story unfolds somewhat differently such that Lincoln is assassinated at a later time, say t_{n+1}, in the actual world, then it turns out that N.N.'s belief at t_{n-1} is safe and thus counts as knowledge because in all nearby possible worlds where he believes that L, his belief is true. After all, given the dissimilarity between the actual world and w, w would not be among the nearby possible worlds.[2] That is to say, L's being true at t_n, at least partly, makes B(L) safe at t_{n-1} and thus K(L) at t_{n-1}.

In sum, the safety theorists' move commits them to the idea that what happens (e.g., L's being true or false) at a later time (e.g., t_n) makes a difference to one's epistemic status (e.g., N.N.'s knowing or not knowing that L) at an earlier time (e.g., t_{n-1}). For the sake of simplicity, this implication might be called "future dependence." I can understand that L's being false at t_n makes it the case that ~K(L) at t_n since knowledge is factive. Nonetheless, how could L's being false, at least partly, at t_n make it the case that ~K(L) at t_{n-1}, and L's being true, at least partly, makes it the case that K(L) at t_{n-1}? After all, t_n is later than t_{n-1}! It is no surprise for epistemic externalists to argue that one's epistemic status might depend on external factors not accessible to the subject such as the reliability of the belief-formation methods and the environment. Nonetheless, the implication that one's epistemic status might depend on what happens in the future is still something hard to swallow.

One might think that it is no surprise that what happens at a later time could make a difference to one's epistemic status. For instance, suppose that we are assessing the reliability of a belief-forming process in its early stages. There has been only one instance of the process. In this case, it seems reasonable to say that its reliability depends on the outputs of the process in the future. Regarding this response, I would say that it is true that the outputs of the process in the future would help us to know more about its reliability. Nonetheless, whether the process is reliable or not and to what extent the process is reliable has already been determined by the facts about the belief-forming process as well as the environment in its early stages. Similarly, it is one thing to say that the head-to-tail ratio in the long run could help us to know more about whether a coin is a fair one; while it is another thing to say that whether the coin is fair or not depends on the head to tail ratio, in the long run, because it is intuitive to say that whether the coin is

---

[2] You may choose a_{n+j} where $j \geq 2$ if you think that the interval of two milliseconds fails to render one's belief at t_{n-1} safe.

fair or not has already been determined by the facts about the coin as well as the environment before the coin has been tossed. Thus, it is still unclear that one's epistemic status depends on what happens in the future.[3]

4. *Conclusion*

When faced with an example such as Dead President, accepting and denying the intuition of knowledge in the example constitute two horns of a dilemma for the safety account of knowledge. If the safety theorists accept the intuition of knowledge, then the safety account is outright false. If the safety theorists deny the intuition of knowledge, then the safety account makes one's epistemic status depend on what happens in the future. In sum, the safety account is problematic.[4, 5]
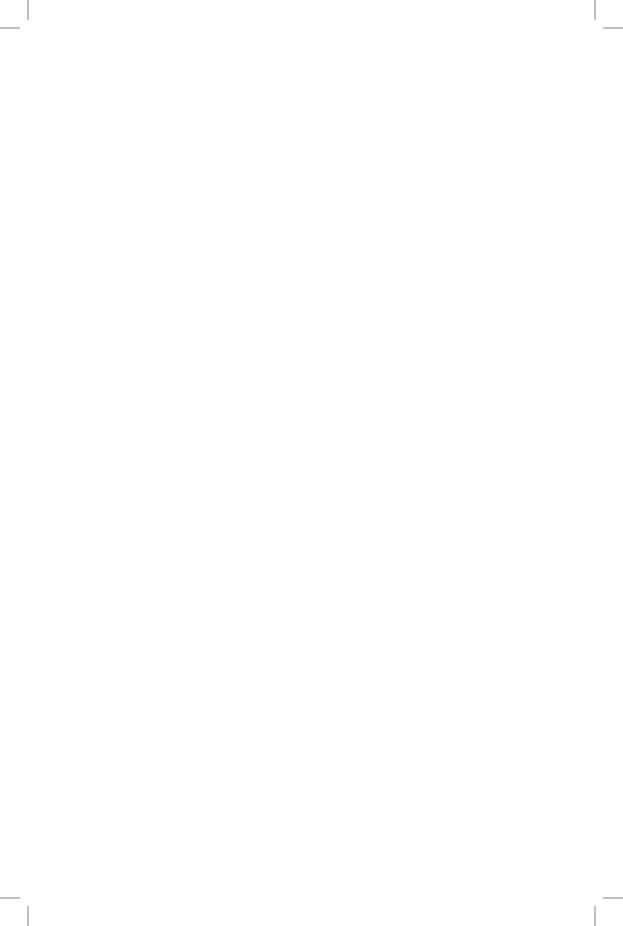
References

Alfano, M. 2009. "Sensitivity Theory and the Individuation of Belief-Formation Methods." *Erkenntnis* 70 (2): 271–281.

Ball, B. 2016. "Knowledge, Safety, and Questions." *Filosofia Unisinos* 17 (1): 58–62.

Baumann, P. 2008. "Is Knowledge Safe?" *American Philosophical Quarterly* 45: 19–30.

Becker, K. 2006. "Reliabilism and Safety." *Metaphilosophy* 37 (5): 691–704.

———. 2008. "Epistemic Luck and the Generality Problem." *Philosophical Studies* 139 (3): 353–366.

———. 2009. "Margins for Error and Sensitivity: What Nozick Might Have Said." *Acta Analytica* 24 (1): 17–31.

———. 2012. "Methods and How to Individuate Them." In K. Becker and T. Black (eds.). *The Sensitivity Principle in Epistemology*. New York: Cambridge University Press, 81–97.

---

[3] It could be conceded that one's epistemic status depends on what happens in the future in cases of future knowledge. Given that knowledge entails truth, it is no surprise that whether one obtains future knowledge or not depends on what happens in the future. For instance, whether I know that *it rains tomorrow* depends on whether it rains tomorrow. If it rains tomorrow, then I could know that *it rains tomorrow*; if it does not rain tomorrow, then I couldn't know that *it rains tomorrow*. However, this kind of future dependence in cases of future knowledge still does not establish the kind of future dependence discussed in this paper. It is one thing to say that one's epistemic status depends on what happens in the future in cases of future knowledge; while it is another thing to say that one's epistemic status depends on what happens in the future in cases of ordinary knowledge even if the truth of the target proposition has been determined, the reliability of the belief-forming process has been determined, the fact about whether the subject obtains good evidence has been determined, the fact about whether subject is in an epistemically friendly environment has been determined, etc.

Beddor, B. and Pavese, C. 2020. "Modal Virtue Epistemology." *Philosophy and Phenomenological Research* 101 (1): 61–79.

Bogardus, T. 2014. "Knowledge Under Threat." *Philosophy and Phenomenological Research* 88 (2): 289–313.

Brueckner, A. and Fiocco, M. O. 2002. "Williamson's Anti-Luminosity Argument." *Philosophical Studies* 110 (3): 285–293.

Comesaña, J. 2005. "Unsafe Knowledge." *Synthese* 146: 395–404.

Dutant, J. 2010. "Two Notions Of Safety." *Swiss Philosophical Preprints* 87: 1–19.

Goldman, A. 1976. "Discrimination and Perceptual Knowledge." *Journal of Philosophy* 73 (20): 771–791.

Grundmann, T. 2020. "Saving Safety From Counterexamples." *Synthese* 197: 5161–5185.

Hirvelä, J. 2019. "Global Safety: How to Deal with Necessary Truths." *Synthese* 196 (3): 1167–1186.

Luper-Foy, S. 1984. "The Epistemic Predicament: Knowledge, Nozickian Tracking, and Scepticism." *Australasian Journal of Philosophy* 62 (1): 26–49.

Manley, D. 2007. "Safety, Content, Apriority, Self-Knowledge." *Journal of Philosophy* 104 (8): 403–423.

Neta, R. and Rohrbaugh, G. 2004. "Luminosity and the Safety of Knowledge." *Pacific Philosophical Quarterly* 85 (4): 396–406.

Nozick, R. (1981). *Philosophical Explanations*. Oxford: Oxford University Press.

Peet, A. and Pitcovski, E. 2018. "Normal Knowledge: Toward an Explanation-Based Theory of Knowledge." *Journal of Philosophy* 115 (3): 141–157.

Pritchard, D. 2005. *Epistemic Luck*. Oxford: Oxford University Press.

———. 2009. "Safety-Based Epistemology: Whither Now?" *Journal of Philosophical Research* 34: 33–45.

Sainsbury, R. M. 1997. "Easy Possibilities." *Philosophy and Phenomenological Research* 57 (4): 907–919.

Sosa, E. 1999a. "How Must Knowledge Be Modally Related to What Is Known?" Philosophical Topics 26 (1&2): 373–384.

———. 1999b. "How to Defeat Opposition to Moore." *Philosophical Perspectives* 13: 141–154.

———. 2015. *Judgment and Agency*. Oxford: Oxford University Press.

Williamson, T. 2000. *Knowledge and its Limits*. Oxford: Oxford University Press.

———. 2008. "Why Epistemology Can't be Operationalized." In Q. Smith (ed.). *Epistemology: New Essays*. Oxford: Oxford University Press, 277–300.

Yamada, M. 2011. "Getting It Right By Accident." *Philosophy and Phenomenological Research* 83 (1): 72–105.

Zhao, B. 2022a. "A Dilemma for Globalized Safety." *Acta Analytica* 37: 249–261.

———. 2022b. "Epistemic Closure, Necessary Truths, and Safety." *American Philosophical Quarterly* 59 (4): 391–401.

———. 2022c. "Knowledge from Falsehood, Ignorance of Necessary Truths, and Safety." *Philosophia* 50: 833–845.

———. 2022d. "Sensitivity, Safety, and Epistemic Closure." *International Journal of Philosophical Studies* 30 (1): 56–71.

———. 2024a. "On Relativizing the Sensitivity Condition to Belief-Formation Methods." *American Philosophical Quarterly* 61 (2): 165–175.

———. 2024b. "On Translating the Sensitivity Condition to the Possible Worlds Idiom in Different Ways." *American Philosophical Quarterly* 61 (1): 87–98.

———. 2025. "On Mentioning Belief-Formation Methods in Sensitivity Subjunctives." *Ergo* 12: 232–246.

———. forthcoming. "On the Tracking Account of Inferential Knowledge." *Inquiry*.

The Holism of Doxastic Justification

ERHAN DEMIRCIOĞLU
Koç University, Istanbul, Turkey

I argue against the orthodox view of doxastic justification, according to which a belief of a given subject is justified for her just in case the subject has a good reason for the belief and she also bases the belief on that reason. The orthodox view is false, I maintain, because there might be unjustified beliefs that are based on the good reasons that support them. The fault lies with the 'particularism' of the orthodox view, which is why it cannot handle those cases where certain 'holistic' considerations render an otherwise justified belief unjustified. Accordingly, I argue for a holistic constraint on doxastic justification: whether a subject that bases a particular belief on a reason that supports it is justified in having that belief depends on what she does, cognitively speaking, with that reason vis-à-vis a considerable portion of her other beliefs.

1. *The Orthodox View of Doxastic Justification*

According to a widely recognized distinction, first clearly introduced by Firth (1978), epistemic justification comes in two kinds: propositional and doxastic. Propositional justification is a relation that holds between subjects and propositions. When a subject S has a good reason R for a proposition p, p is justified for S, irrespective of whether she believes p or not.[1] When Holmes and Watson have the same pieces of evidence

---

[1] Surely, with the caveat that there are no defeaters (or, equivalently, the good reason in question is *overall*), which I will take for granted in what follows. And, here are two points about what I mean by "a reason." First, by "a reason," I don't necessarily mean a reason that does not involve any other reasons as constituents. So, when a subject has, say, three different reasons that jointly (but not necessarily individually) support a proposition, those reasons, collectively taken, constitute *a* good reason for that proposition. Second, I take "a (good) reason" to be whatever it is that is a *source* of propositional justification, be it an experience, a belief, or what have you.

jointly suggesting that the butler is guilty, the proposition that the butler is guilty is justified for both Holmes and Watson, irrespective of whether they believe it or not. However, it does not follow from the fact that S has R, which justifies p, and S believes p, that S's believing p is justified *tout court*: there might still be a clear sense in which her believing p is not justified, a possibility that demands the recognition of a distinct kind of epistemic justification, *viz. doxastic* justification. If, for instance, the reason for which Watson comes to believe that the butler is guilty is not the evidence that he shares with Holmes but the butler's "suspiciously trimmed moustache," then there is a clear sense in which his believing thus is not justified (not "well-founded"), despite the fact that the proposition that the butler is guilty is justified for him. Under these circumstances, Watson's believing that the butler is guilty is not *doxastically* justified; however, Holmes, an epitome of epistemic perfection, surely does not commit Watson's (or any other) blunder and his believing that the butler is guilty is doxastically justified.

What does it take for S's believing p to be doxastically justified? It is uncontroversial that doxastic justification requires propositional justification (and not *vice versa*): in order for S's believing p to be justified, S must have a good reason R to believe it. However, as Watson's "suspiciously-trimmed-moustache" case illustrates, propositional justification does not suffice for doxastic justification. In addition to having propositional justification for p, doxastic justification requires a proper connection between S's believing p and R (i.e., what makes p justified for her): the former must be *based on* the latter.[2] Watson's believing that the butler is guilty is not doxastically justified because it is not based on the good reasons he has for the belief, while Holmes's believing is doxastically justified because it is based on those reasons. According to an orthodox view of doxastic justification, it is necessary and sufficient for S's believing p to be doxastically justified, that p is propositionally justified for S and S's believing p is based on what makes p propositionally justified for her. Briefly put, on this view, doxastic justification is propositional justification *plus* the basing relation.[3]

[2] There are two main competing accounts of the basing relation. On the *causal* account, the reason had by the subject for a given belief must causally initiate or sustain the subject's believing in order for the subject to be justified in having that belief. For defenses of (some different versions of) the causal account, see Moser (1989), Turri (2011), and Ye (2020). And, on the *doxastic* account, having a meta-belief to the effect that the reason is a good reason for the belief is either necessary or sufficient for the belief's being based on the reason. For a defense of the doxastic account, see Tolliver (1982) and Foley (1984). In what follows, nothing much depends on the distinction between these different accounts of the basing relation.

[3] Turri (2010) provides a rich list of representative quotations from prominent epistemologists endorsing the orthodox view of doxastic justification. Here are only two of those. Kvanvig writes: "Doxastic justification is what you get when you believe something for which you have propositional justification, and you base your belief on that which propositionally justifies it" (2003: B1). And, Pollock and Cruz write: "To be justified in believing something it is not sufficient merely to *have* a good reason

The orthodox view of doxastic justification appears to be highly plausible. After all, one might reasonably wonder, what *else* can possibly be required for S's believing p to be justified, other than it being the case that her believing p is based on the good reason R she has for it? If the reason S has for p is good, then it follows that p is justified for S. And, furthermore, if the reason for which S believes p is the good reason she has for p (that is, if she bases her believing on that reason), then it appears that she has all that is required for having a doxastically justified belief that p.

In what follows, I will argue that the orthodox view of doxastic justification is mistaken: doxastic justification cannot be conceived as propositional justification plus the basing relation. Before presenting my argument, however, I would like to first briefly point out how it differs from an influential objection raised by Turri (2010) against the orthodox view. On the basis of some (purported) counter-examples to the orthodox view, Turri argues that that view is fundamentally mistaken because it gets the order of explanation between propositional and doxastic justification wrong. A characteristic feature of the orthodox view is that it holds that the primary unit of justification is propositions (and not actual believing attitudes or, simply, beliefs) and, accordingly, that propositional justification is explanatorily prior to doxastic justification: a theory of epistemic justification must, on this view, *first* explain propositional justification and *then* explain doxastic justification in terms of propositional justification and the basing relation. And, according to Turri, what makes the orthodox view vulnerable to the counter-examples he presents is its definitive commitment to the explanatory priority of propositional justification; and, from this, Turri derives the conclusion that propositional justification (what it is for a proposition to be justified for a subject) must be explained in terms of doxastic justification (what it is for a subject to have a justified belief in a proposition) rather than *vice versa*.

The argument I will offer below proceeds, like Turri's own argument, by way of providing counter-examples to the orthodox view: there are, surprisingly, cases in which S is not justified in having the belief that p, despite the fact that S bases her belief that p on the good reason she has for it.[4] Accordingly, I agree with Turri that, contrary to

for believing it. One could have a good reason at one's disposal but never make the connection. What is lacking (in such a case) is that you do not believe the conclusion *on the basis* of those reasons" (1999: 35–36).

[4] It is important to distinguish two different ways in which the thesis that doxastic justification is propositional justification plus the basing relation might be taken. According to one, the 'thesis' is a definitional truth, according to which the basing relation acts like a place-holder, being *whatever it is* that must be added to propositional justification in order to have doxastic justification. According to the other, the thesis purports to express a substantive truth, which presupposes that we already have a grasp, however tenuous or rudimentary it might be, of the basing relation, and claims that *that* basing relation (perhaps tenuously grasped) is what must be added to propositional justification in order to have doxastic justification.

the orthodox view, doxastic justification cannot be conceived as propositional justification plus the basing relation: propositional justification and the basing relation, jointly taken, are not sufficient for doxastic justification. However, I take the explanatory priority of propositional justification for granted,[5] and I see no good reason to believe that my counter-examples below point toward anything to the contrary. As will become clear, the lesson straightforwardly suggested by those counter-examples is that the orthodox view's fault lies with its 'particularism' (or 'atomism'), according to which whether a belief is doxastically justified is simply a matter of whether it is based on a reason that supports it, regardless of what the subject does, cognitively speaking, with that reason *vis-à-vis* some of her *other* beliefs. The upshot is that there is a significant holistic constraint on doxastic justification, which goes unnoticed by the orthodox view.

This paper is hereafter divided into four sections. In Section 2, I present two counter-examples to the orthodox view of doxastic justification. In Section 3, I address four objections that might be raised against the proposed counter-examples. In Section 4, I elaborate on the sort of holism about doxastic justification ('doxastic holism') suggested by those counter-examples and distinguish it from some better-known holisms in the vicinity. Section 5 sums up the lesson.

2. *Two Counter-Examples to the Orthodox View*

Now, consider the following case:

> *Weather Belief* I ask my friend, Thomas, what he thinks the weather will be like tomorrow. He tells me that it is going to rain tomorrow because a reliable weather report has just said that it is going to rain tomorrow. Then, perhaps out of boredom, I decide to ask him some other questions concerning the weather of a more distant future, questions like "So, what will the weather be like two days from now? One week from now? One month from now?" What I come to realize after these queries is something rather unexpected: to each question concerning the weather of a particular future day,

---

The distinction here is analogous to a distinction between warrant and justification. The 'thesis' that knowledge is true belief plus warrant is definitional, given that warrant acts like a place-holder, being *whatever it is* that must be added to true belief in order to have knowledge. However, the thesis that knowledge is true belief plus justification is substantive and, as Gettier-cases purport to show, might fall prey to counter-examples. Clearly, if the thesis that doxastic justification is propositional justification plus the basing relation is taken as a definitional truth, then it cannot fall prey to counter-examples. In this paper, I take it, plausibly I believe, that that thesis is conceived by the defenders of the orthodox view as purporting to express a substantive truth.

[5] For a defense of the thesis that propositional justification is explanatorily prior to doxastic justification, see Kvanvig (2003) and Silva (2015). It is worth mentioning that Silva (2015) argues that Turri's purported counter-examples can be disarmed by a relatively straightforward revision of the orthodox view (namely, by endorsing that doxastic justification is propositional justification plus the *proper* basing relation). However, Silva's "fix" does not work against the counter-examples I present below.

Thomas gives the answer that it is going to rain on that day because the weather report has just said that it is going to rain tomorrow. For instance, in response to my question "what will the weather be like one month from now?", Thomas says that it is going to rain one month from now because the weather report has just said that it is going to rain tomorrow!

In this case, Thomas intuitively has a good reason for his belief that it is going to rain tomorrow (i.e., the reliable weather report) and might be plausibly taken as basing that belief on that reason.[6] Hence, the orthodox view delivers the result that Thomas's "tomorrow-weather" belief (i.e., the belief that it is going to rain tomorrow) is doxastically justified. However, contra the orthodox view, it seems clear that Thomas's belief is *not* justified. And, this is, to a first approximation, because Thomas has some other "future-weather" beliefs which are also based on the report about tomorrow's weather, which is evidently *not* a good reason for those beliefs. Surely, this should give us a pause. The fact that there are various other future-weather beliefs of Thomas's, each of which is based on what the report says about tomorrow's weather, appears to disqualify his tomorrow-weather belief that is also based on that report from being justified for him.

What might be plausibly drawn in the first instance from the Weather Belief case is this: if a subject bases a significant number of her other beliefs on R, which does not support those beliefs, then her particular belief that p on R, which is a good reason for p, is not (doxastically) justified for her. In other words, whether a subject that bases her belief that p on R, which is a good reason for p, is justified in having that belief depends on what she does, cognitively speaking, with R *vis-à-vis* a considerable portion of her *other* beliefs. Hence, the general lesson is that there is a necessary holistic condition for the doxastic justification of a particular belief, which goes unrecognized by the orthodox view, a condition that concerns whether the subject bases some of her other beliefs on the reason that she bases that particular belief on.

The general lesson above drawn from the Weather Belief case can also be arrived at from an opposite direction. We can imagine cases in which a good reason R supports both p and some other propositions q, r, s, and so on, and also in which a given subject, basing her belief that p on R, does *not* base her beliefs that q, r, s, and so on, on R. Consider, for instance, the following case:

Fortune Belief There is currently nothing in Michael's bank account, but he believes that it will have at least $100000 in three years and he bases this belief on his well-supported belief that he will get promoted in a few months to a top position in the company, which annually pays $200000. Interestingly, however, Michael has various other beliefs (e.g., that he will be able to afford his $100000 dream car in three years, that he will be in a position

---

[6] I don't mean to suggest, nor do I intend to argue, that basing requires articulating (or even having the capacity of articulating) one's own reasons. The case at hand can be suitably revised to fit the reader's preferred account of the basing relation.

to lend $100000 to his brother in three years, and so on), each of which is as clearly supported by his promotion belief as is his future-bank-account belief, while he does not base them on that promotion belief.

In this case, Michael bases his future-bank-account belief on the promotion belief that supports it. Hence, the orthodox view delivers the result that the future-bank-account belief is justified for Michael. However, contra what the orthodox view entails, it seems clear that the future-bank-account belief is *not* justified for Michael. And, this is, to a first approximation, because Michael has various other beliefs that he does not base on the promotion belief, despite the fact that the promotion belief *is* as clearly good a reason for those beliefs as it is for his future-bank-account belief. This should give us a pause. The fact that Michael does not base his other beliefs on the promotion belief that supports them appears to disqualify his future-bank-account belief that is supported by the promotion belief from being justified for him.

What might be plausibly drawn in the first instance from the Fortune Belief case is this: if a subject does not base a significant number of her other beliefs on R, which supports those beliefs, then her particular belief that p on R, which is a good reason for p, is not (doxastically) justified for her. In other words, whether a subject that bases her belief that p on R, which is a good reason for p, is justified in having that belief depends on what she does, cognitively speaking, with R *vis-à-vis* a considerable portion of her *other* beliefs. Hence, the general lesson is the same as the one drawn from the Weather Belief case, *viz.* that there is a necessary holistic condition for the doxastic justification of a particular belief, which goes unrecognized by the orthodox view, a condition that concerns whether the subject bases some of her other beliefs on the reason that she bases that particular belief on.

The Weather Belief case is intended to show that *basing* some beliefs on a reason that does *not* support them might disqualify a distinct belief that is based on the very same reason that actually supports it from being justified. And, the Fortune Belief case is intended to show that *failing* to base some beliefs on a reason that *supports* them might disqualify a distinct belief that is based on that reason that actually supports it from being justified.

The argument from the two cases presented for a holistic constraint on doxastic justification proceeds in two main steps. The first step is that the intuitive reaction to the cases is, I take it, that there is something epistemically wrong with the target beliefs of the subjects in question—more specifically, that those beliefs are unjustified. It is not merely that, in the Weather Belief case, for instance, Thomas's other future-weather beliefs are unjustified but also that there is something epistemically wrong with his tomorrow-weather belief—that is, that that belief is unjustified. The reaction I have (and expect that the reader will have) is something along the following lines: "Wait! If Thomas has all these other future-beliefs based on the report about

tomorrow's weather, that surely renders his tomorrow-weather belief epistemically problematic!" (or, more specifically: "Wait! If Thomas has all these other future-beliefs based on that report, that surely renders his tomorrow-weather belief unjustified!") Furthermore, given that the subjects have propositional justification for their target beliefs, what is epistemically wrong with those beliefs must be that they are not *doxastically* justified.[7] And, the second step of the argument is that what explains what renders the beliefs in question doxastically unjustified is that there is a holistic constraint on doxastic justification: what the subject does with a reason *vis-à-vis* some of her other beliefs might render an otherwise justified belief unjustified.

Let me now summarize the main result of this section. The orthodox view of doxastic justification is particularist (or atomistic) in the sense that it presumes that whether a particular belief is justified depends solely on whether its subject has a good reason for it and she bases *it* on that good reason. The presumption here is that whether a particular belief is justified has nothing to do with how its subject, so to speak, cognitively stands with respect to some of her other beliefs. However, the cases presented above point toward the conclusion that doxastic justification is holistic in the sense that whether a particular belief is justified for a subject cannot be decided in isolation from her cognitive doings with some of her other beliefs, in particular from whether she bases those other beliefs on a reason which supports that particular belief but does not support them.[8]

3. *Objections Answered*

There are four objections I want to address in this section that might be levelled against the cases proposed above. The first and second objections reject that the proposed cases are genuine counter-examples to the orthodox view as it currently stands, while the third and fourth objections acknowledge that they are genuine counter-examples but they hold that the orthodox view might be saved by a suitable revision preserving its particularism and nothing as dramatic as a holistic constraint on doxastic justification is needed.

The first objection I want to consider is the most straightforward one, which simply insists that in the given cases, the target beliefs that are based on the good reason that supports them (e.g., Thomas's tomorrow-weather belief) *are* doxastically justified, despite the fact that the other beliefs (e.g., Thomas's other future-weather beliefs) are not

[7] I assume, plausibly I believe, that there are only two kinds of epistemic justification relevant here and that they are propositional justification and doxastic justification.

[8] Since I believe that the general moral to be drawn from the Fortune Belief case is the same as the one drawn from the Weather Belief case, I will, for convenience's sake, formulate my discussion below mainly only in terms of the latter.

doxastically justified. According to this objection, what the orthodox view delivers with respect to these cases is correct and, therefore, no revision is called for.

My response to this objection is that it is based on a misunderstanding of the current dialectic. The question at issue is not what the orthodox view delivers with respect to the proposed cases but whether its deliverances are correct, and merely insisting that they are correct does not make them so. The intended dialectic has indeed a familiar thought-experimental structure, where an intuitively compelling case is presented against a certain view. I take it as intuitively clear that in the Weather Belief case, for instance, Thomas's basing his other future-weather beliefs on tomorrow's weather report casts serious doubt on the epistemic standing of his tomorrow-weather belief: there is clearly something epistemically wrong with his tomorrow-weather belief, given what he does, cognitively speaking, with his other future-weather beliefs.[9] Furthermore, I take it that the obvious explanation of what it is that is epistemically wrong with Thomas's tomorrow-weather belief is that it is *not* doxastically justified (given especially that it *is* propositionally justified). So, it appears that the orthodox view delivers an intuitively false verdict for the Weather Belief case: while it seems that Thomas's tomorrow-weather belief is not doxastically justified, it delivers the result that it is. A defender of the objection that the proposed counter-examples do not call for any revision of the orthodox view needs to argue either that there is nothing epistemically wrong with Thomas's tomorrow-weather belief or that what is epistemically wrong with it has nothing to do with doxastic justification. And, I take it that neither claim is compelling.

The second objection maintains that the proposed cases are psychologically implausible and thus fail to be genuine counter-examples to the orthodox view. Why would anyone act, cognitively speaking, in the way Thomas does? Thomas's cognitive behavior is highly eccentric and hard to make sense of: human beings do not normally cognitively act like *that*. How can Thomas fail to recognize that the report about tomorrow's weather is not a good reason for the belief that it is going to rain one month from now? Thomas's failure is, one might insist, so peculiar that it is hard to take his case seriously as a genuine challenge to the orthodox view.

I agree with this objection that Thomas's failure is peculiar, but it does not follow from this peculiarity that the Weather Belief case fails to be a genuine counter-example to the orthodox view of doxastic justi-

---

[9] I suspect that some readers might be unmoved by the proposed cases, shrug their shoulders and maintain "Everything seems to be in order with Thomas's tomorrow-weather belief!" This sort of 'dismissive' attitude is to be expected (as it is the case with many other thought-experiments in epistemology and other fields—cf. for instance, reactions to the fake barn cases (see Lycan 2006)), but my prediction is that it will be limited. I have nothing to offer in order to convince the readers that do not feel the pull of the proposed cases.

fication. The crucial point about the Weather Belief case is not whether it is psychologically plausible but whether its possibility is excluded by the orthodox view, and the answer to the latter question is no. Similar considerations apply to some famous philosophical thought-experiments. The crucial point about the Knowledge Argument, for instance, is not whether it is psychologically plausible (e.g., not whether it is psychologically plausible to assume that a human being, Mary, can live in an isolated room, without having no visual contact with colors, for a very long period of time and learn, basically through reading, everything physical about colors), but whether it is *possible*, given what physicalism says about how the world is. And, its possibility, not its psychological plausibility, makes the Weather Belief case a genuine counter-example to the orthodox view.

The third objection acknowledges the intended morale of the proposed cases, *viz.* that in the Weather Belief case, for instance, Thomas's tomorrow-weather belief is doxastically unjustified; and, it builds on a certain diagnosis about what goes wrong in those cases. Why is it that the fact that the subject bases some of her other beliefs on a reason that does not support them disqualifies *this* particular belief of hers based on that reason, which actually supports it, from being justified for her? According to a proponent of the third objection, there is a plausible answer to this question that is consistent with the particularism of the orthodox view, which goes as follows. If a subject bases a significant number of some of her beliefs on R, which does not support those beliefs, then this indicates something significant about her basing the particular belief that p on R, which supports it, *viz.* that particular basing does not result from its being the case that R supports p.[10] This is because, surely, if the subject's basing her belief that p on R resulted from it being the case that R supports p, then she would not base those other beliefs of hers on R, which does not support them. Consider a 3-year-old child who gives "4" as an answer to every multiplication problem. Clearly, those answers indicate that his saying "4" in response to "2×2=?" does not result from (his appreciation of) the truth that 2×2=4. Similar considerations apply, for instance, to our subject, Thomas, who bases all his beliefs about future weather on what the report says about tomorrow's weather. Furthermore, if the subject's basing p on R, which supports p, does not result from it being the case that R supports p, then the belief that p is not justified for the subject. If, for instance, the subject's basing p on R, which supports p, results from it being the case that the subject likes the sound of R, then the belief that p is clearly not justified for the subject. And, what explains why this is so is that, in such a case, what is responsible for the basing is not that R supports p.

And, if this diagnosis of what is wrong with, for instance, Thomas's tomorrow-weather belief, is correct, then, the third objection goes, what

10 For an extended and illuminating discussion of this idea, see Wedgwood (2006).

fundamentally matters to the doxastic justification of that particular belief is not what Thomas does, cognitively speaking, with some of his other future-weather beliefs but whether Thomas bases that belief on the tomorrow's weather report *because* that report supports it. Furthermore, it seems that this accords well with the particularism of the orthodox view, and all that is needed to save the view is to strengthen it by adding the following constraint on doxastic justification: a subject's belief that p is doxastically justified by R only if she bases her belief that p on R because R supports p. According to the proposal suggested by this objection, what a subject does with some of her other beliefs might serve as *evidence* that there is something wrong with a particular belief of hers that is based on a reason that supports it (more specifically, as evidence that it is not based on the reason that supports it *because* the reason supports it), but this does not mean, the objection goes, that there is a holistic constraint on doxastic justification.

I believe that this 'causal-responsibility' diagnosis—*viz.* what is wrong with Thomas's tomorrow-weather belief is that what is causally responsible for its being based on the weather report is not that the weather report supports the belief—is attractive and deserves careful attention; however, I don't believe, unfortunately for the orthodox view, that it survives close scrutiny. Note that it is not clear at all how the epistemic support relation between a reason and a belief can be causally responsible for the latter's being based on the former. The epistemic support relation appears to be causally inert and as such does not appear to be capable of being causally responsible for anything, let alone a belief's being based on a reason.[11] In response to this observation about the causal powers of the epistemic support relation, a natural move available to a proponent of the causal-responsibility diagnosis is to go 'doxastic' and appeal to the subject's *belief* regarding the support relation in question (given that beliefs about the support relation can be causally effective even if the support relation itself cannot). Given this move, the causal-responsibility diagnosis amounts to this: what is wrong with Thomas's tomorrow-weather belief is that what is causally responsible for it being based on the weather report is not that *Thomas believes that* the weather report supports the belief. Now, however, the obvious problem is that there is nothing in the Weather Belief case that excludes the possibility that what is causally responsible for Thomas's tomorrow-weather belief's being based on the weather report *is* that

---

[11] I do not take this reply that appeals to the causal inefficacy of the normative to be a decisive argument against the objection at hand. The point of the reply is rather that endorsing the objection comes with a significant *cost*, which is that such a move needs to explain how normative facts can figure in causal explanations and that it is not intuitively clear at all how this can be done. A thesis of the causal efficacy of the normative is perhaps ultimately defensible (see, for instance, Wedgewood (2006)) but I see it as a serious, if not deadly, bullet to bite. Thanks to an anonymous reviewer for pressing on this issue.

Thomas believes that the weather report supports the belief. Thomas might believe that the weather report supports his tomorrow-weather belief, and his basing the tomorrow-weather belief on the weather report might result from this supporting belief of his, *despite the fact that* he also bases his other future-weather beliefs on the very same weather report. And, if this is so, then the causal-responsibility diagnosis is mistaken: what is wrong with Thomas's tomorrow-weather belief is not what the causal-responsibility diagnosis says it is.

There is a further point I want to make against the objection at hand. I have just argued that the causal-responsibility diagnosis is mistaken; however, one might reasonably wonder whether, even if that diagnosis were correct, it would follow that there is no holistic constraint on doxastic justification. Assume, for the sake of the argument, that what goes wrong in the proposed cases is that what is causally responsible for the basing is not that the reason supports the belief. Now, let us raise the following question: How can we tell whether a given subject bases her belief that p on R because R supports p? We now assume that the fact that R supports p can be causally responsible for the fact that the subject bases her belief that p on R; however, this assumption surely does not entail that we can tell whether the causal-responsibility claim in question holds in a particular case without taking a look at what the subject does with R *vis-à-vis* some other beliefs. It might well be that the *only* way to tell whether the subject bases her belief that p on R because R supports p is to check whether, for instance, the subject bases some of her other beliefs on R, which does not support those beliefs: if she does, then we can plausibly conclude that what is causally responsible for her basing the belief that p on R is not that R supports p; and if she does not, then not. And, if the only way to tell whether the causal-responsibility claim holds in a particular case is to do such a holistic checking, then it seems reasonable to argue that the truth of causal-responsibility diagnosis as such does not guarantee the particularism of the orthodox view. Consider: if the only way to tell whether something is an X is, say, by figuring out whether that thing has blue eyes, then it seems reasonable to say that being an X is nothing other than having blue eyes. In a similar vein, if the only way to tell whether a belief is doxastically justified is by looking at what other beliefs the subject bases on the reason, then it seems reasonable to say that what makes a belief doxastically justified is at least in part a matter of what other beliefs the subject bases on the reason.

However, I expect that some philosophers might not be inclined to agree with my appeal in the previous paragraph to a sort of 'verificationism:' they might object that the question of what it is to be an X should not be conflated with the question of how one can tell whether something is an X. This brings me to the last point I want to make against the objection at hand. If the only way to tell whether the causal-responsibility relation holds in a particular case is by doing some 'holis-

tic checking,' then you can perhaps save the *letter* of the particularism of the orthodox view but not its *spirit*. That is to say, a proponent of the orthodox view might insist that doxastic justification is particularistic in nature while accepting that a sort of holistic checking is required to tell whether it obtains. I take this conception of the orthodox view as a serious departure from the original one, where holistic checking is absent. And, if this paper succeeds in persuading some proponents of the orthodox view that holistic checking is required, then I would be glad that it has served its purpose.[12]

According to the fourth, and final, objection I wish to consider, the cases presented in the previous section (such as the Weather Belief case) are best conceived as pointing at the need to impose, contra the causal-responsibility diagnosis examined above as well as the holistic thesis I propose, not a condition directly on the doxastic justification of a particular belief but a condition directly on its basing on the relevant piece of reason. According to this objection, a subject can only *base* her belief that p on R, which supports that belief, if she does not base some of her other beliefs on R, which does not support them. This objection maintains that, in the Weather Belief case, for instance, Thomas fails to base his tomorrow-weather belief on the report because he bases a significant number of his other future-weather beliefs on the report. The proposal that the Weather Belief case points at the need to impose a condition *directly* on the basing relation (and thereby a condition *at one remove* on doxastic justification) saves the orthodox view from the threat posed by that case: if Thomas fails to base the tomorrow-weather belief on the report, then the orthodox view delivers the intuitively correct result that that belief is not doxastically justified.

There are two points I wish to make against the move suggested by this objection. First, this move can only save the particularist letter of the orthodox view of doxastic justification but not its particularist spirit. The spirit of the orthodox view is particularist regarding the conditions for the basing relation in that that account takes it that whether a particular belief is based on a reason has nothing to do with whether some other beliefs are based on that reason.[13] However, the move at hand incorporates a holistic constraint on the basing relation, thereby sacrificing the particularist spirit. Second, placing holistic constraints along the lines suggested by this move on the obtaining of the basing relation is implausible. Whatever the correct account of the basing relation eventually turns out to be, it certainly appears to be possible, to

[12] This rather extended reply of mine to objection 3 has benefited greatly from the critical observations of an anonymous reviewer about its previous version. I thank the reviewer for insightful remarks.

[13] It is worth noting here that both of the two main (namely, the *causal* and the *doxastic*) accounts currently available of the basing relation are particularist in this sense (see fn. 2 above). On both of these accounts, whether the basing relation holds between a reason and a belief has nothing to do with what the subject does with the reason *vis-à-vis* some other beliefs.

be a datum if you wish, that Thomas bases both the tomorrow-weather belief and some other future-weather beliefs on the report about tomorrow's weather, something which is deemed impossible by the move under consideration.

4. *Reflections*

Having answered what I take to be the most significant objections against the proposed counter-examples to the orthodox view of doxastic justification, I would now like to further elaborate on the holistic thesis suggested by those counter-examples, which I henceforth call "doxastic holism" ("DH"). Recall that DH is the thesis that whether a subject that bases her belief that p on R, which is a good reason for p, is doxastically justified in having that belief depends on what she does with R *vis-à-vis* a considerable portion of her other beliefs.

The sort of holism proposed in this paper concerns doxastic justification and must be carefully distinguished from holism about propositional justification (or, propositional holism). The core idea of propositional holism is that what makes a particular proposition p justified is not another particular proposition but a feature of the set of propositions to which p belongs. As is well-known, a straightforward kind of propositional holism is offered by coherentism, *viz.* the thesis that p is justified for a subject S just in case the set of propositions, which are available to S and of which p is a member, is coherent (to a certain degree).[14] It must be clear that doxastic holism does not require (but is surely consistent with) propositional holism.[15] Doxastic holism is the thesis that there are holistic conditions that apply to doxastic justification, and this thesis can be true without there being holistic conditions that apply to propositional justification. Consider "propositional particularism," which allows for the possibility that p might be justified for S despite the fact that the set of propositions, available to S and of which p is a member, is not coherent (to the degree deemed required by coherentism). The point is that doxastic holism is consistent with propositional particularism.

It is worth emphasizing that there are three different readings of the DH, and it is the weakest reading that is intended here. The strong DH claims that the justification for a subject of a particular belief re-

[14] For defenses of coherentism, see BonJour (1985) and Lehrer (1974).

[15] This is good news for doxastic holism because it means that doxastic holism is not necessarily vulnerable to some of the objections that might be raised against propositional holism. Consider the question whether we can base our beliefs on holistic facts about our belief sets. If the answer is no, then it follows from propositional holism that our beliefs cannot be doxastically justified. (*Cf.* Cohen (2002) and Korcz (2000).) And, given that doxastic justification is required for knowledge, this in turn would entail skepticism. So, assuming that skepticism is false, that would give us an excellent reason to reject propositional holism. However, surely, such an argument against propositional holism does not have a direct bearing on doxastic holism as such, given that doxastic holism does not require propositional holism.

quires that she *have* some other beliefs (and also that she does not base those beliefs on a reason which supports that particular belief but does not support them). However, the weak DH imposes only a conditional requirement having to do with what must obtain if the subject has some other beliefs (e.g. that, as the Weather Belief case is intended to show, *if* the subject has some other beliefs, then she does not base those beliefs on a reason which supports the target particular belief that she bases on but does not support them). Unlike its strong cousin, the weak DH does not deny that a subject that has only one belief can be justified in believing it (more on this below). Finally, the moderate DH imposes a counterfactually-strengthened conditional requirement having to do with what must obtain if the subject has or *had* some other beliefs (e.g., that if the subject has or had some other beliefs, then she does not or *would not*, respectively, base those beliefs on a reason which supports the target particular belief that she bases on but does not support them). The moderate DH is weaker than the strong DH in that it does not deny, unlike the strong DH, that a subject that has only one belief can be justified in believing it. And, it is stronger than the weak DH in that it does not exclude, unlike the weak DH, the possibility that a subject that has only one belief, based on a reason supporting it, might be unjustified in believing it. It is clear that only the weak DH is supported by the Weather Belief and Fortune Belief cases as they are presented,[16] and by "DH," I will henceforth mean the weak version.

Another clarificatory point I wish to make is about the qualification "a considerable portion" involved in the thesis. This qualification is significant because it seems unclear whether what the subject does with R *vis-à-vis* one or two of her other beliefs, if she has any, has any undermining effect on the justificatory status of her belief that p based on R, which is a good reason for p. For instance, it seems unclear whether Thomas's basing his belief that it is going to rain one week from now on the report about tomorrow's weather, *taken by itself*, disqualifies his tomorrow-weather belief based on that report from being justified for her. However, it seems clear that if Thomas bases a considerable number of his future-weather beliefs on that report, then the disqualification effect takes place.[17]

DH is substantively different from a sort of holism about epistemic justification inspired by Quine's classic "Two Dogmas of Empiricism" (1951) and also from the sorts of holism about the mental defended famously by Davidson in various works including, for instance, "Radi-

[16] This does not of course mean that there are no cases in the vicinity that might be taken as supporting one of the other (stronger) versions of DH. I am especially sympathetic to the idea that the cases presented in section 2 might be rephrased in a way that supports the moderate DH. However, I wish to remain non-committal about this issue in this paper.

[17] How many beliefs make up a considerable portion? Evidently, no definite answer is forthcoming and vagueness plagues here as it does many other areas.

cal Interpretation" (1973). Quine maintains that "our statements about the external world face the tribunal of sense experience not individually but only as a corporate body" (1951: 38). According to Quinean holistic epistemology, the basic unit of justification is "theories" as a whole but not particular beliefs: evidential support relates systems of beliefs, but not particular beliefs, to pieces of evidence. However, the orthodox view I have targeted above presumes that a given subject might have a *good reason for a particular belief* and hence that evidential support relates particular beliefs, but not systems of beliefs, to pieces of evidence. And, my argument against the orthodox view shares with that view these presumptions. The sort of holism I propose about justification is particularist in *this* sense. The holism of DH stems from the claim that whether the belief that p based on R, which supports p, is justified for the subject cannot be decided in isolation from her cognitive doings with R regarding some of her other beliefs (if she has any). However, this is consistent with the sort of particularism rejected by Quinean holism.

Davidson defends a number of non-equivalent theses, each of which might be properly viewed as a sort of holism about the mental; and, here I will address two of those, both of which pertain to belief possession. One is about what it takes to have a *single* belief. Davidson straightforwardly argues from what he takes to be the correct theory of individuation of concepts as components of beliefs that beliefs necessarily come as "a matched set" (1982: 319) and that it thereby does not make sense to have only one belief. DH is about what it takes to have a belief that is justified and makes no commitments about the conditions for having a belief as such, and it is consistent both with the thesis that there is no such thing as a singleton belief set and with its negation.

DH does not entail that having a single belief requires having many, nor does it entail that having a single justified belief requires having some. For all DH claims, not only might there be a subject that has only one belief but there might also be a subject that has only one belief, which is justified. DH does not deny that neither the having nor the justification of a particular belief requires neither the having nor the justification of other beliefs. It is important to note here that in the Weather Belief case, what makes it true that Thomas's tomorrow-weather belief is not justified is *not* that Thomas's other future-weather beliefs are not justified (on account of the fact that they are based on the report about tomorrow's weather). Even if Thomas's other future-weather beliefs *were* justified, the tomorrow-weather belief would *not* be justified if those other beliefs were based on the tomorrow's weather report. To see this, suppose that Thomas's other future-weather beliefs are based not only on the tomorrow's weather report, which does not support them, but also on some other pieces of evidence that support them (e.g., on some other future-weather reports). Under these circumstances, it is plausible to think that Thomas's other future-weather

beliefs, being based both on a reason that does not support them and on a reason that supports them, are justified. However, it still intuitively seems that Thomas's tomorrow-weather belief is not justified, given that Thomas bases some of his other future-weather beliefs on the report about tomorrow's weather, the sole basis of the tomorrow-weather belief, which does not support them.

The other Davidsonian holistic thesis I wish to consider concerns what it takes to have a belief *set* (rather than a single belief). Davidson argues from the conditions that a radical interpreter finds herself in that belief possession as well as belief attribution is subject to some (minimal) rationality (especially, coherence) constraints. It might be observed in this connection, for instance, that rather than attributing both the belief in a particular proposition and its negation to a subject, who asserts both that proposition and its negation, we tend not to attribute her either. It is, for Davidson, only against the background of at least some degree of overall rationality had by a subject's belief set that we can speak of her having a belief set at all (and *a fortiori* of her having an irrational belief). However, considerations similar to the one I have made in relation to the former sort of Davidsonian holism also apply to this sort: DH is about what it takes for a belief to be justified and makes no commitments about the conditions for having a belief set. DH is consistent both with the thesis that a belief set necessarily satisfies some degree of rationality and its negation.

It is worth noting that the fact that a given belief set satisfies some minimal rationality constraints does not guarantee that the condition imposed by DH on the justification of particular beliefs is thereby satisfied. We can imagine Thomas in the Weather Belief case, for instance, having a deductively consistent belief set. Indeed, nothing in the description of the Weather Belief case entails that Thomas's belief set is deductively inconsistent. However, Thomas's tomorrow-weather belief is not justified, which entails that satisfying some minimal rationality constraints does not entail that the condition imposed by DH is thereby satisfied.

It might be proposed, perhaps by a tough-minded Davidsonian holist, that the rationality constraints that need to be satisfied for there to be a belief set include *inter alia* the condition imposed by DH on the justification of particular beliefs. However, this is highly implausible. If this proposal were true, then the Weather Belief case as described would be impossible since in that case, Thomas is described as *having* a belief set one of whose members is a belief (namely, the tomorrow-weather belief) that does not satisfy the condition imposed by DH. However, the Weather Belief case appears to be clearly conceivable; and, therefore, irrespective of the final verdict about its plausibility, Davidson's insight that there are some basic rationality requirements for having a belief set should not be stretched to its breaking point, where we lose the distinction between having a belief set that satis-

fies those rationality requirements and having a belief set all of whose members are justified.

   DH claims that a given subject is justified in believing that *p* on *R*, which supports *p*, only if she does not base a considerable portion of her other beliefs on *R*, which does not support those beliefs. DH is not to be confused with the claim that a subject is justified in believing that *p* on *R*, which supports *p*, only if she does not base a considerable portion of her other beliefs on (a different piece of evidence) *R´* which does not support those beliefs. DH concerns what the subject basing a particular belief on a piece of evidence does with *that very same* evidence *vis-à-vis* some other beliefs, and the latter concerns what that subject does with a *different* piece of evidence *vis-à-vis* some other beliefs. Evidently, DH is consistent both with this latter "different-evidence" claim and its denial. I think (but will not argue here) that the different-evidence claim is implausible; however, another claim in its vicinity, sometimes attributed to Davidson (e.g., by Kim (1988: 393)), is plausible, *viz.* that a subject that does not get evidential support relations right in general (or a subject whose cognitive "output" is not regulated and constrained in general by those relations) cannot be said to *have* a belief system.[18] Again, DH is consistent both with this "regulation-by-support-relations" claim and with its denial.

5. *Conclusion*

Let me conclude by highlighting the central points that have emerged in our discussion. The orthodox view of doxastic justification maintains that basing a belief on a reason that supports it suffices for the justification of that belief. I have argued above that the orthodox view is false: there might be unjustified beliefs that are based on reasons that support them. This is because, contra what is presumed by the orthodox view, there is a holistic constraint on doxastic justification: the justification of a belief based by the subject on a reason that supports it depends on what she does with that reason *vis-à-vis* some of her other beliefs. The sort of epistemic holism defended in this paper, I have argued, neither entails nor is entailed by coherentism, Quinean epistemic holism or Davidsonian holisms about the mental. Its distinct character makes it, I believe, all the more interesting and its hitherto absence in the epistemological literature all the more curious.

References

BonJour, L. 1985. *The structure of empirical knowledge*. Cambridge: Harvard University Press.

Cohen, S. 2002. "Basic knowledge and the problem of easy knowledge." *Philosophy and Phenomenological Research* 65 (2): 309–329.

---

[18] This is a way of cashing out a fundamental insight mentioned above of Davidson's, namely, that having a belief system requires having minimal rationality.

Davidson, D. 1973. "Radical interpretation." *Dialectica* 27: 314–328.

Davidson, D. 1982. "Rational animals." *Dialectica* 36: 318–327.

Firth, R. 1978. "Are epistemic concepts reducible to ethical concepts?" In A. Goldman and J. Kim (eds.). *Values and Morals*. Dordrecht: Reidel Publishing Company, 215–229.

Foley, R. 1984. "Epistemic luck and the purely epistemic." *American Philosophical Quarterly* 21 (2): 113–124.

Kim, J. 1988. "What is 'naturalized epistemology'?" *Philosophical Perspectives* 2: 381–405.

Korcz, K. 2000. "The causal-doxastic theory of the basing relation." *Canadian Journal of Philosophy* 30 (4): 525–550.

Kvanvig, J. 2003. "Propositionalism and the perspectival character of justification." *American Philosophical Quarterly* 40 (1): 3–18.

Lehrer, K. 1974. *Knowledge*. New York: Oxford University Press.

Moser, P. 1989. *Knowledge and evidence*. Cambridge: Cambridge University Press.

Pollock, J. and Cruz, J. 1999. *Contemporary theories of knowledge*. New York: Rowman & Littlefield.

Quine, V.W.O. 1951. "Two dogmas of empiricism." *The Philosophical Review* 60 (1): 20–43.

Silva, P. 2015. "On doxastic justification and properly basing one's beliefs." *Erkenntnis* 80 (5): 945–955.

Tolliver, J. 1982. "Basing beliefs on reasons." *Grazer Philosophische Studien* 15 (1): 149–161.

Turri, J. 2010. "On the relationship between propositional and doxastic justification." *Philosophy and Phenomenological Research* 80 (2): 312–326.

Turri, J. 2011. "Believing for a reason." *Erkenntnis* 74 (3): 383–397.

Wedgwood, R. 2006. "The normative force of reasoning." *Nous* 40 (4): 660–668.

Ye, R. 2020. "A causal-doxastic theory of epistemic basing." In J. Carter and P. Bondy (eds.). *Well-founded Belief: New essays on the epistemic basing relation*. New York: Routledge, 15–33.

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 clay-works, 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*. PerfectBound.

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.

The Relational and Doxastic Approach to Religious Diversity[1]

DANIELE BERTINI
University of Rome Tor Vergata, Rome, Italy

The main purpose of my paper is to work out an experiential notion of religious diversity. This means characterising religious diversity in terms of relational and doxastic features. Such a proposal differs from mainstream approaches to religious beliefs (at least from a philosophical viewpoint) because these handle the epistemic dimension of faith as a purely epistemic matter. On the contrary, my idea consists of highlighting how the epistemic evaluation of opposing religious propositions is the outcome of an interpersonal process of evidence sharing wherein the particularities of the involved individuals matter. In the introductory section, I will define the topic of my paper. In the subsequent one, I will characterise how the mainstream approach to religious diversity and my own contrast. In the third section, I will develop in a few details the main reasons for why the epistemic approach is unsatisfactory. In the fourth one, I will provide a thought experiment for religious diversity, and I will set forth my considerations about what the thought experiment shows, with a focus on issues about the epistemology of religious disagreements. I will conclude my paper with a brief overview of the main consequences of my proposal.

1. *Introduction*

The purpose of my paper is to raise several doubts about common assumptions framing the debate on religious diversity. According to the mainstream understanding of the notion, religious beliefs are to be addressed in terms of the ordinarily assumed epistemic intuitions that govern the epistemological evaluation of believing. This means that

---

[1] Dedicated to Ginevra Zoni, my May Queen. She gives me light, and for her I carry my fire.

dealing with religious beliefs involves the same conceptual analysis at work in any other domains of theoretical inquiry. My view is that such a project is questionable because religious beliefs are sui generis beliefs in reason of their irreducibility to a univocal meaning shared across different individuals engaging with them (Bertini 2020). As such, a bit of relational and doxastic features of religious beliefs, which are generated by the phenomenology of the assent to them, reveal relevance as much as purely epistemic characterisation (Bertini 2018).

Before entering in the argument, few statements on what the paper is concerned with and what is not. Religious diversity is a notion characterising a multiplicity of very different phenomena. From societal and historical viewpoints, distinct religious groups of believers adhering either to the same tradition or to alternative traditions may conflict and have social and institutional relationships due to their belonging either to the same context or competing ones. From a taxonomical viewpoint, religions that differ in core beliefs as they are expressed by the frame of narratives concerning religious referential targets give voice to different mindsets and cultural inheritances. From a content-of-faith viewpoint, beliefs addressing a religious object, ritual, or institutional device exhibit what individuals hold to be the case relating to their religious ideas.

Each of these domains of inquiry defines religious diversity in its terms and is concerned with specific domain-related questions. My efforts will be entirely devoted to saying something about the last of the domains mentioned above. There is no epistemic assumption of primacy towards philosophy or implicit hierarchical belief in my statement. Simply, as a philosopher of religion, I am interested in the truth of religious propositions. As a consequence, I am committed to theorising about what happens when individuals disagree over what they hold the truth of their beliefs is. Particularly, as an epistemologist of religious beliefs, I aim to provide the form that a rational answer to such disagreements should have. I do not believe that my task is of more importance nor more informative about religious diversity than those related to alternative approaches: it is indeed a matter of fact that relevant considerations from the three different domains which I distinguished above can be found within the theorisations developed by the use of the notion of religious diversity.

In what follows, I will refer to religious diversity as means to characterise the semantic issues relating to the divergent content of apparently incompatible beliefs, and I will not presume to have anything important to say about how religious diversity is drawn in other fields of inquiry addressing religions.

The peculiarity of my approach to religious diversity consists of characterisation of it in terms of relational and doxastic features. Such a proposal differs from mainstream approaches to religious beliefs (at least from a philosophical viewpoint) because these handle the epis-

temic dimension of faith as a purely epistemic matter. I assent to the trivial claim that the epistemic dimension of faith is an epistemic matter. However, I do not accept that it is purely such an affair: the relational and doxastic features of religious beliefs play an equally relevant role.

2. *Relational and Doxastic vs. Purely Epistemic Approach to Religious Beliefs*

Let me begin with sketching the notions of *religious affiliation* and *belief justification* that prevail in mainstream literature. Such a clarification will help distinguish my own approach from rival ones.

Religious affiliation, first. Any tradition can be picked out by reason of a doctrinal system (Pouivet 2013). Such a system is constituted by a set of framework propositions that express a structured worldview. All adherents to a tradition assent to such a set of propositions (Alston 1992; Plantinga 1999; Harrison 2006; van Inwagen 2010). This implies that the religious beliefs of individuals adhering to opposing traditions may be dealt with by intuitions about the exclusivity of truth and other non exotic principles concerning how to think about the semantic incompatibility of beliefs. A notable consequence of this notion of religious affiliation is that any individuals adhering to a tradition may be represented as prototypical exemplars that stand for any other individual adhering to that very tradition. This being the case, from an epistemic viewpoint, all members of a doxastic group are interchangeable. This is exactly where the doxastic dimension of believing vanishes into the epistemic one. Suppose that all adherents to a tradition T give a literal assent to belief P. Real differences concerning how to understand the content of P are obscured by the alleged epistemic stability of the semantic value of P. That is, according to a purely epistemic approach whichever interpretation of P which is thought to be legitimate in T, has the same semantic value and represents the same content.

I will now move to the notion of justification of religious beliefs in the face of controversy. Having a religious belief consists of taking the content of the relevant proposition as true. As such, believers are justified if they have reasons in support of the truth of the proposition. Evidence works epistemically here: believers access a body of evidence containing *pro* and *contra* items to their belief, assess the weight of each element within the body of evidence to which they have access, and, finally, proportionate their evaluation of the truth of the proposition to evidence. Such evaluation of evidence is computational. According to the two main opposing parties in the epistemology of religious disagreement (i.e., the steadfast view and conciliationism), if (you hold that) your reasons are evidentially much stronger than those of your adversary, you do not have any epistemic obligation to revise your belief. Problems arise about whether the mere fact of peer disagree-

ment counts as contrary evidence, and how much weight such a mere fact should be given. While steadfast theorists think that, once certain epistemic requirements are met, there is no obligation to proportionate evidence in response to contrary beliefs (Gellman 2000; Bogardus 2013; Pittard 2014; Choo 2021), conciliationists hold that disagreements ask for such an obligation (Feldmann 2010). The relevant point is that common consensus in the epistemological literature makes little or no appeal to understanding adversary views when a disagreement is fully disclosed, namely, to having a first-person comprehension of why your adversary thinks their reasons are as good as yours.

According to the above assumptions, religious beliefs are ordinary beliefs concerning the doxastic field of religious ideas, and religious diversity is a relation between the doctrinal systems of different traditions. Any approach relying on similar stipulations about religious affiliation and justificatory processes for beliefs is merely epistemic because it assumes that religious propositions have a semantic content easy to grasp, are unambiguously shared across believers, and, at least in principle, are subject to ordinary manners of evaluation. Consequently, the evidential support relation between religious propositions and facts consists simply of listing which facts (would) make the proposition true: whoever is epistemically positioned in the right way (i.e. whoever accesses all the relevant pieces of information about the content of the proposition) reasons simply by assessing whether the content of the proposition correctly represents facts. Moreover, whoever is in a so happy epistemic position achieves the same conclusion as any other epistemic agent standing in the same epistemic position.

However, this approach to religious affiliation and justification of religious beliefs seems to be problematic from both empirical and normative viewpoints. First, traditions are (among other things) epistemic fields wherein a plurality of doxastic groups have relations of different kinds. Thinking about religions as homogeneous conceptual objects is unsupported by empirical evidence (Bertini 2019). Second, things do not change if the analysis focuses on the denominations of a tradition as objects of inquiry. Picking out a doxastic group is a matter of the fineness of analysis (call a *religious doxastic group* any group identified in terms of the acceptance of a set of seminal claims; Potter 2013; Bertini 2021). Denominations have internal diversity too (e.g., think about denominational divisions in the interpretation of Vedanta within Vedantin classical schools of philosophy). Third, several lively disagreements between denominations do not express normative disagreements relating to points of doctrine. For example, it is a common fact that people from different denominations may have the same fundamental beliefs. Consider the Christian debate concerning the Trinity: there are both Latin and Social Trinitarians within any Christian denominations. As a consequence, assenting to one or the other interpretation does not have any normative consequence for the religious affiliation to one or

the other denomination. Fourth, traditions are historical facts. Long-standing historical facts. Interpretive efforts of framework beliefs have accumulated an enormous quantity of reasons, arguments, and ideas in the passing of time: accordingly, no human being can have real access to full evidence (Bertini 2022). This seems to imply that any justificatory process seems mainly anecdotal, at least *prima facie*. By the notion of anecdoticity, I mean to characterise (in line with the common understanding of the term) that justification is based on personal reasoning and subjective experiences, random investigation, particular access to literature and case studies, and unsystematic appreciation of evidence.

My proposal consists of highlighting how the epistemic evaluation of opposing religious propositions is the outcome of an interpersonal process of evidence sharing wherein the particularities of the involved individuals matter. Each of these individuals has an anecdotal understanding of the content of their beliefs. It follows that their reaction to religious diversity basically depends on such anecdotal features. In my view, there is a plain conclusion to draw: religious diversity is not a (logical) relation between the doctrinal systems of different traditions (or denominations); rather, it is an (interpersonal) relation between individuals. Saying that religious diversity is a relational fact means to express that concrete individuals are engaged in a mutual process of epistemic comparison. Saying that it is a doxastic fact means to refer to the anecdotal nature of such a process. The point I'm making is not that the purely epistemic approach ignores tout court the particular epistemic situations of any two disagreeing individuals, of course. All the work made in the epistemology of disagreement literature concerns how to handle such epistemic situations and discriminate whether sound epistemic requirements are met to evaluate contrary evidence. Rather, my point is that such epistemic work on evidence comes after the construal of the meaning of the beliefs on which disagreements arise. Disagreements between beliefs are not given apriori in the opposition of verbal phrasing of the propositions involved. The interpersonal relation between individuals is the fact that originates the meaning of the disagreeing propositions. This is the main reason why my relational and doxastic approach to religious diversity contrasts the purely epistemic one. The epistemic approach brackets the relational and doxastic features of religious diversity and jumps to an idealized epistemic evaluation: comparing opposing propositions consists of assessing their semantic value according to the assumption that their literal phrasing explicitly expresses their content. Such evaluation is idealised because it stipulates that the content of belief can be captured from its propositional literal expression. On the contrary, my proposal argues for the claim that the required epistemic evaluation should be performed by starting from the relational and doxastic features of the relation.

3. *Issues with* Aprioricity *within the Purely Epistemic Approach*

Religious diversity is an endogenous phenomenon of religion. Any historical revelation is set forth within a context of competing religious worldviews. It is a common pattern of canonical texts to refer to the preaching of a doxastic leader, whose main concern is to defend their view in the face of alternative understandings of the divine. The existence of a plurality of traditions is always primitively assumed as a matter of fact. As such, religious diversity is a constitutive feature of religious belief and demands an epistemic answer (among other attitudes to the challenges that it poses).

From a sociological perspective, the contemporary way of experiencing religious diversity is qualified by the notion of secularity. Grossly speaking, the traditional embedding of religion in a state, which makes the pair with the unchallengeability of belief in the religious target of the institutionalised religion, has progressively eroded from inside during the last centuries, to the extent that the process left open the room for a plurality of competing beliefs (Taylor 2007). As a result, due also to the particular multicultural setting of modern societies, religious beliefs and unbeliefs are fragmented across the epistemic field within any doxastic context.

Such a situation has promoted the emergence of novel approaches to the study of religious diversity (Warner 1993, 2008; Woodhead 2009) which give up the traditional *church model* (i.e., religious experience is embodied into an institutionalised community at a national level of analysis) for *a congregational model* (i.e., religious experience is embodied in congregations mutually related within a religious marketplace): the qualifying feature of religious experience within nowadays societies is the constitutive enjoyment of relational frameworks, both from a theoretical and a practical viewpoint. The substantive consequence of such social facts is that individuals engage in personal ways to experience their religious commitments and life (Cipriani 2009).[2]

In order to provide conceptual distinctions useful to thinking epistemically about religious diversity, I will now introduce a tentative taxonomy for the varieties of the relation. Suppose adherents to different religious doxastic groups meet and debate over the truth of a proposition P. P is a seminal claim for one of the groups, but is denied by others. For example, P is the claim that there is only one divine entity that can be predicated of being God. In such a case, strict monotheists as Jews and Muslims disagree *prima facie* with Trinity monotheists like Christians or openly non-monotheists as Brahmanic Hindus (i.e., mainstream Hindus following Brahmanic traditions in rituals endors-

---

[2] For reasons of space, I cannot discuss in more details these relevant issues. I addressed the issue of how to approach the sociological study of religion in a chapter-length discussion in Bertini (2016). Further, I provided an extensive discussion of how to account for what religions are in Bertini (2019).

ing the pluralistic pantheon of texts such as the Vedas). This provides an instance of *global epistemic* religious diversity. A further example of disagreement is the following (this is the *interdenominational* variety): let P be the claim that a transcendent awareness without content is the only existing reality, and has no proper parts. While Advaita Vedantins accept P, Vedantins adhering to the Viśiṣṭādvaita Vedanta school deny it. Finally (here is the *intradenominational epistemic* disagreement), P is the claim that the Trinity of God should be accounted for by starting from the notion of onefoldness. Latin Trinitarians and Social Trinitarians oppose by reason of their epistemic reactions to P.

Although scholars differently construe, categorise, and address religious diversity, there is a wide consensus that such diversity (whichever way it is understood) is an indisputable matter of fact (McKim 2012). Particularly, the consensus view seems to be that, even if you are sensitive to the fact that religious traditions cannot be reduced to their belief systems (e.g., denominational diversity in religion consists often in a difference in rituals and governmental institutions, the acceptance of the same set of seminal claims notwithstanding), the inner core of their diversity is always to be exemplified in terms of differing beliefs (independently of that they address epistemic or ritual and institutional matters).

Now, I do not question the assumption that beliefs are expressive of how religious individuals understand their faith. I accept that differences in things other than beliefs (e.g., rituals, historical contexts, governmental institutions), are at least conceptually represented by differences in beliefs. However, I have a few doubts that, when one is reasoning about religious disagreements, this suffices to state that different religious doxastic groups generate, develop, and endorse incommensurable religious worldviews.

My perplexities can be motivated by the following reasons:

A. The purely epistemic approach overestimates the relevance of global religious diversity in comparison to the other ones. As a consequence, it models interdenominational and intradenominational epistemic diversity on features of the global variety. This being the case, disagreements about couples of incompatible beliefs are understood as expressive of opposing viewpoints on basic matters. However, this is hardly the case in religious diversity. Often, individuals in such a relation disagree on local issues within a context of shared overlapping assumptions (e.g., most examples of interdenominational religious diversity can be traced back to this pattern). Further, beliefs stating different claims are in a lot of cases expressing different nuances of the same content (e.g., Jews, Christians, and Muslims characterise their God in opposing manners, nonetheless they all qualify as monotheists, and ground their faith on different interpretations of the same religious narratives). The moral is that people adhering to the same worldview (i.e., assenting to the same block of framework propositions for a

given domain of knowledge (Malcolm 2000)) may heavily disagree over grounding topics notwithstanding their shared assumptions of basic beliefs. As such, interdenominational epistemic diversity, for instance, is hardly a matter concerning incompatible worldviews (e.g., the evidence for this claim is that the victims of religious intolerance are often affiliated to the same religion as the persecutors).

B. The purely epistemic approach deals with what individuals believe in terms of their adherence to a religious doxastic group. A Jew is representative of Judaism, a Christian of Christianity, and so on. The concrete difference between individuals vanishes, and changes into a diversity of token and type: each individual adhering to a tradition is an ideal exemplar of a certain faith and differs from an individual adhering to another faith in reason of the difference of types of faith to which they adhere. Nonetheless, that co-religionists assume univocally the same set of seminal claims is controversial (Bertini 2020): the history of any religious tradition shows that religions are doxastic battlefields wherein debates never come to an end (e.g., theologians and scholars in doctrine usually disagree on a high number of seminal topics).

C. The purely epistemic approach tends to understand the meaning of a belief in literal terms (Pouivet 2013), contrary to the massive historical evidence that religious thinkers within any tradition have constantly defended the view that religious beliefs are not to be assumed literally (several Revelations consist properly in novel interpretations of previous religious materials).

The problem addressed by (A), (B), and (C) is that the purely epistemic approach qualifies religious diversity as an a priori matter, that is, a logical fact relying on the relation of compatibility and incompatibility among beliefs. For example, if you believe that *there is only one divine entity that can be predicated of being God*, whoever accepts that there are three or more divine entities that can be predicated of being God, *prima facie* denies your belief; if you believe that *a transcendent awareness without content is the only existing reality*, whoever accepts that a transcendent awareness of something is the only existing reality, *prima facie* denies your belief; finally, if you believe that *the Trinity of God should be accounted for by starting from the notion of onefoldness*, whoever holds that the Trinity of God should be accounted for by starting from the notion of threefoldness, *prima facie* denies your belief. What is common to all these cases is that accepting a proposition in terms of its literal meaning is not compatible with accepting a proposition whose literal meaning either is or involves the negation of the literal meaning of the first. An alethic relation of exclusion among the propositions occurs: they cannot be both believed at once because what is stated by one of them is denied by the other.

However, such understanding of diversity in religion collides with an unsurpassable difficulty: religious beliefs do not have a clear literal meaning, that is, they are required to be semantically analysed to express informative contents (I name *semantic analysis* the epis-

temic procedure of attributing meaning to a proposition).[3] In order to account for this claim (which admittedly may sound strongly controversial), consider the notions of *God*, *divine entity*, and *transcendent awareness*, which occur in the contested propositions. Each of these has an ordinary manner of being employed in non-scholarly language (e.g., something in the neighbourhood of the vocabulary definitions). This can be equated with the *prima facie* understanding of such notions. Nonetheless, experts in religion within any tradition and denomination usually debate about the correctness of such ordinary readings of the notions, aim at reforming them in light of more adequate and legitimate interpretations of the relevant revelation, and, unavoidably, establish different and incompatible characterisations of them.[4] For example, soteriological degree pluralists such as Karl Rahner or Shahid Mutahhari assented to orthodox statements of faith and accepted their literal utterance, although they gave them a meaning differing from that which fundamentalists or exclusivists accept (Legenhausen 2013).

This being the case, it is completely a contingent matter whether two religious individuals making explicit the meaning of their beliefs will agree or disagree. Since it cannot be established a priori which interpretation of the relevant notions involved they accept, it should not be inferred from the *prima facie* reading of their belief that a conceptual difference occurs here (the contrary holds too: two individuals assenting to the same claim may mean different things). The conclusion follows: the notion of religious diversity in terms of the logical relations of compatibility and incompatibility among beliefs should not be accepted as a given. It can turn out that such a notion is too raw to have any theoretical virtue of capturing facts about religious disagreements.

---

[3] My point is that, within the religious field, the relation between meanings and linguistic propositions is not a one-to-one function, primitively stipulated by literality. Meaning supervenes over linguistic proposition, in such a way that any proposition instantiates a plurality of divergent meanings. Literal readings are possibilities in the face of several other ones. My claim is then not that no literal reading of religious beliefs is meaningful, but that literal readings of religious beliefs are assessed as any other readings in terms of a semantic analysis. This implies that there are no primacy relations at all between literal and nonliteral readings of religious beliefs.

[4] The history of religions provides plenty of anecdotal evidence for my claim. Consider the exemplary case of the establishment of the legitimate interpretation of Christianity during the Council of Ephesus (431 AC). The Bishops attacked those Christians who confessed the Nicene Creed assuming an incorrect interpretation of it. The problem was that "some pretend to confess and accept the Nicene Creed, while at the same time distorting the force of its expressions to their own opinion" (The Definition of Faith). Although the writers of the text understand correctness and incorrectness in terms of true and false readings, according to standards of the age, it is clear from the amount of literature which they quote, that the evidence of semantic indeterminacy of religious beliefs was perfectly in their view (Bertini 2018). A few years will pass, and the exegetical schools of spiritual readings of biblical inconsistent passages will flourish, and will establish nonliteral interpretations as common rules for making sense of ambiguity of beliefs.

4. *A Thought Experiment in Support of the Relational and Doxastic Approach to Religious Diversity*

Regardless of whether an epistemic religious disagreement over a proposition P is global, interdenominational or intradenominational, and regardless of how such categories should be qualified, disagreements over a religious proposition begin with the awareness that the contested beliefs challenge each other. There is an acknowledgement here: individuals become doxastic opponents when they are ready to recognise that the rival's view provides something important to think about, because it offers evidence that someone satisfies their spiritual needs in a manner that is oblique, diverging, and far away from that of the other. Such a claim means to point out that, contrary to the received approach by which disagreements provide evidence to be managed on the sole basis of the epistemic contents of opposing beliefs, acknowledging that a belief challenges one's views is not an apriori matter; rather, it is the result of a relational process engaging disagreeing individuals (Bertini 2021). The assumption is that religious beliefs propositionally represent, publicly establish, and intersubjectively give voice to how individuals understand their religious life. Accordingly, religious diversity has to do with a challenge towards one's commitments: individuals understand that they enjoy different religious experiences; the representation by beliefs of such a difference questions their own experiences; and, consequently, people react to the fact that others think and live religious matters differently.

4.1. *Discovering Religious Diversity*

I will introduce a thought experiment to develop a few considerations on the epistemology of religious diversity. A methodological statement is required. My story is a hypothetical epistemic situation, which provides an idealised doxastic comparison between peers (although I hope that the idealisation is a sufficiently rich scenario for permitting the flow of substantive informative content). Naturally, the story is not an empirical one. I do not assume that individuals experiencing religious diversity behave as my characters do, nor that the outcomes of empirical research on actual phenomena of inter-religious relations cohere with my scenario. As usual in epistemology, my work focuses on a "what-if" situation, namely, what would be the rational answers to an epistemic issue if a given situation occurs. Consequently, my main concern is with plausibility: if my scenario is a real possibility, the story can highlight which legitimate epistemic attitudes are appropriate. The focus is on the notion of peerhood. Elsewhere I suggested reasons against a few mainstream assumptions on what it is to be epistemic peers (Bertini 2021) by setting forth doubts in reason of their abstractness and aprioristic commitments. My claim is that the richness of particularities on which the story relies can point conclusively at the ne-

cessity of quitting apriori reasoning in religious diversity and set forth the way for an empirical understanding of what is epistemic peerhood.

Saul, John, and Mohammed are friends. Saul is a Jew, John a Christian, and Mohammed a Muslim. Sometimes they speak about religion. When they do, they are contented with thinking that they are all monotheists. Possibly, they hold that they all believe in the same God: differences in their doctrines are simply due to differences in the social context wherein their revelation occurred and in the history that has developed from then on. However, they accept that such differences play a role in their religious affiliation. Each of them thinks that while the religion of the others may provide them with salvation, their revelation is more fit to achieving such a condition because of some epistemic virtues. Inclusivist readings of Judaism, Christianity, and Islam (which have developed throughout the history of these traditions) show that my hypothesis has empirical evidence in support (Race 1986; Cohn-Sherbok 1996; Khalil 2016). In this situation, John, Saul, and Mohammed do not think that they experience a real difference in religious matters. Naturally, they are aware that their religious commitments are different; however, they understand such diversity as something relating to the surface of the issue. If they go in-depth, they cannot find reasons for believing that they are different from a religious standpoint: they have faith in the same God, they apply similar ethical codes (Küng 2002), they evaluate what is relevant and what is not by analogous criteria, and so on. In this scenario, none of them appears to be challenged by the religious beliefs of the others. This does not imply epistemic apathy; rather, each of the three friends is ready to acknowledge a degree of truth to the views of the others. On one side, they epistemically behave in a kind of pragmatical pluralism, on the other they hold that their disagreement is not a crucial fact challenging their faiths. However, things suddenly change one day. That evening, Pasolini's movie *The Gospel According to St. Matthew* was scheduled at their favourite movie theater. None of the three friends ever saw the masterpiece. For this reason, John asked the others whether they would like to watch it with him. Naturally, they were positive. During the show, both Saul and Mohammed remained astonished. They were both fascinated by how Jesus appeared to their eyes: he was a real, perfect, and sublime man and the natural object of an unstoppable love. Nonetheless, Saul felt also disappointed by how Hebrew authorities are painted in the movie, and, accordingly, began to develop contrasting feelings about whether Jesus was so fascinating. Mohammed, for his part, had something like an epiphany: for the first time, he understood why his Christian friend was so attracted by a God who became a man. After the show, the three friends engaged in a fueled conversation about who Jesus was and established their different beliefs over the topic. From that day on, they have sunk into the experience of their religious diversity: their different ideas on Jesus

represent the different perspectives on which they focus in living their religiosity. What changed from the former to the latter situation is that now the three friends cannot confine differing beliefs on the surface: Saul and Mohammed have felt how appealing Christianity may be, and perfectly understand that it is a challenge to their faith. John has not stood firm either: while his friends defended their views, he reacted to such apologies, and became familiar with the attractiveness of their ideas. The direct experience of how the faith of others challenges their own grounds a new understanding of what their beliefs represent, and generates an epistemic reaction.

4.2. *Epistemic Anxiety*

The basic form of religious dissent can be represented in a propositional manner. Saul claims that Jesus was a saint, but he denies that he showed any supernatural feature in his life; John claims that Jesus was the Son of God, namely, the incarnation of the second divine Person of the Trinity; Mohammed claims that Jesus was the last prophet before the revelation of Islam, and that, although he was not the Son of God, a supernatural relation to God was the substance of his life.

However, such a propositional approach sets aside the essential of the three friends' thought experiment: they acquired awareness of their theoretical disagreements by having personally enjoyed a lively experience of how much the beliefs of others challenge their own. Consequently, the matter at issue does not simply consist of evaluating which of their belief is the more probable, given their evidence; rather, because of their uncertainty towards beliefs which they are not willing to give up, they are pushed to determine definitively how things stand.

The crucial point is that they experience a kind of anxiety. From a psychological viewpoint, anxiety is an emotional state caused by uncertainty. Basically, anxiety is a response to a threat flowing from an internal conflict generated from a plurality of incompatible imaginations of future scenarios. It is a modal psychological state of mind; something like randomly accessing different possible worlds wherein challenging states of affairs occur, and being incapable of managing such a rich realm of different and incompatible possibilities in the meantime (Steimer 2002). Now, when disagreements are real, namely, when they are first-person experiences as those enjoyed by John, Saul, and Mohammed, disagreements are someway analogous with anxious states of mind. Indeed, an undetermined uneasy psychological state related to the incapability of silencing the challenge set forth by the beliefs of others flows from disagreements and makes dissenters unable to react properly to such a plurality of possible realisations, similar to what happens when anxiety is experienced in front of different and torturing possibilities. To a certain respect, disagreements externalise what anxiety internally is to one single mind. Participants in a disagreement situation make it possible for others to access those alternative

scenarios which hold for any of them. If such participants acknowledge each other's epistemic peerhood, any of them show the reasons why their epistemic situation is attractive. They invite others to take part in a lively possibility. As a consequence, epistemic anxiety acquires an overriding epistemic value. My favorite way to characterise such value consists in focusing on its normative nature: whenever individuals disagree on grounding topics because of experiential access to their religious diversity, their anxiety towards the matter should incline them to pursue indefinite inquiry on the issue. This means that epistemic anxiety plays the role of a fuel of any proper process of evidence evaluation between disagreeing religious individuals (who personally experienced their religious diversity).[5]

4.3. *The Relational and Doxastic Approach to Religious Diversity*

In my view, the following are the main facts attested by the three friends thought experiment. First, Saul, Paul, and Mohammed cannot simply evade, ignore, or forget their doxastic opposition, because they had the experience of how much the beliefs of others challenge their own, and they have mutual esteem for their ability to draw correct evaluations concerning several important things, among which there is religion (remember that before the acknowledgement of their diversity, each of the three had a good opinion of the religiosity of the others). Such impossibility of bracketing the awareness that their beliefs are challenged is constituted by the reactive generation of religious diversity: they have a certainty that their beliefs are threatened because they enjoyed a relevant experience (call this feature *the experiential origin of epistemic anxiety*).

Second, the focus of their debate is not scholarly. They are not intellectuals having a professional exchange of opinions. Rather, they are sincere religious individuals who want to be correct in their views, because they experienced a strong epistemic anxiety concerning the soundness of their beliefs. Particularly, they are not actualising global instances of religious diversity: their issue originates from a particular topic (i.e., the religious meaning of the life of Jesus) which is

[5] Quantitative sociological research attests that a common reaction to disagreements is steadfastness free from anxiety. Such a result depends on that doxastic opponents are not peers in most ordinary epistemic encounters. Imagine introducing in the scenario a new epistemic agent claiming that Jesus is an alien. Naturally enough, there would be no reason to think that such a claim would provoke anxiety in the three friends. The reason is that such a claim will qualify the new epistemic agent as a non-peer. Epistemic anxiety is the outcome of epistemic peerhood. Quantitative inquiries manifest steadfastness free from anxiety in interreligious relations because they concern non-peerhood relations. On the contrary, literature on relevant epistemic debates in religious matters, such as *Scriptural Reasoning* or historical cases of authentic religious debate for instance, provides *prima facie* evidence that, when doxastic opponents deal with each other as peers, the epistemic process of evidence sharing is moved on by attitudes which are a variety of what I taxonomise as epistemic anxiety.

interpreted by partially overlapping views. As such, the issue cannot be solved by means of a logical analysis of the doctrinal body of three mutually excluding religions. Rather, the focus of their inquiry is the way competing different understandings of Jesus's life substantiate the concrete religious lives of each of the three friends, and how much each of these interpretations is attractive to any of the three friends (this is *the interpersonal nature of religious diversity*).

Third, they do not have any arguments capable of persuading others to change their view. This is a very frustrating thing: they would like to confirm from a cognitive viewpoint what they experientially live every day, but they cannot, because the others do not give their assent to the claims which are so essential to each of them (label this fact *the undecidability of religious disagreements*).

Fourth, even worse is that none of them assumes to be in a better position than the others. Possibly, after a few debates, they can rationalise their dissent (I use the term rationalise according to its use in psychology and sociology), and can find rest to their anxiety by deluding themselves to have a privileged access to evidence. However, although some scholars defend such an answer (notably, Plantinga 2000), this position is not supported by strong reasons: it is a *petitio principi*, where it is assumed that a side in a disagreement is right because either it has better access to evidence or evaluates the available evidence in the correct manner (the disjunction begs the question because it assumes what should be proved). I am not claiming that debates do not ordinarily go this way (e.g., van Inwagen 1996). Naturally, from time to time parties in a disagreement presume to be right because their position is their own. Nonetheless, even if such a move may also have a descriptive value, it is unmotivated from a prescriptive standpoint (call this *epistemic symmetry between the disagreeing parties*). Naturally, steadfast theorists relying on partisan reasoning as a basic ground for not conciliating do not think that standing firm is an act of rationalised delusion motivated by the need to stop the flow of epistemic anxiety. Their point is that assenting to framework propositions stipulating a determinate epistemic viewpoint makes an individual able to appreciate precisely that evidence in support of their claims, and that, in the absence of conclusive reasons against such evidence, there is no reason to dismiss a belief which could be the right one. Nonetheless, steadfast theorists epistemically suggest in this way to behave in the face of intractable disagreements as if rationalisations of dissent is the good option: you accept your belief in terms of your having access to a given epistemic situation; you cannot be persuaded that contrary evidence is defeating your belief because you do not access the epistemic situation of your opponents; you and your opponent have no reason to dismiss your beliefs because none of you has the access to the evidence that is at the disposal of the other; the evidence is there, but you evaluate such evidence non challenging for reasons which are not relating to the

content of the controversy – simply because it is not your own, it is not what you can appreciate given your epistemic standpoint (van Inwagen 2010).

Fifth, epistemic symmetry is anecdotally experienced by religious individuals involved in a disagreement. That is, since each of the three friends anecdotally accesses evidence, none of the three friends is a prototypical exemplar of the tradition to which he adheres. Their epistemic sharing of evidence concerns particular facts, which have a particular justification.

4.4. *Which Answer to Religious Diversity*

The epistemological debate on disagreement has been largely shaped by the use of the uniqueness principle in support of conciliationism and permissivism in support of the steadfast view (Christensen 2009). Such bifurcation can certainly be relaxed in both camps. Conciliationists may argue for their preferred option even in case permissiveness is assumed (Christensen 2007). Steadfast theorists may refer to permissiveness only as a supplemental reason for standing firm (Frances and Matheson 2024). Nonetheless, relating uniqueness to conciliationism and permissiveness to steadfastness is somewhat constitutive in the epistemology of religious disagreements. Therefore, I will develop my following considerations by such assumption. Suppose that you are a friend of the view that a certain situation supports one epistemic answer alone. This being the case, whenever a disagreement occurs, there is only one correct position to hold. Your opponent and you are both epistemically competent, have comparable access to evidence, and cannot arrive at a decision over which belief is the sound one. Consequently, you should conciliate, namely, you should withhold from assuming a view and suspend your judgement (Feldman 2006; Elga 2007; Kornblith 2010). On the contrary, suppose that you accept that more than one justified position is permitted by one and the same situation. You can then answer the disagreement by standing firm in your belief: it may be that the doxastic opposition does not concern commensurable beliefs (Pittard 2014; Choo 2021), that being spineless does not favour any epistemic benefit and success (Elgin 2010), or that conciliating bootstraps the opposite view (Aikin, Harbour, Neufeld and Talisse 2010). In any case, the steadfast view suggests being resolute towards your beliefs.

Apply all this to the disagreement between Saul, John, and Mohammed. When their religious diversity was not in focus, they considered each other equally good religious individuals. Since their interest in religion is not scholarly but practical, from their mutual acknowledgment that they are equally good religious individuals, it follows that each of them attributes a sufficient degree of capabilities in religious matters to the others. As a consequence, none of them can assume that

the others *prima facie* assess their evidence in the wrong way. Now, the core belief generating their disagreement concerns the nature of Jesus. They started sharing their evidence on such a topic after the movie show. They have interacted longly (and their debate is still going on) over how their different religious lives depend on their understanding of the nature of Jesus. The different characterisation of him is a representation of their difference in faith. Therefore, they have at present comparable access to the body of evidence. Nonetheless, the belief about the nature of Jesus is undecidable to them, that is, there are no conclusive arguments at their disposal to which John, Paul and Mohammed assent.

Conciliation is not a possibility here, contrary to the intuition that it could be one. If your opponent and you are equally good in epistemic matters, and your shared evidence is not sufficient to decide which of a couple of contrary beliefs is the true one, it seems rational to conclude that the evidence at stake is not good evidence in support of either conclusions. Now, John, Paul, and Mohammed are in a relation of epistemic symmetry to evidence (each of them accesses the same evidence with a comparable competency, and they nonetheless cannot provide definitive arguments in support of their view in the face of the objections of others), and the contested beliefs are undecidable. However, they cannot withhold their beliefs, due to their experiential access to their disagreement, and the generation of epistemic anxiety which follows it. Particularly, contrary to the purely epistemic approach to conciliation assuming that undecidability depends apriori on the paucity of evidence in support, the relational and doxastic features of religious diversity show that the paucity of evidence is not an empirical character of the disagreements generated by religious diversity.

When each of the three friends experiences that the viewpoints of the others radically challenge theirs, they do not indeed access the epistemic fact that they have no good evidence in support of their claims. Rather, each of the three friends *experiences* that others have as good evidence as their own. This is the shocking outcome of their incapability of coming to a shared belief over Jesus. That is, the situation is not about the paucity of evidence in support. Religious disagreements such as John's, Paul's, and Mohammed's, are not an instance of a Quinean scenario wherein beliefs are evidentially underdetermined (Quine 1959). On the contrary, each of the three friends shows to the others that his way of understanding Jesus' life is compelling, challenging, and adequately supported by good arguments and good religious reasons. The problem is then that there is too much good evidence, namely, the abundance of evidence overdetermines what the three beliefs about Jesus represent. When the three friends use the term Jesus, the sense by which they refer to Jesus is not scarcely determined; rather, there are many over-detailed and competing senses representing Jesus in incompatible manners, although all of these are working. Consequently, given that Feldman's, Elga's and Kornblith's reasons for

conciliation rely on the paucity of evidence, their prescriptive answer does not apply here.

A further reason for the inapplicability of the conciliationist prescription relies on the normative nature of epistemic anxiety: doxastic opponents feel anxiety because they are required to answer the disagreement to hold their beliefs, but they cannot do it. Anxiety is a mark of something going wrong with the support relation between evidence and belief. Particularly, while the contested beliefs are mandatory to each of the three friends, the experiential access to the fact that others reasonably understand Jesus may mean that different manners to refer to something could provide alternative representational contents cohering with the same body of evidence. The suggestion is that if the interpersonal nature of religious disagreements generates epistemic anxiety, given that such anxiety arises from the experiential access to the reasonableness of the views of others, the interpersonal nature of religious disagreements is an experientially grounded reason for the claim that a plurality of replies to evidence are epistemically permitted.

Unfortunately, the resolute option is not more promising than the conciliatory. The motivations for standing firm are indeed defeated by the following reasons. First, John, Paul, and Mohammed cannot think that the views of the others are not commensurable to their own. If they thought this way, they would have not experienced epistemic anxiety. But, if they did not experience epistemic anxiety, they could have no reason to think that more than one response to evidence is epistemically permitted. Therefore: either they do not feel anxiety, and, accordingly, they could conciliate, or they do feel anxiety, and, accordingly, they should not stand firm because of the incommensurability of their beliefs. As in the previous case, the purely epistemic approach to steadfastness cannot provide a good prediction of how real disagreements behave.

Second, I may argue that conciliationism should be opposed because I hold that disagreements are bearers of epistemic benefits (Elgin 2010; Dormadandy 2020; Bertini 2021). Nonetheless, if this is my strategy, I cannot endorse steadfastness either. Disagreements favour epistemic improvements of beliefs because they push doxastic opponents to re-evaluate their first-order evidence and to include objections and replies to objections within their body of evidence. However, when this strategy is pursued, the doxastic opponents decide not to stand firm, but to check again whether they have good reasons for their belief. Such a move obviously requires that each party in the disagreement is ready to learn from the rival one. It would be odd if someone said: *I do not agree with you, I am right and I do not move from my viewpoint; in any case, let me see whether your reasons are good.* I cannot see how it can be that my opponent is not joking, and they are willing to learn from me and accept that they could be wrong and I can be right. The relational and doxastic features of the empirically informed view of

religious diversity cohere with the way a disagreement may promote epistemic benefits to both sides of a dispute.

Third, it is true that unilateral conciliation may cause the undesired consequence that the steady party of a disagreement would find additional reason for standing firm: if someone conciliates in the face of a disagreement, the epistemic opponent refusing to conciliate seems to be rational in taking their evidence stronger than those in support to the other side, given that the other gave up their view. However, this fact does not constitute a real reason for steadfastness. It would be one if either conciliationism or standing firm were the only epistemic options. What really happens when individuals discover their religious diversity shows that a third view is on the marketplace, namely, engaging in an epistemic debate on the truth of the matter.[6]

5. *Conclusions*

The main consequence of my way of modeling religious diversity is that apriori normative prescriptions towards how to manage religious disagreements should be set aside in favour of pursuing an epistemic process of evidence sharing. The dichotomy consists in that the former assumes that opposing beliefs reveal evidentially their contrast by simply stating the propositions at stake, and the latter asks for interpersonal in-depth work on meaning. That is to say, if my description of how religious individuals disagree is empirically more on target than the purely epistemic approach, suggestions to conciliation or holding firm springing from apriori reasoning about the logical properties of incompatible beliefs should be evaluated unjustified in providing a reasonable answer.

According to my proposal, although religious disagreements naturally concern the truth-value of the involved propositions, namely, they are ultimately epistemic disputes about which of the contested propositions is the adequate representation of the matter at issue, such disagreements have relational and doxastic features which cannot be left on the side. This means that religious diversity is not a global affair

[6] Notoriously, the Rawls-Habermas debate addresses issues in the neighbourhood of the hoped practical outcomes of my work (Kedziora 2019). Both working within a neo-Kantian setting, the two thinkers analise and criticise each other on how to manage the fragmentation of beliefs in contemporary societies. While my epistemic approach is obliquous to the socio-political focus of their debate, my considerations can be found in line with Habermas' (and Apel's) inspiring principles towards the construal of a theoretical ethics of discourse. Particularly, translating ideally generalised doxastic comparisons between groups into empirical cases of actual relationships between individuals is a way to fill in relations of epistemic diversity with contents. This also answers a possible objection to my approach, namely, the difficulty of generating inductive results in religious studies of diversity. Good epistemological practices in the face of diversity suggest focusing on empirical particularities and fixing controversies in mutual knowledge. Qualitative research is the main asset to give voice to my proposal.
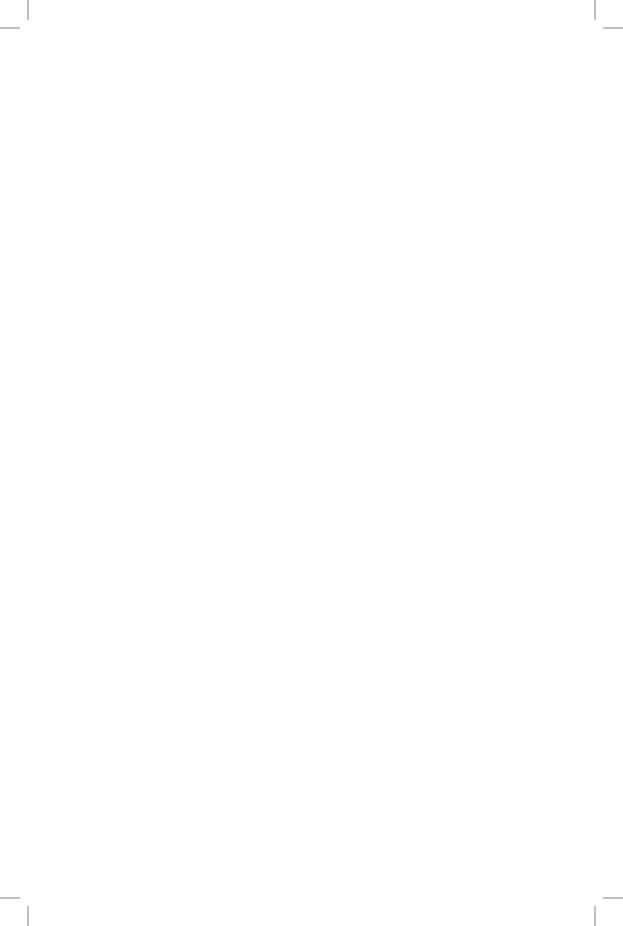
relating to the doctrinal body of competing traditions. Rather, it is a reactive awareness by means of which concrete individuals accept that their particular (i.e., anecdotal) views are challenged, and, accordingly, start engaging in a process of evidence sharing. This process is motivated by the willingness to make an experience of the opponent's standing point. As a consequence, you cannot simply bear that someone thinks differently from how you do, and remain untouched by this fact: you should act responsibly to your belief, and begin exploring the in-depth binds of an interpersonal confrontation.

References

Aikin, S. F., Harbour, M., Neufeld, J., and Talisse, R. B. 2010. "Epistemic Abstainers, Epistemic Martyrs, and Epistemic Converts." *Logos & Episteme* 1 (2): 211–219.

Alston, W. P. 1992. *Perceiving God: The Epistemology of Religious Experience* (Reprint edition). Ithaca, NY: Cornell University Press.

Bertini, D. 2016. *Tradizioni religiose e diversità*. Verona: Edizioni Fondazione Centro Studi Campostrini.

Bertini, D. 2018. "The Anecdotal Nature of Religious Disagreements." *International Journal of Philosophy and Theology* 80 (3): 215–229.

Bertini, D. 2019. "On What a Religion Is Not." *Religions* 10 (1): 29.

Bertini, D. 2020. "The Vagueness of Religious Beliefs." *European Journal for Philosophy of Religion* 12 (2): 181–210.

Bertini, D. 2021. "Anecdotal Pluralism, Total Evidence and Religious Diversity." *Philosophia. Philosophical Quarterly of Israel* 49 (1): 155–173.

Bertini, D. 2022. "Anecdotal Pluralism." *Logos & Episteme* 13 (2): 117–142.

Bogardus, T. 2013. "Disagreeing with the (Religious) Skeptic." *International Journal for Philosophy of Religion* 74 (1): 5–17, 19.

Choo, F. 2021. "The Epistemic Significance of Religious Disagreements: Cases of Unconfirmed Superiority Disagreements." *Topoi* 40 (5): 1139–1147.

Christensen, D. 2007. "Epistemology of Disagreement: The Good News." *The Philosophical Review* 116 (2): 187–217.

Christensen, D. 2009. "Disagreement as Evidence: The Epistemology of Controversy." *Philosophy Compass* 4/5: 756–767.

Cipriani, R. 2009. "La religione dei valori diffusi." *La società degli individui* 34 (1): 40–52.

Cohn-Sherbok, D. 1996. "Jewish Religious Inclusivism." *CrossCurrents* 46 (3): 326–342.

Dormandy, K. 2020. "The Epistemic Benefits of Religious Disagreement." *Religious Studies* 56 (3): 390–408.

Elga, A. 2007. "Reflection and Disagreement." *Noûs* 41 (3): 478–502.

Elgin, C. Z. 2010. "Persistent Disagreement." In R. Feldman and T. A. Warfield (eds.). *Disagreement*. Oxford: Oxford University Press, 53–68.

Feldman, R. 2006. "Epistemological Puzzles about Disagreements." In S. Hetherington (ed.). *Epistemology Futures*. Oxford: Oxford University Press, 216–236.

Feldman, R. 2010. "Reasonable Religious Disagreements." In L. M. Anthony (ed.). *Philosophers Without Gods: Meditations on Atheism and the Secular Life*. New York: Oxford University Press, 194–214.

Frances, B. and Matheson, J. 2024. "Disagreement." In E. N. Zalta and U. Nodelman (eds.). *The Stanford Encyclopedia of Philosophy (Winter 2024 Edition)*.

Gellman, J. 2000. "In Defence of a Contented Religious Exclusivism." *Religious Studies* 36 (4): 401–417.

Harrison, V. S. 2006. "Internal Realism and the Problem of Religious Diversity." *Philosophia* 34 (3): 287–301.

Kedziora, K. 2019. "Habermas and Rawls on an Epistemic Status of the Principles of Justice." *Acta Universitatis Lodziensis* 34: 31–46.

Khalil, M. H. 2016. "Islam and the Salvation of Others." In R. McKim (ed.). *Religious Perspectives on Religious Diversity*. Leiden; Boston: Brill, 149–161.

Kornblith, H. 2010. "Belief in the Face of Controversy." In R. Feldman and T. A. Warfield (eds.). *Disagreement*. Oxford: Oxford University Press, 29–52.

Küng, H. 2002. *Wozu Weltethos? Religion und Ethik in Zeiten der Globalisierung*. Freiburg: Herder, Freiburg.

Legenhausen, M. 2013. "Nonreductive Pluralism and Religious Dialogue." In M. H. Khalil (ed.). *Between Heaven and Hell. Islam, Salvation, and the Fate of Others*. Oxford: Oxford University Press, 153–179.

Malcolm, N. 2000. "The Groundlessness of Belief." In B. Davis (ed.). *Philosophy of Religion*. Oxford: Oxford University Press, 115–122.

McKim, R. 2012. *On Religious Diversity* (1 edition.). New York: Oxford University Press.

Pittard, J. 2014. "Conciliationism and Religious Disagreement." In M. Bergman and P. Kain (eds.). *Challenges to Moral and Religious Belief: Disagreement and Evolution*. Oxford: Oxford University Press, 80–97.

Plantinga, A. 1999. "Pluralism: A Defense of Religious Exclusivism." In P. L. Quinn and K. Meeker (eds.). *The Philosophical Challenge of Religious Diversity*. New York: Oxford University Press, 172–192.

Potter, D. 2013. "Religious Disagreement: Internal and External." *International Journal for Philosophy of Religion* 74 (1): 21–31.

Pouivet, R. 2013. *Epistémologie des croyances religieuses*. Paris: CERF.

Quine, W. van O. 1959. "Meaning and Translation." In R. A. Brower (ed.). *On Translation*. Cambridge: Harvard University Press, 148–172.

Race, A. 1986. "Christianity and Other Religions: Is Inclusivism Enough?" *Theology* 89 (729): 178–186.

Steimer, T. 2002. "The Biology of Fear- and Anxiety-related Behaviors." *Dialogues in Clinical Neuroscience* 4 (3): 231–249.

Taylor, C. 2007. *A Secular Age*. Cambridge: Harvard University Press.

van Inwagen, P. 1996. "It Is Wrong, Everywhere, Always, and for Anyone, to Believe Anything upon Insufficient Evidence." In J. Jordan and D. Howard-Snyder (eds.). *Faith, Freedom, and Rationality*. Lanham, Md: Rowman & Littlefield Publishers, 137–153.

van Inwagen, P. 2010. "We're Right. They're Wrong." In R. Feldman and T. A. Warfield (eds.). *Disagreement*. Oxford: Oxford University Press, 10–28.

Warner, S. R. 1993. "Work in Progress Toward a New Paradigm for the Sociological Study of Religion in the USA." *American Journal of Sociology* 98 (5): 1044–1093.

Warner, S. R. 2008. "Parameters of Paradigms: Toward a Specification of the U.S. Religion Market System." *Nordic Journal of Religion and Society* 21 (2): 129–146.

Woodhead, L. 2009. "Old, New, and Emerging Paradigms in the Sociological Study of Religion." *Nordic Journal of Religion and Society* 22 (2): 103–121.

Rejoinder to Wysocki and Dominiak on Blackmail Law and Austrian Economic Welfare Theory

WALTER E. BLOCK
Loyola University New Orleans, New Orleans, USA

Wysocki and Dominiak (2023) defend the thesis that blackmail is only beneficial, relatively, since the target of this offer values his secrecy more than the money he must pay. However, he loses absolutely, since the blackmailee would have been even better off if the blackmailer had disappeared from the scene, and/or never made his demand in the first place. In their view, the blackmailer is thus akin to the highwayman who offers you the choice of your money or your life. You value the latter more highly, so you "gain" by complying, but only in this relative sense. However you lose absolutely, since you would have been even better off had there been no highwayman threatening you, in the first place. This is in sharp contrast to the person who sells you a car. There, you gain in both cases: you value the auto more highly than its price, otherwise you would not make the purchase, and, you would not be better off if he disappeared. The present paper is critical of this thesis.

Keywords: Justice; Austrian economics; economic welfare; libertarianism.

Wysocki and Dominiak (2023, hereafter WD) make an important contribution not only to libertarian theory, but, also, to Austrian as well as mainstream economics. They bring new intellectual energy to the study of, and the interactions between, the freedom philosophy and the praxeological school of thought, as it pertains to welfare economics.

WD start off on the right, well, correct, foot. They state:

> This paper argues—contra some Austro-libertarians—that whether a given exchange is welfare-enhancing or welfare-diminishing does not depend on whether that exchange is just or unjust, respectively. Rather, we suggest that in light of our two thought experiments, Austro-libertarianism has at

least a pro tanto reason to conceive of justice and welfare as two logically distinct ideals.

We don't need any new thought experiments to establish this, however welcome, and clever, are the ones supplied by WD.[1] The fact value distinction, that between positive and normative economics, is already well established within the economics literature. There is all the world of difference between "A causes B" and "A is justified, B is not." However, all too often this distinction is honored more in the breach than in any other way, so, hats off to these authors for reminding us of this important understanding.[2]

However, from this foundation of political economy, WD "… predict the possibility of (a) just but welfare-diminishing exchanges and (b) unjust but welfare-enhancing ones." Here, I fear, they go astray.

In this regard they mention the following quote from Rothbard:

> It so happens that the free-market economy, and the specialization and division of labor it implies, is by far the most productive form of economy known to man, and has been responsible for industrialization and for the modern economy on which civilization has been built. This is a fortunate utilitarian result of the free market, but it is not, to the libertarian, the prime reason for his support of this system. That prime reason is moral and is rooted in the natural-rights defense of private property we have developed above. Even if a society of despotism and systematic invasion of rights could be shown to be more productive than what Adam Smith called 'the system of natural liberty,' the libertarian would support this system. Fortunately, as in so many other areas, the utilitarian and the moral, natural rights and general prosperity, go hand in hand. (Rothbard 2006: 48–49)

Here, I must depart from Rothbard. In my view, it would be logically impossible for "a society of despotism and systematic invasion of rights … to be more productive than" laissez faire capitalism. For under the latter regime, each and every commercial interaction necessarily enhances economic welfare at least in the ex ante sense.[3] This includes buying, selling, trading, bartering, investing, gift giving, etc. In each case, all parties to the transaction expect to be made better off by engaging in it, otherwise they would scarcely agree to do so. Nor is there any other system that can truthfully make this claim. Certainly, when rights are violated, at least one person, the victim, necessarily loses by the interaction.

---

[1] See on this below.

[2] Our authors use the phrase "Austro-libertarians." Strictly speaking, this is a violation of the normative positive distinction, as the former deals with facts, cause and effect, etc., while the latter focuses on rights. However, this phrase is so well-entrenched in the literature, I use it myself, that I can hardly blame them for so doing. My interpretation of this phrase is that it is simply a short hand version of characterizing those who support both Austrian economics and libertarianism. Just because facts are philosophically different from values does not mean we cannot have views on both.

[3] And very often ex post as well, although this is not a necessary condition of the system.

It is puzzling that Rothbard takes this position since his own Unanimity Rule has it that "[w]e can only say that 'social welfare' (or better, 'social utility') has increased due to a change, if no individual is worse off because of the change (and at least one is better off)" (Rothbard 2011: 23).

One possible counterexample is the harm of automation. The computer comes along, and pretty much bankrupts all producers and sellers of typewriters.[4] The response is that these people are no longer part of the market. They are not able, any more, to make an offer other people find attractive. I now offer you, gentle reader, this here pencil for the designed to sell price of one million dollars. What? No takers? Does this mean I am a market participant who is worse off? No. I am not a market participant at all, defined as someone who can find someone else to interact with, commercially.

Another possible counterexample is pollution. A produces steel and sells it to B. But C is harmed by the pollutants A places into the air, and into his lungs. The best refutation of this claim against the market is Rothbard (Rothbard 1982a: 152), who demonstrates that this is a trespass of private property rights, and not an aspect of the free enterprise system at all.

It is clear that WD are onto something important when they write that "(1) the free market always increases social utility and that (2) no governmental intervention can ever increase it. In other words, the above two statements have it that just exchanges are always mutually beneficial and that unjust exchanges can never be welfare enhancing."

Unhappily, their paper, instead, "… makes a positive argument for market inefficiencies and mutually beneficial injustices, and hence for the position that justice and welfare should constitute two independent ideals within the Austro-libertarian framework. This in turn predicts that there can indeed be (a) just but welfare-diminishing exchanges and (b) unjust but welfare-enhancing ones."

We wait with baited breath for these authors to show that "Just exchanges are not necessarily welfare-increasing." They make this attempt starting with the highwayman example, who demands your money at the point of a gun and threatens to kill you if you resist. You pay, since you value your life more than the money.[5] Our authors say of this situation: "… the recipient seems to be rendered worse off when compared to the situation in which the gunman would have nothing to do with the recipient at all."

Nothing to do with? But there are lots of other cases, where justice prevails, and yet there is indeed a loss in utility of this sort. Competition for example. A and B compete for the custom of C. A wins. B loses out.

---

[4] Other examples immediately come to mind regarding the computer, the horse and buggy industry, cameras, etc.

[5] Comedian Jack Benny whose schtick was that he was a miser, when confronted with this choice would whine in response: "I'm thinking, *I'm thinking.*"

Yes, in Nozick's (Nozick 1974: 84) drop dead theory, blackmail is "non-productive." The blackmailee would be better off if the blackmailer didn't exist, or didn't know his secret. B would be better off if A didn't exist. But it is productive when the blackmailee pays for silence.[6]

You would be better off if your neighbor and the big tree on his property that keeps growing and now blocks your view didn't exist. But he, and it, does. So, is it productive when you pay him to shorten the height of his tree so you can have a better view? Sure it is. Yet, according to WD, while this is a just contract, it is not welfare enhancing. I find it difficult to understand why it is not productive, not mutually beneficial. Certainly it is justified for the neighbor to have a big tree, or even a spite fence, on his property.

Take another case. You are dating a girl. She has a younger brother. You would just as soon as he would disappear over the hill, so that you can have more quality time with your lady friend. So you spring for a movie, dinner, popcorn, whatever it takes, for the young lad. This is certainly a productive exchange on your part. You regard his absence as of far more value to you than the cost of the film and some food. Yet, he has done nothing wrong just by being at home while you wish he were away. In the view of WD, again, this is a non-productive exchange, only beneficial in the relative but not the absolute sense, since you would have been even better off were he not present at his home in the first place; then, you would not have to bribe him to disappear for a few hours.

Let us try one more example. You purchase some shoes for $50. You would be much better off if the store owner gave them to you for free, out of the goodness of his heart. In this case, also, as in the previous one, there is nothing untoward going on; all that takes place is legal. But if we follow the logic being put forth by WD, you didn't really absolutely benefit from the sale of the shoes, even though, obviously, you valued them at $60, and thus made a $10 relative profit on the deal. Why not? That is because you would have gained even more welfare had you received the shoes as a free gift.

The parallels with blackmail are obvious. Yes, the blackmailee would have been better off had the blackmailer not unearthed his secret. In actual point of fact, the former pays the latter a given sum because he values the silence of the blackmailer to a greater extent than the cost of purchasing his quiet. Revealing a secret is not a crime. It is part and parcel of free speech. It is to engage in mere gossip.

WD then offer the case of the Car Dealer where both buyer and seller gain from their transaction.[7] These authors characterize this as

[6] For a defense, not of blackmail per se, but of legalizing it, see Block (1972, 1976: 53–58, 1986, 1997, 1998, 1999a, 1999b, 1999c, 1999d, 2000a, 2000b, 2000c, 2000d, 2001a, 2001b, 2002–2003, 2009, 2013), Block and Anderson (2001), Block and Gordon (1985), Block, Kinsella and Hoppe (2000), Block and McGee (1999a, 1999b), Mack (1982), Murphy (2019), Rothbard (1982b: 124–126).

[7] Necessarily in the *ex ante* sense, but only presumably, also, *ex post*.

both an absolute gain (since the buyer would not be better off if the seller disappeared) and also a relative one (since voluntary trade is always mutually beneficial).

Next they wheel in their big guns and apply their insights to blackmail. They state:

> Suppose that a blackmailer makes the following proposal to the blackmailee:
>
> (1) If you pay me \$1.000.000 (demand), I will let your reputation remain untarnished (relative benefit).
>
> (2) If you don't pay me (refusal), I will gossip about your secrets (threat).

First of all, since the threat element promises an action that would not violate the criminal's rights, the blackmailee paying the blackmailer \$1.000.000 would result in a just distribution. To see that, consider the following assessment of justice of blackmail proposals by Block (Block 1999: 124), who has it that in blackmail scenarios "a valuable consideration is demanded, under the threat of doing something entirely licit, something that everyone would agree is legitimate if it occurred in any other context." Moreover, our author also notes that under blackmail "money is usually the valuable consideration demanded" and that "the threat is to engage in entirely legal gossip."

Second, since the blackmailee paying the blackmailer is an action, the blackmailee must have benefited relatively by transferring money. Otherwise, he would not have paid. However, contrary to Car Dealer, the blackmailee did not benefit in absolute terms because he would have been better off when compared to the situation in which the blackmailer had had nothing to do with him at all (since then he would preserve his reputation for free).

"Thus, in this respect, the blackmailee is in the same position as the highwayman's victim in Highwayman. That is, he benefits only relatively but not absolutely."

Into which philosophical category does the blackmailer fit? WD's car salesman or their highwayman? These authors plump for the latter. But, in my view, the former is correct. The highwayman threatens violence; he is a rights violator. Does the car salesman threaten violence? Of course not. What about the blackmailer? Again, we must answer in the negative. If he does any such thing as to threaten violence, he is an extortionist, not a blackmailer. The only thing the blackmailer threatens is to engage in gossip. But gossip is an aspect of free speech, it is not at all a rights violation.[8] This position strengthens when we realize that sometimes it is the blackmailee who approaches the blackmailer, and not the other way around. The supposed victim initially offers money to the blackmailer, in order to shut him up. So it is not, merely, the case that the blackmailer more resembles the car salesman than he does the

---

[8] By the way, the car salesman also makes a "threat." He threatens that if you do not pay him for the purchase of the automobile, he will not give it to you. You, also, "threaten" him: if he does not give you the vehicle, no money will be forthcoming from you.

highwayman. Rather, it is a fact that the blackmailer and the car sales-man are indistinguishable in this regard; neither threatens a rights violation. In very sharp contrast indeed, the highwayman occupies a completely separate category, a threatener of unjustified violence.[9]

As for being better off if someone disappears, WD are in danger of conflating the highwayman with the young brother of the lady the suitor wishes would disappear. They aver as follows:

> The only relevant difference between the two cases is justice of the threat element and, therefore, of the subsequent distribution. Hence, blackmail ex-changes would be just, although welfare-diminishing in the relevant sense.

But precisely the same analysis applies to the younger brother and also to the tree case. Are WD prepared to carry through with these examples, also, and maintain that welfare decreases there, too? Some-times, the world is not exactly as we would like it. We might wish all sorts of people to disappear: younger brothers, tree owners, competi-tors of all sorts in business, in sports, in romance. There would seem to be no limit in this regard. But this hardly demonstrates that there is something untoward going on in any of these cases, nor in the black-mail instance either, the views of my two philosophical colleagues to the contrary notwithstanding.

Saith WD:

> Thus, we have a case that seems to run counter to Rothbard's first welfare theorem that just exchanges always increase social utility. To illuminate further why we contend that blackmail exchanges do not increase black-mailees' welfare, we should come back to our distinction between benefit-ting relatively and benefiting absolutely. We might also call benefiting (sic) relatively benefiting in a weak sense, whereas benefiting absolutely benefit-ing in a strong sense. Now let us define benefitting in a weak sense as maxi-mizing one's welfare under a newly imposed budget constraint. In fact, little wonder this sense of benefitting is weak. For we should bear in mind that every instance of human action benefits its doer at least in the weak sense.

These authors are to be congratulated upon creating a new distinction, benefitting in the weak and in the strong sense, relatively versus abso-lutely. The former in each pair is compatible with wishing the trading partner to have vanished, the latter, not. I commend them upon their creativity. This is indeed an important philosophical breakthrough. And, I concede to them that in the case of blackmail, the target[10] does indeed wish that his partner in this commercial interaction did not exist, along with the younger brother, the tree owner and all the rest. But the lesson usually drawn is invalid: that it ought to be prohibited by law. The proof is a reductio ad absurdum: if you ban blackmail, you must also do so with regard to all these other cases I have been adum-brating. It is highly problematic to ban younger brothers, owners of growing trees, etc.

---

[9] Categorization is important in philosophy. See on this Barnett and Block (2008).

[10] I will not say victim, since he gains at least in the weak or relative sense

Our learned authors further aver:

> Whatever economic agents do, they maximize their expected welfare under the occurrent circumstances, whether welcome or not. However, were Austro-libertarians to adopt the weak sense of benefitting in their defence of the presumed social-welfare-enhancing character of blackmail exchanges, they would at the same time prove too much. For then, it would transpire that the gunman's proposal "Money or your life" is welfare-enhancing too. After all, whatever the gunman's victim happens to choose under the thus imposed constraint will automatically increase his expected welfare.

As a libertarian, I cannot of course countenance highwaymanship. However, fair is fair. Some of these robbers just out and out murder their victims, not offering them any choice at all. The criminal resorted by WD, in contrast, is a relatively decent chap. He offers his prey a choice. I would much prefer to be in the hands of this relatively benevolent felon than under the gun of his nasty brother in arms who offers no option at all, just pulls the trigger. But this observation on the part of my friends and colleagues WD does not undermine my examples of the younger brother, tree owner, attempt at a reductio ad absurdum against their position. I congratulate them on their relative absolute distinction, but fail to see how it undermines the case for legalizing blackmail.[11]

Next in the batter's box is this statement from WD:

> it is a matter of course that no Austro-libertarians would be ready to bite the bullet and thus concede that the victim's exchange with the gunman constitutes a Pareto-superior move. Besides being extremely counterintuitive, this move would violate the second welfare theorem, which has it that no unjust exchanges ever increase social utility.

It all depends upon the starting point. Handing money over to the extortionist most certainly is Pareto-superior compared to being shot to death. Both parties "gain" from this "exchange" when compared to that other option. The criminal saves the cost of a bullet, reduces the risk of being punished as a murderer, the victim keeps his life. Another difficulty with the position laid out by these scholars is that they keep insisting that blackmail is an "unjust exchange." The blackmailer is not at all an extortionist. His threat is to do something entirely legal: gossip. Yes, the blackmailee would be better off if the blackmailer did not exist, or did not know his secret, or had no desire to engage his vocal cords in any such manner. But none of that renders his offer to be paid for silence "unjust."

But WD are not having any of this. Rather, they maintain:

> … to establish whether the blackmailee actually benefits from the blackmailer's proposal we should compare this situation to a merely possible situation in which the actual blackmailee does not have to deal with the actual blackmailer at all, everything else equal. It seems quite clear that the ac-

---

[11] Our authors do not call for a legal prohibition of blackmail. However, to call something "unjust" is to at least hint at this conclusion.

tual blackmailee would be better off if no blackmailer were around, for in this situation the former would not even have to pay to preserve his reputation. By contrast, once the blackmailer appears on the stage and makes his blackmail proposal, there is no chance for the blackmailee to preserve his good reputation and keep the money. Therefore, it stands to reason that the blackmailee does not benefit absolutely when given a blackmail proposal. And, rather unsurprisingly, the same remark applies to Highwayman. The highwayman's actual victim would have been better off had he had nothing to do with the highwayman at all in the first place. Once confronted by the highwayman, the victim can no longer preserve his money and his life.

True, true, and well said to boot.[12] However, this analysis continues to be vulnerable to the cases of the tree owner, the younger brother etc. There, too, there can be no absolute benefit. But that failure should not be the criterion of economic welfare. We should be more modest and content ourselves with relative benefit. Absolute benefit is a will-o-the-wisp; not always to be attained; often, not. There must be thousands of examples where someone would be better off if someone else simply did not exist, or engage in action that is entirely legal. For example, in the last National Basketball League Championships, the Denver team beat the one from Florida in the final contest. Surely, the latter would have been better off if the former did not play as well as they in fact did play.[13]

In the next section of their paper, WD switch gears and attempt to demonstrate that "Unjust exchanges are not necessarily welfare-diminishing." I fear I cannot say they have succeeded. They argue as follows:

> Suppose A has an old broken fridge in his backyard, which is an economic bad for him. He would like to get rid of it, but it takes disposing of it in a faraway junkyard. Selling it would also be burdensome for him due to high transaction costs. So, the fridge just sits there in the backyard spoiling its owner's view. One day he sees, to his delight, a thief absconding with the fridge. Having realized his fridge is thus being removed for free, he decides not to interfere. First of all, this exchange of an old fridge for the satisfaction of having it removed is unjust. Clearly, our thought experiment stipulates that person A holds a property right in the fridge. Additionally, the above scenario assumes that A has never waived his ownership rights.

This is all well and good. It must be acknowledged that it is also far more than passing clever. The example exhibits great creativity, for which these authors are to be highly congratulated. However, to what extent is this even an "exchange." An exchange, contrary to WD, implies the human action of at least two parties. Here, there is only one person who is doing anything: the thief. This is an example of robbery. True, from a common-sense point of view, A has indeed benefitted. How

[12] Given the illicit basic premise on which it is predicated

[13] Of course this applies to any opponent in a competitive game; if they did not exist, there would have not been able to be any contest. No one can be deemed a champion if they cannot outplay their rival.

do we know that? Why, WD have told us that. However from an Austrian perspective, we are entitled to reach no such conclusion. All we know as praxeologists is that the thief absconded with property owned by someone else. We are not at all entitled to assume that the owner regarded this as a garbage good, and was happy to be rid of it.[14]

Here is the second error committed by my learned friends in this fridge scenario. They state:

> However, there is a worry that the putative theft cannot count as rightvio-lating (sic) simply because A welcomes it, which might translate into a tacit waiver. But this charge is unavailable for Austro-libertarians, who repudi-ate the juridical significance of tacit or implicit consent.

Yes, as Hoppe (Hoppe 2006: 389–390), Evers (Evers 1977: 193), Nozick (Nozick 1974: 287) Rothbard (Rothbard [1982b] 2002: 164–166) and Barnett (Barnett 1986: 317) correctly maintain, most representative political theorists, from John Locke to James Buchanan and John Rawls, have tried and failed to justify the existence of government on the basis of implicit contracts. But we can hardly deduce from this un-doubted fact that there can be no such thing from a libertarian, if not an Austrian point of view.

Our authors opine: "the above thought experiment assumes that the exchange in question involves an action on the part of A. After all, A omitted to interfere with the process of stealing and as Mises (Mises 1998: 13) famously contented, all omissions are actions."

Methinks WD stretch this undoubted insight of Mises's beyond rea-sonableness. Suppose that the thief stole the fridge while the owners were away from home. No longer could it be even hinted that they ac-quiesced in the theft. Yet, nothing substantive has changed from alter-ing the thought experiment offered therein. The thief is still equally a thief.

Consider the following: Jones goes to a restaurant, orders a cup of coffee and drinks it. Whereupon he is presented with a bill for $1 mil-lion dollars. Must he pay this amount of money? Of course not. There is a tacit agreement in effect that if the eatery wants to charge that amount, indeed, anything out of the ordinary, they must notify the cus-tomer and obtain his explicit agreement. It does not logically follow from the undoubted fact that some tacit agreements are invalid, the ones proving that government is a voluntary organization, that all im-plicit contracts are without legal foundation.

Strangely enough, our authors themselves admit as much. To wit, they allow that when you take a taxi, or order a drink, you are implic-itly agreeing to pay for this good or service; that when friends shake hands, they tacitly agree to do so but when a bully grabs a woman's hand from behind, no such accord is in operation.

[14] Often, owners in A's position would place a sign on the fridge, something along the lines of "free" or "take this." Then, we could go along with WD's assessment. But, then, taking this refrigerator would hardly be considered theft.

In the next section of their paper our authors continue their attempt to undermine the case for the legalization of blackmail. They do so on the basis of the "Involuntariness charge." They state:

> However, as we remember, Austro-libertarians adhere to the Nozickian (Nozick 1974: 262) rights-based contrast, in the blackmail scenario, the agreement on the part of the blackmailee secured by the blackmailer's proposal is voluntary since there is no right violation looming in the case of the blackmailer spreading the unwelcome gossip. In other words, in the blackmail scenario, the blackmailer's threat is legitimate and it is for this reason that when the blackmailee agrees to pay, he does so voluntarily. Now because he agrees voluntarily, Austro-libertarians could try to argue that the exchange is welfare-enhancing, regardless of the fact that he loses in absolute terms.

But the blackmail target does not at all "lose" from the actual exchange of money for silence. Rather, he is disadvantaged from the fact that someone knows his secret and is about to blab about it. The exchange itself is mutually beneficial and hence promotes welfare. WD are not sufficiently distinguishing between the exchange itself, which is necessarily mutually beneficial, and the fact that the blackmailer threatens to gossip, which is not an exchange at all. It is the same, to return to the view-blocking tree example. The buyer benefits from the tree owner pruning his tree. That is the exchange, that is the only exchange. There is no other exchange. None. The fact that the tree grows and blocks his view is not at all an exchange. Again, WD are failing to distinguish these two very, very different phenomenon. It is not at all the case that there are two exchanges, an absolute one and a relative one, and that the former is beneficial, the latter not. I repeat, there is only one exchange: the purchase of silence in blackmail, the purchase of tree pruning in order to improve the view. The fact that trees grow, and that people sometimes learn secrets about us that we do not want revealed, are not exchanges. They are just facts of life.

I must now conclude. This is a brilliant paper. It is radical, in that it gets to the very root of the libertarian political economic philosophy. Its distinction between relative and absolute economic welfare is singularly important. It uses the legalization of blackmail as a vehicle to overturn the Rothbardian claim that this exchange is both justified and welfare enhancing. It allows this on the basis of relative, but not absolute exchange. The fridge example is a masterful one in attempting to tease apart justice (is it a clear crime), from welfare enhancement (WD claim it does have this effect).

I have given reasons to reject much of the claims of these scholars: tree growing, younger brother, general competition. But I acknowledge that reading their essay and writing about it has been an exhilarating experience. I have been in the presence of first class minds, and I hope and trust that together, the three of us, have advanced our knowledge of libertarian that proverbial once millionth of an inch closer to the Truth with a capital "T."

References:

Barnett, R. E. 1986. "A Consent Theory of Contract." *Columbia Law Review* 86 (2): 269–321.

Barnett, W. and Walter B. E. 2008. "Economic categorization." *Laissez-Faire* 28 (1): 4–12.

Block, W. E. 1972. "Blackmailer as Hero." *Libertarian Forum* 4 (10): 1–4.

Block, W. E. 1976. *Defending the Undefendable*. New York: Fox and Wilkes.

Block, W. E. 1986. "Trading Money for Silence." *University of Hawaii Law Review* 8 (1): 57–73.

Block, W. E. 1997. "The Case for De-Criminalizing Blackmail: A Reply to Lindgren and Campbell." *Western State University Law Review* 24 (2): 225–246.

Block, W. E. 1998. "A Libertarian Theory of Blackmail." *Irish Jurist* 33 (10): 280–310.

Block, W. E. 1999a. "Blackmailing for Mutual Good: A Reply to Russell Hardin." *Vermont Law Review* 24 (1): 121–141.

Block, W. E. 1999b. "The Crime of Blackmail: A Libertarian Critique." *Criminal Justice Ethics* 18 (2): 3–10.

Block, W. E. 1999c. "Replies to Levin and Kipnis on Blackmail." *Criminal Justice Ethics* 18 (2): 23–28.

Block, W. E. 1999d. "Blackmail and Economic Analysis." *Thomas Jefferson Law Review* 21 (2): 165–192.

Block, W. E. 2000a. "Blackmail is Private Justice." *University of British Columbia Law Review* 34 (1): 11–37.

Block, W. E. 2000b. "Reply to Wexler: Libertarianism, Blackmail and Decency." *University of British Columbia Law Review* 34 (1): 49–53.

Block, W. E. 2000c. "Threats, Blackmail, Extortion and Robbery And Other Bad Things." *University of Tulsa Law Journal* 35 (2): 333–351.

Block, W. E. 2000d. "The Legalization of Blackmail: A Reply to Professor Gordon." *Seton Hall Law Review* 30 (4): 1182–1223.

Block, W. E. 2001a. "The Logic of the Argument in Behalf of Legalizing Blackmail." *Bracton Law Journal* 33: 56–80.

Block, W. E. 2001b. "Toward a Libertarian Theory of Blackmail." *Journal of Libertarian Studies* 15 (2): 55–88.

Block, W. E. 2003. "Berman on Blackmail: Taking Motives Fervently." *Florida State University Business Review* 3 (1): 57–114.

Block, W. E. 2009. "Reply To Matt Mortellaro On 'Block's Paradox': Causation, Responsibility, Libertarian Law, Entrapment, Threats And Blackmail." *Libertarian Papers* 1 (33): 1–12.

Block, W. E. 2013. *Legalize Blackmail*. New Orleans: Straylight Publishing.

Block, W. E. and Anderson, G. 2001. "Blackmail, Extortion and Exchange." *New York Law School Law Review* 44 (3): 541–561.

Block, W. E. and Gordon, D. 1985. "Extortion and the Exercise of Free Speech Rights: A Reply to Professors Posner, Epstein, Nozick and Lindgren." *Loyola of Los Angeles Law Review* 19 (1): 37–54.

Block, W. E., Kinsella, S. and Hoppe, H. 2000. "The Second Paradox of Blackmail." *Business Ethics Quarterly* 10 (3): 593–622.

Block, W. E., Kinsella, S. and Whitehead, R. 2006. "The Duty to Defend Advertising Injuries Caused by Junk Faxes: An Analysis of Privacy, Spam, Detection And Blackmail." *Whittier Law Review* 27 (4): 925–949.

Block, W. E. and McGee, R. W. 1999a. "Blackmail as a Victimless Crime." *Bracton Law Journal* 31: 24–48.

Block, W. E. and McGee, R. W. 1999b. "Blackmail from A to Z: A Reply to Joseph Isenbergh's 'Blackmail from A to C.'" *Mercer Law Review* 50 (2): 569–601.

Evers, W. 1977. "Social Contract: A Critique." *Journal of Libertarian Studies* 1 (3): 185–194.

Hoppe, H. 2006. *The Economics and Ethics of Private Property*. Auburn: Ludwig von Mises Institute.

Mack, E. 1982. "In Defense of Blackmail." *Philosophical Studies* 41: 273–284.

Mises, L. 1998. *Human Action: A Treatise on Economics. The Scholar's Edition*. Auburn: Ludwig von Mises Institute.

Murphy, R. P. 2019. "The Case for Legalizing Blackmail." *Mises Wire.*
https://mises.org/wire/case-legalizing-blackmail

Nozick, R. 1974. *Anarchy, State, and Utopia*. Oxford: Basil Blackwell.

Rothbard, M. N. 1982a. "Law, Property Rights, and Air Pollution." *Cato Journal* 2 (1): 55–99.

Rothbard, M. N. 1982b. *The Ethics of Liberty*. Atlantic Highlands: Humanities Press.

Rothbard, M. N. 2006. *For a New Liberty. The Libertarian Manifesto*. Auburn: Ludwig von Mises Institute.

Rothbard. M. N. 2011. "Toward a Reconstruction of Utility and Welfare Economics." In M. Rothbard: *Economic Controversies*. Auburn: Ludwig von Mises Institute, 289–333.

Wysocki, I. and Dominiak, L. 2023. "How Does Justice Relate to Economic Welfare? A Case Against Austro-Libertarian Welfare Economics." *Croatian Journal of Philosophy* 23 (67): 51–67.

Must Pornography Be Passed Over in Silence?

CAROLA BARBERO and ALBERTO VOLTOLINI
University of Turin, Turin, Italy

This paper critically examines leading feminist philosophical arguments asserting that inegalitarian pornography inherently perpetuates the objectification and silencing of women, thereby warranting moral condemnation or legal restriction. While recognizing the seriousness of these concerns, we argue that neither objection holds, regardless of how objectification or silencing is conceptualized. Central to our position is the distinction between fictional and non-fictional pornography. As fiction, we contend, pornography does not intrinsically validate real-world beliefs or behaviors regarding women's subordination. Even in non-fictional ("documentary") contexts, the purported causal link between pornography and harm remains unsubstantiated. The paper deliberately sets aside ethical concerns about coercion in pornography's production (e.g., exploitation, abuse) to focus on its alleged social effects. By interrogating the assumed mechanisms of influence—whether through fictional representation or documentary realism—we challenge the foundational premises of anti-pornography arguments and advocate for a more nuanced assessment of pornography's role in shaping social norms.

Keywords: Pornography; objectification; silencing; feminist philosophy; fiction; illocutionary acts.

Introduction

In some influential papers, some feminist philosophers have claimed that pornography, or better *inegalitarian* pornography—so qualified: "sexually explicit representations that as a whole eroticize relations (acts, scenarios, or postures) characterized by gender inequity" (Eaton 2007: 676) (from now on, we will take this specification for granted)[1]—

---

[1] For more on the distinction between egalitarian and inegalitarian pornography, cf. Eaton (2007: 676–679).

should be, if not legally banned, at least civilly actionable, or overall morally blamed.[2] In this context, different theses have been put forward. First of all, by representing reluctant women who give in to sexual pleasure upon being objectified (i.e., in a first approximation, not treated as persons, but merely as objects for sexual gratification, Dworkin 1985; Langton 1995; MacKinnon 1987; Nussbaum 1999; Vadas 2005), this form of pornography incentivizes in their appreciators' beliefs that women enjoy being treated in that way. Moreover, by letting women fantasize only about satisfying men's desires, this form of pornography makes its appreciators to not only subordinate them (Langton 1993), but also take their "no" utterances to actually mean "yes," as a form (to be properly cashed out) of silencing (Dworkin and MacKinnon 1988; Langton 1993; Langton and West 2009; McGowan 2009, 2012).

So first of all, if such an objectification really took place in pornography, women represented by pornography would be characterized as instruments of pleasure that can be undressed, disguised, handled, ceded, shared, or violated without taking into account the fact that they are persons. Moreover, because of their subordination, they would be induced to silence. For on the basis of what one reads or sees in pornography, women apparently denying their consent to sex would be taken as meaning the opposite, thereby preventing them from successfully performing certain illocutions (e.g., from refusing sexual advances), in their being silenced in some sense, hence in their failing to secure uptake (Langton 1993; Hornsby and Langton 1998).

If all the above were what *always* happened in pornography, then it would be hard to disagree with the aforementioned positions. As for objectification, how can one remain indifferent to the fact that a man reduces a woman to a mere instrument of pleasure to satisfy his desires? Ditto for silencing. How can one tolerate that women cannot express their will on sexual issues and are not taken seriously in their refusing their consent on such matters?

Clearly enough, the debate on what objectification and silencing respectively amount to is very complex and subtle,[3] but to enter in this

[2] For example, Longino (1980: 45), following Kant, explains how pornographic depictions, by showing women subordinated by men and seen as mere instruments of male's desires, actually promote an immoral (and henceforth unacceptable) treatment of women. Dworkin's (1985) and MacKinnon's (1987) proposals incorporate a Kantian model of sexual morality, too. On the limits of such an approach together with a critique to Kant's original account see Shrage (2005).

[3] As regards objectification, Nussbaum (1999) highlights its seven characteristics at stake in pornography: instrumentality, denial of autonomy, inertia, fungibility, violability, ownership by third parties, and denial of the individual's subjectivity. Langton (2009) rectifies this characterization. As regards silencing, Bianchi (2008) and Caponetto (2016) explicitly focus on the connection between pornography and silencing by resorting to illocutionary acts. Bianchi starts from Saul's (2006) critique to Langton's (1993) idea that pornography can be construed as illocutionary acts (specifically acts of subordinating/silencing women). According to Saul, works

debate is not the matter of this paper. For we want instead to show that, whatever objectification and silencing really amount to, it is not really the case that pornography leads either to objectification, or to subordination, or even to silence. According to our view, pornography neither displays such reproachable situations concerning women, nor induces one to have weird beliefs about such situations. For first, since we maintain that pornography is basically a matter of fiction, given its fictional character no such consequence derives. Second, this would not even be the case if one were involved with documentary, hence non-fictional, pornography. Section 1 deals with pornography considered as fiction. Section 2 deals with documentary pornography.

Before starting, a caveat. In this paper we will not consider possible objections to pornography based on eventual coercion and exploitation of women in the *production* of pornography. We will not focus on this aspect, since we obviously agree that any sexual action that is not grounded on shared consensus is simply illegal and illicit. Unfortunately, as the #metoo movement has abundantly showed, sexual abuse, sexual harassment and rape culture are typical of many workplaces and not just limited to the pornography industry.

1. *Pornography as Fiction*

To begin with, we claim that pornography is basically a *piece of fiction*, an artifact (as Mikkola (2013) argues by following Thomasson (2003)), in which whatever is represented, explicitly or implicitly, is primarily *fictionally*, hence not really, true; i.e., it is primarily true in the world of the fictional context of make-believe that is activated in the pornographic setting. Indeed, a pornographic piece is basically either a literary story in which a writer makes believe that sexual acts of any kind occur, or a movie whose actors play the role of people involved in such acts. Granted, to say that pornography is fictional is not a new idea (for some previous defenses of it, cf. Cooke 2012; Liao and Protasi 2013), and is supported as well by some insiders, such as the famous director of porno movies Erika Lust (https://www.bbc.co.uk/programmes/p09cbtjy). Yet, by articulating this idea in more detail, we aim to show how it may resist criticisms of various sorts that can be addressed, and have been already addressed, against it.

So, the *fictional* (and not the supposedly real) context, in which women are portrayed or presupposed as submissive or deprived of

of pornography are not to be considered as speech acts, because only utterances in specific contexts can qualify as such. While agreeing with Saul that only utterances in contexts can be considered speech acts, Bianchi explains how Saul's reformulation can be seen as not undermining Langton's thesis, provided that one appeals to Predelli's (1998) distinction between context of utterance and context of interpretation. Caponetto identifies four categories of silencing (essential, authority, sincerity, and seriousness) and sees illocutionary silencing as a failure in speech acts representing as such a genuine instance of illocutionary disablement, thereby constituting a harmful manifestation of discursive injustice.

their rights, should be taken into account (Saul 2006; Cooke 2012; Heck 2023).[4] For the fact that things are told here in a fictional context makes a fundamental difference. It is one thing to say to a woman, in a pornographic film, "You are my sexual slave," but it is quite another to say the same to a woman we meet in the street. In the first case, the utterance occurs within a fictional context in which the subject is acted out.[5] Hence, it is true, but in the fictional world of that context, thereby at most determining objectifications, or illocutions of subordination, or even silencings that are merely *fictional*. Whereas in the second case, the utterance occurs in a real context. Hence, it would be a genuine problem if it were true in the real world of that context, by determining those phenomena as *real*. Yet, considering an utterance in the fictional context as if it occurred in a real context would be incorrect, since fiction allows utterances to be true *in its own context*. To consider a similar case, just think of the many novels with wrong and offensive content. By adopting the *fictive stance* (Lamarque and Olsen 1994; Davies 1997), one can certainly ignore the fact that such a novel sometimes says something really wrong by means of this content, just to enjoy the fiction, regardless of that content. The same goes for pornographic narrative. If, when reading *History of O*, one realizes that in this fiction a woman—O—likes to be humiliated, one can, in adopting the fictive stance, ignore the fact that this is not *really* the case, but only *fictionally*, in order to be aesthetically involved in the plot. From this point of view, a pornographic narration is no different from a horror or a violent narration (Liao and Protasi 2013), or even from 'bad' jokes (Dennett *et al.* 2011), in which whatever is represented is true, but in the world of the fictional context of such narrations. Usually, the audience is well aware of that.

Certainly, pornography is a case of fiction showing that, as Austin originally understood (1961: 240–241), to be fictionally the case that *p* does not entail that it is really the case that *not-p*. For example, in a porno movie, although it is not really the case that two protagonists make love by being emotionally involved as it is instead fictionally the case, still, if it is fictionally the case that such protagonists copulate, it is also really the case that they do so. Moreover, although it is not explicitly said in the movie, it is also fictionally true that when copulating, their pleasure brain receptors are active, since this is a real truth that is imported in the movie. But *pace* Langton and West (2009) and McGlynn (2021), the fact that real situations also occur in pornography does not undermine the pornography's fictionality. For this is on a par with

[4] *Pace* Heck (2023), that context is not a mere fantasy context, but a properly *fictional* context. For the imagination taking place there does not float free, but is prescribed by the pornography authors.

[5] According to Zamir (2013: 78), there is an analogy between pornography and advertisement because both *use acting* in order to achieve their goal rather than being actually interested in acting.

historical novels. A piece of pornography in which it is both fictionally and really the case that *p*, as in the above example, is just like a historical novel—if you like, it is 'historical' fictional pornography. Indeed, in this respect, historical novels match what happens in pornography. Not only it is fictionally the case in the historical novel of Alessandro Manzoni, *The Betrothed*, that a branch of Lake Como ranges towards the south, but it is also really the case. Likewise, it is both fictionally and really the case in *The Bethrothed* that the city of Milan is about ten hours walk from the city of Como, although Manzoni is silent about that. For this real truth can be also imported in *The Bethrothed* as a fictional truth. Given that, one may take a historical novel as a piece of history telling how things have unfolded in the real world. But the novel can also be read as a fictional account of how things might have gone in a fictional world parallel to, but independent of, the real one.[6]

On the basis of the above, we could even say that pornography, like any work of fiction, is ultimately about fictional characters who are, metaphysically speaking, objects of the same kind as Madame Bovary or Mickey Mouse. Therefore, the women we focus on in pornographic works are also fictional objects. Indeed, those generated via pornography are not, strictly speaking, ordinary women, but fictional characters whose rights there is no point in worrying about; no more, at least, than there is in worrying about Balthus's *Girl with a Cat* or Sade's sisters we find in *Justine* ("*or the misadventures of virtue*") and *Juliette* ("*or the prosperities of vice*"). Fiction has reasons (and objects) that reason (focused too much on reality) does not (always) recognize.

But there is no need to appeal to fictional characters created by fiction to stress the gap between fiction and reality, as is the case with pornographic fiction. Fictional contexts do not usually involve fictional characters; individuals in those contexts are usually taken to be flesh-and-blood individuals like you and me. But, as we said, whatever is true of those individuals within such contexts is not automatically true outside such contexts. If it is true outside those contexts, it is *independently* true, as with historical novels and, if you like, 'historical' fictional pornography.[7]

---

[6] Our claim echoes Currie's (1990: 46) idea that a fiction, if actually true, is accidentally so.

[7] Clearly enough, illicit pornographic contents even located in a fictional space may raise the problem of *imaginative resistance*, i.e., the problem of whether one may be unwilling to even fictionally endorse such contents (Barbero and Voltolini 2024). Granted, as some people say (e.g. Stokes 2006), imaginative resistance is *subject-relative*: what is unimaginable for some could be imagined by others. So, to circumvent this relativity, the problem can only sensibly arise in a *prescriptive* form; namely, whether one *should* imaginatively resist such contents. Obviously, immoralists with respect to fiction would see no problem with such contents (Kieran 2002; Eaton 2012; Zhen Li 2021). But even if we put immoralism aside, we may face the problem in a Freudian vein by saying that endorsing such contents fictionally may be a way of fulfilling bad drives *fictionally* rather than *really*, hence of *sublimating* such drives.

But even if one accepts the idea that pornography is basically fictional, many criticisms of it seem to remain untouched. They concern, first, the unbearable things that one can allegedly learn by pornography; second, the bad things that pornography can make you believe anyway; third, the bad actions and emotions that pornography may trigger one to perform or to entertain; fourth, the fact that repeated exposure to pornography can lead even ordinary people to forget the distinction between fiction and reality, by following what it suggests as if it were real. We will address these criticisms in turn.

To begin with the first point, an opponent endorsing the afore-mentioned feminist perspective may retort that it is always possible that people learn wrong lessons from such works. This learning may be taken literally, as amounting to the knowledge of a real situation, or non-literally, as amounting to an appreciator's belief in such a situation.

Let us start with the first, the literalist option. Our opponent might remark that pornography, by making something fictionally true, is made to learn something from what happens in real life, in this case to do with sexual facts. In particular, this has to do with pornographic narrative as a *factory* (Gendler 2000); namely, as an active generator, in its case, of new bad cognitive content, producing new misunderstandings, misleading perspectives, or moral misconceptions.[8]

Yet we may reply that, first, in general, learning something from fiction is a matter of conversational implicatures *to be really true* (Voltolini 2021b; Barbero and Voltolini 2024).[9] For such learning is for us a form of *propositional* knowledge; since knowledge is factive (knowing that *p* entails *p*), those implicatures must be true. But secondly, in the above pornographic cases, even if we accept that their authors want to convey something perverse as being true in the real world (which should not be taken for granted: Cooke 2012: 234), nothing really true is actually conveyed, because no such implicature is really true! For example, suppose that, by fictionally writing, in *Fifty Shades of Grey,* that Anastasia Steele derives pleasure by being submitted and humiliated by Christian Grey, her author, E.L. James, had wanted to convey the idea that for women it is in general extremely exciting to be so treated. Yet, James would have tried to convey something *that is re-*

---

[8] For Gendler there is another sense in which we can learn something from fiction; namely, when the fictional narration is taken as a *clearinghouse*, which allows us to learn ordinary facts from fiction, e.g. when we read from the Holmes stories how far is Paddington station from Waterloo station in London. Yet as far as pornography is concerned, this sense is irrelevant for the present debate. There is no particular problem in learning from pornography about, say, how human anatomy is made. On the importance of taking into account genre variations for better evaluating different effects of pornography consumption, see Liao and Protasi (2013: 110–113).

[9] If pornography is fiction, one cannot say that one learns from pornography that women have an objectified status, as construed by pornography (McGowan 2005; Jenkins 2017). For again, this cannot be learnt, since it is false. At most, as we have seen before, pornography may *fictionally* construe that objectification.

ally false, since real women do not find that treatment exciting. Hence, given such falsity, no such moral could be learned from her narration.

By defending this point, we do not want to deny in general that one can learn something from pornography, if one manages to derive some true implicatures from it. Yet curiously enough, such a derivation could even have benign effects. Suffice it to say that, thanks to pornography, one could learn about sexual mechanics, explore sexual identities and orientations, and thus conclude that some sexual practices are actually not to be condemned *per se*, as one may originally believe erroneously (maybe for cultural, religious, political, and social reasons).[10]

If the above is the case, it is also wrong to say that pornographic fiction prescribes that something morally bad *ought* to be *really* the case, whether this is supposed to hold for inegalitarian pornography in general (Longino 1980; Eaton 2007) or just in the case of so-called mainstream, response-realistic, pornography (Liao and Protasi 2013).[11] For if *descriptive* implicatures do not hold, *normative* implicatures do not hold either. If it is false that women like to be submitted and humiliated, it is even more false that women *ought* to like being submitted and humiliated, if by chance E. L. James had wanted to convey this prescription via her *Fifty Shades*.

At this point, moreover, comes the second point. Our opponent may advance the second, non-literalist, option about learning from fiction and retort that what counts for distancing from pornography is not whether the kind of real beliefs one draws from pornography are really true, but the mere fact that from it one *may export* such beliefs in real life, independently of whether they are true (McGlynn 2021). This determines, for example, a perlocutionary effect of subordinating women, says Langton (1993). Just as, after reading Goethe's *The Sorrows of Young Werther*, some people thought it appropriate to see suicide as the proper affirmation of individual freedom, likewise some other people, after watching pornographic movies, may think it appropriate to go around and treat women as they have seen in the movies. Should one then conclude that pornographic works should be censored, or at least morally blamed? In this respect, let's consider again the strongest view against pornography[12] according to which, by implying the subordina-

---

[10] Granted, it is far from easy to find *clear* evidence of what and how consumers do learn from pornography, as emphasized by Litsou et al. (2020).

[11] What grounds the distinction between response-realistic and response-irrealistic pornography, i.e., the amount of importation of truths from the real world (Heck 2023), is irrelevant for our purposes. For that importation only makes it the case that what is really true is also fictionally true in the pornographic context. Yet, as we saw, for us evil *fictional* truths are not problematic. We are focusing here on exportation, not on importation.

[12] The one defended by MacKinnon according to which in pornography "(i) women are presented dehumanized as sexual objects, things, or commodities; or (ii) women are presented as sexual objects who enjoy humiliation or pain; or (iii) women are presented as sexual objects experiencing sexual pleasure in rape, incest

tion of women, pornography forces them into silence, violating their most fundamental rights in objectifying them. Many arguments have been put forward in support of this position, but we will only examine one of them (Langton 1993; Hornsby 1993), by generalizing its import. This argument claims that the primary effect of pornographic works is to subordinate and relegate women to a state of inferiority and silence and that the secondary effect is to foster in society the belief that women are merely sexual objects without the right to speak out or rebel against their oppressors. This second effect is further supposed to have three different consequences (MacKinnon 1987): (1) being aware of the fostered hostility against them, women would develop the propensity to express themselves as little as possible (e.g. by not telling about the very discriminatory acts of which they are the victims); (2) being considered in low regard, women would often be mocked and ridiculed; (3) since in pornography women's utterances are given the opposite meaning to what they mean—when they say "no" they mean "yes," when they say "enough" they mean "again"—this would make it the case that the explicit intentions manifested by women in their real utterances are hardly ever taken seriously.

Admittedly, if all this were the case, one would have to take a negative attitude towards pornography. Indeed, one might even adopt a paternalistic attitude and think that because of a few consumers of pornographic works who, on the basis of their pornography-dependent weird beliefs, engage in morally deviant behavior, all possible users should be blamed. However, two problems must be addressed. First, wouldn't this attitude run the risk of infringing the freedom of all other responsible people who do not derive wrong lessons from the works they appreciate? Second, is it plausible to treat spectators as a mass of individuals unable to resist the temptation to project into reality what they see in fiction, as if Bovarysme were a pandemic, while the recent data seem to prove the opposite (Heck 2023: 19–20)? Our opponent argues: in pornographic narrations, women are inferior beings with whom one can have fun, therefore spectators will conclude that women are objects of entertainment; in pornographic narrations, women are happy to be raped, then spectators will conclude that women like to be raped; in pornographic narrations, women are humiliated and silenced, therefore spectators will think it normal to humiliate and silence women. But, really, who are these spectators? Shouldn't we question the idea that pornography is *responsible* for arousing morally deviant beliefs and hence violent or discriminatory behavior in spectators (Garry

or other sexual assault; or (iv) women are presented as sexual objects tied up, cut up or mutilated or bruised or physically hurt; or (v) women are presented in postures or positions of sexual submission, servility, or display; or (vi) women's body parts [...] are exhibited such that women are reduced to those parts; or (vii) women are presented being penetrated by objects or animals; or (viii) women are presented in scenarios of degradation, humiliation, injury, torture" (MacKinnon 1987: 176).

1978; Cooke 2012: 237–239)? If there are suggestible people who develop such beliefs, wouldn't it be safer to arrange appropriate debriefing with such people to help them change such beliefs (Saul 2006: 246)? To move to a similar case yet not involving pornography, consider the case of a disturbed spectator who, after watching *Clockwork Orange* by S. Kubrick, besides beating vagrants in the streets, also rapes women while singing "Singin' in the Rain." Didn't the movie simply made him reactivate a perverse tendency of his that he entertained completely independently and that must be primarily extirpated?

If the above is correct, it is hardly the case that, by prompting in disturbed people deviant beliefs and sexual arousal along with possibly deviant behavior, pornography suggests that such a behavior is acceptable and merited (Eaton 2007: 682). This kind of is-ought connection, which is notoriously problematic in general, is present in many other similar cases. Consider *The Godfather* by Francis F. Coppola, interesting in its portrayal of the gangs through the perspective of the gangsters and in its depictions of Mafia as a reaction to a corrupt society. This representation of a criminal counterculture includes unapologetic gender stereotypes (such as when Vito Corleone intimates to a tearful Johnny Fontane to "act like a man"), the cult of vengeance (the figure of Michael Corleone is essentially based on revenge), and the idea that anything is permissible to protect one's business and loved ones (as when Michael orders his brother Fredo's murder), which are undoubtedly significant factors in the film's appeal. Again, it may engender in some mafia-oriented spectators a feeling of approval, sharing, imitation, and revenge as well as the corresponding behavior. Yet, it does not certainly suggest that in reality such behavior is acceptable and merited (however it might be in fiction).

Yet furthermore, the third point comes to the fore. Our opponent may still retort that, unlike standard cases of fiction, pornography *precisely* aims to *trigger* sexual arousal, hence to have *real effects* on its spectators (Mumford 2013: 62). By inducing admittedly fictional sexual desires, pornography also wants to induce *real* sexual desires, in order finally to affect the spectators' *real* behavior. Consider the trivial fact that pornography leads its spectators to masturbate (Liao and Protasi 2013).

Granted, pornography has such an aim (Cooke 2012: 230). For some, this aim must be incorporated in the very definition of pornography (McGlynn 2021; Zamir (2013: 77) explicitly says: "pornography is a graphic (pictorial, cinematic, photographic, acoustic, staged) depiction of bodily display and action that is projected to generate sexual excitement in its beholder"). Yet first, it must not be taken for granted that such an aim purports to trigger emotions *outside* the fictional context in which the pornographic narration is set. In this respect, appearances notwithstanding, pornography is like other fictional cases: since they prompt to generate emotions only in the context of fiction, no real piece

of behavior must follow those emotions. Consider horror again. It may be the case that a horror movie generates fear in the context of the movie (Walton 1978, 1990, 1997). Yet such an emotion is not accompanied by the typical behavior it would prompt outside that context. For one thing, spectators' heart beating, sweating, and trembling while attending the horror movie do not prompt them to get out of the cinema running and screaming or calling the police. For they remain in their cinema seats eating popcorn and drinking Coke, perfectly aware about the fictional status of what they are watching.

Second, even if pornography managed to induce *real* emotions accompanied by *real* behavior, as in the masturbation case, one may note that the sort of dangerous and violent behavior that might follow watching pornography is not induced by pornography *per se*, but by the specific overall attitude of some of its spectators. Saying that pornography reinforces such deviant reactions does not take into account the actual dispositions of such spectators, which they have independently of pornography. There is just a mere *correlation*, but not a well-established causal relationship, between attending pornography and behaving morally badly (Cooke 2012: 250; on doubts concerning pornography research together with methodological flaws, see Eaton 2007: 697–710). Even if following Eaton (2007) one appeals to a multifactor probability-based model of causality, one may acknowledge that in the above cases, it is more likely that such a piece of behavior is caused by one's independent weaknesses/confusions rather than by the very fact of attending pornography.

Granted, iterated exposition to pornography may raise the probability that already aggressive men perform sexual crimes (Malamuth et al. 2012; Eaton 2017). But again, one may expect that the same happens with, say, violent movies watched by aggressive people. So, just as one should not blame violent stories as such, one should not blame pornography *per se*, but again, help its spectators to remove such independent weaknesses/confusions. To go back to a previous example, *The Sorrows of Young Werther* should not be blamed because some people committed suicide after having read it. They did so not because Goethe's work prompted them to do so, but because they mistook the (fictional) ending of the story as having a sort of moral to be derived from it, and therefore felt authorized to apply it to reality. Yet such a consequence does not highlight a dangerous characteristic of Goethe's work, but instead a deviant appreciation from some (evidently) confused readers.

Finally, and this is the fourth and last point, our opponent may note that watching pornography, in their typically reiterated and compulsive way, is not merely problematic for disturbed people. For it prompts spectators *in general*, even ordinary people, to obliviate the fact that it is fictional and to take the relevant pornographic narration as a documentary narration, to be reacted upon as if what it shows and tells were models to be somehow reproduced in real life, that is, as if the charac-

ters involved in that narration were showing a way people could follow in reality (for some—admittedly disputable—evidence on this concern, see Cooke 2012; Eaton 2007: 707–709). Some works of pornography are precisely aimed at blurring the distinction between fiction and reality. For Langton and West (2009), this explains why in pornography, a huge amount of real truths is imported into the fictional truths.

Yet, if this were the case, as some people have already noted in the debate (Langton 1993), the proper moral to be drawn should be to endorse a pedagogical attitude enabling ordinary spectators, who are not mentally disturbed, to reactivate the awareness that what they attend to is fictional and not real (Voltolini 2021a). By the way, this is usually the kind of strategy adopted in other similarly problematic cases. If, after having watched too many terrifying zombie movies, a subject gets out in the street fearing that she could be attacked and eaten at any hidden corner, we would simply address her with a kind tone of voice saying something like "calm down, relax yourself, it was only make-believe, there are no zombies here." In the case of consumers of pornography who confuse fiction and reality, the situation would be similar to that of someone who makes a cataloguing error, hence not different in substance from the one we find, for example, in the movie *Betty Love*. Betty, a waitress from Kansas who avidly views *A Reason to Love*—a soap set in hospitals—after losing her sense of reality due to a trauma, decides to leave for Los Angeles in search of one of the soap's protagonists, Dr David Ravell, in order to reveal him her love. People would react to Betty as a person that must be rescued from her confusion.

2. *Documentary Pornography*

So far, so good. If we are right, pornography as fiction is not responsible for the weird effects it is taken to engender. Yet at this point, our opponent may altogether reject, wholly or partially, our original assumption about the fictional status of pornographical narrative and say that pornography is, or at least some pornography is, documentary, hence non-fictional (McGlynn 2021). If this were the case, it seems that at least for documentary pornography, there would be no way of avoiding the bad repercussions that we have attempted to discard by relying on the fictionality of pornography. People watching amateurish reproachable documentary movies that had circulated via whatsapp may take them as examples of how to sexually behave. Just as after having watched a lot of whatsapp videos about birthday parties one may think that it is a good thing to arrange such parties.

Once again, first of all, we will not consider documentary pornography that is not based on sexual consensus among its actors. To repeat, if someone films himself raping a woman and shares the video with his friends, this is just a crime to be legally prosecuted.

This said, not even documentary pornography has the above reper-
cussions. Let us allow for this kind of documentary pornography. Yet as
a documentary, any piece of pornography would be a mere *singular* wit-
ness that would have no *general* motivational force, as any piece of fic-
tion is instead supposed to do—as is widely understood, *Anna Karenina*
wants us to believe not that a *particular* family, but that *any* family is
happy in the same way, while each family experiences its own unique
form of unhappiness. Here iteration is irrelevant. For no number of
projections is enough to determine a behavioral *model*. This utterly
agrees with Aristotle's distinction in the *Poetics* between history and
poetry: "The one tells what happened and the other what might hap-
pen. For this reason, poetry is something more scientific and serious
than history, because poetry tends to give general truths while history
gives particular facts" (2013: 1451b). In this vein, consider a terrorist
documentary. That documentary would represent something that re-
ally happened, hélas, if the documentary is true. Now, terrorists may
circulate a huge number of such videos. But one would neither learn
from it the general lesson that terrorism will prevail in the world—for
that conversational implicature is far from being true—nor would it *eo
ipso* prompt any corresponding general real belief. For one may react
to it by saying that although the documentary producers may have that
belief, watching the documentary does not make it the case that that
very belief is to be universally shared. Granted, one might be scared af-
ter having watched the documentary. But how one would behaviorally
react to it depends again on many other factors that have to do with
one's *specific* psychological history. Likewise for documentary pornog-
raphy, if there is any. Consider gangbang movies supposedly report-
ing the real sexual practices of certain men simultaneously copulating
with a single woman. First, those movies would hardly impart the gen-
eral lesson that women like to copulate with a male group simultane-
ously. For there is no such lesson, since that implicature would be false.
Nor second, would they prompt the corresponding *general* real belief.
One might at most be prompted to believe that *those very real women*
involved in such movies like that practice. Third, even if one were sex-
ually aroused by watching such movies, *ceteris paribus* this arousal
would hardly lead one to look for other mates to have sex together with
some woman or other.

Yet our opponent can finally retort that, just as any documentary
narration, documentary pornography may also vehiculate possibilities,
and therefore have general effects. Just as a terrorist group may send
terrible videos of tortured prisoners in order to frighten people, cannot
a member of a gang send his terrible videos of him raping women in
order to suggest emulation?

Here we must repeat what we have said about pornographic fiction.
Not only would even a pornographic documentary conversationally im-
plicate no weird belief, but either it would be also unable to prompt it,
or it would effectively induce emotions whose behavioral consequences

are not only difficult to determine, but whose actors should also be tested for their weaknesses. To stress a point already made, if by watching the news, in seeing the video of a crime someone thinks that he could do to the same, we should not only be concerned about the video, we should take care of him.

Conclusion

All in all, we conclude that pornography *qua* fiction concerns only what is true in its own fictional scope. In this respect, our moral beliefs can be subverted, but only *within* that scope; just as it happens, say, in 'bad' jokes. Moreover, appreciators are not authorized to export morally bad false implicatures outside of fiction, nor do they run the risk of being somehow invited to endorse bad beliefs, or adopt a morally bad behavior. Granted, pornography is successful when able to stimulate sexual arousal, unsuccessful if not. Yet this does not make it even a mere trigger of a morally bad way of life, whether it is a piece of fiction or not.[13]

References

Aristotle. 2013. *Poetics*. Oxford: Oxford University Press.

Austin, J. L. 1961. "Pretending." In Austin, J. L. *Philosophical Papers*. Oxford: Oxford University Press, 253–272.

Barbero, C. and Voltolini, A. 2024. "How One Cannot Participatively Imagine What One Could Cognitively Imagine." *Philosophia* 52: 643–660.

Bianchi, C. 2008. "Indexicals, Speech Acts and Pornography." *Analysis* 68: 310–316.

Caponetto, L. 2016. "Silencing Speech with Pornography." *Phenomenology and Mind* 11: 182–191.

Cooke, B. 2012. "On the Ethical Distinction between Art and Pornography." In H. Maes and J. Levinson (eds.). *Art and Pornography*. Oxford: Oxford University Press, 229–253.

Currie, G. 1990. *The Nature of Fiction*. Cambridge: Cambridge University Press.

Davies, D. 2007. *Aesthetics and Literature*. London: Continuum.

Dworkin, R. 1985. "Do We Have a Right to Pornography?" In Dworkin R. *A Matter of Principle*. Cambridge MA: Harvard University Press, 335–372.
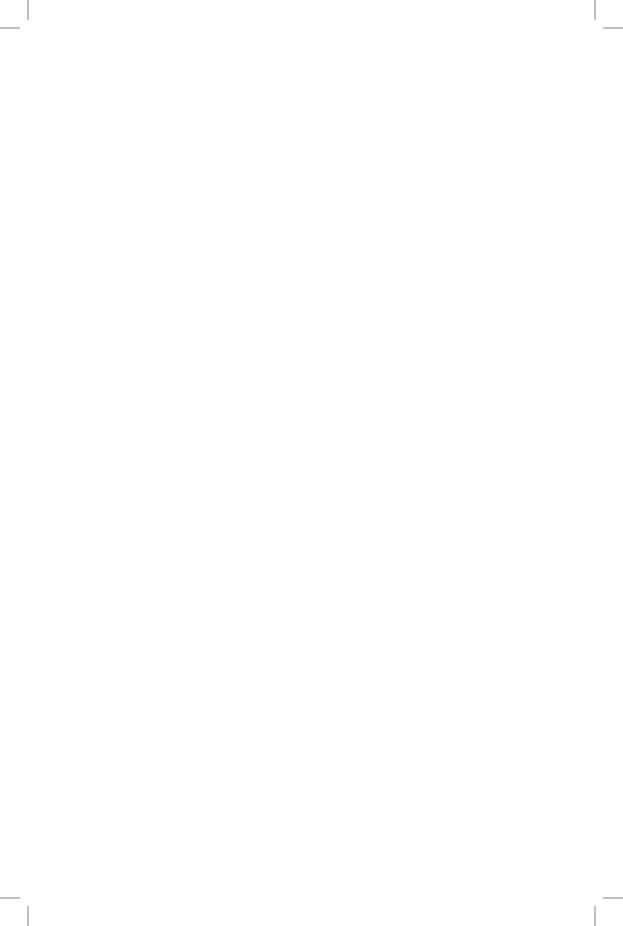
Dworkin, A. and MacKinnon, C. 1988. *Pornography and Civil Rights. A New Day for Women's Equality*. Minneapolis: Organizing Against Pornography.

Eaton, A. W. 2007. "A Sensible Antiporn Feminism." *Ethics* 117: 674–715.

Eaton, A. W. 2012. "Robust Immoralism." *The Journal of Aesthetics and Art Criticism* 70: 281–292.

Eaton, A. W. 2017. "Feminist Pornography." In M. Mikkola (ed.). *Beyond Speech*. Oxford: Oxford University Press, 243–257.

Garry, A. 1978. "Pornography and Respect for Women." *Social Theory and Practice* 4: 395–421.

Gendler, T.S. 2000. "The Puzzle of Imaginative Resistance." *The Journal of Philosophy* 97: 58–81.

Heck, R. K. 2023. "Does Pornography Presupposes Rape Myths?" *Pacific Philosophical Quarterly* 1–25.

Hornsby, J. 1993. "Speech Acts and Pornography." *Women's Philosophy Review* 10: 38–45.

Hornsby, J. and Langton, R. 1998. "Free Speech and Illocution." *Legal Theory* 4: 21–37.

Hurley, M., Dennett, D. and Adams, R. B. jr. 2011. *Inside Jokes*. Cambridge MA: The MIT Press.

Kieran, M. 2002. "On Obscenity: The Thrill and Repulsion of the Morally Prohibited." *Philosophy and Phenomenological Research* 64: 31–55.

Jenkins, K. 2017. "What Women Are For: Pornography and Social Ontology." In M. Mikkola (ed.). *Beyond Speech*. Oxford: Oxford University Press, 91–112.

Lamarque, P. and Olsen, S. H. 1994. *Truth, Fiction, and Literature*. Oxford: Clarendon Press.

Langton, R. 1993. "Speech Acts, and Unspeakable Acts." *Philosophy and Public Affairs* 22: 293–330.

Langton, R. 1995. "Sexual Solipsism." *Philosophical Topics* 23: 149–187.

Langton, R. 2009. "Autonomy-Denial in Objectification." In R. Langton (ed.). *Sexual Solipsism: Philosophical Essays on Pornography and Objectification*. Oxford: Oxford University Press, 223–240.

Langton, R. and West, C. 2009. "Scorekeeping in a Pornographic Language Game." In R. Langton (ed.). *Sexual Solipsism: Philosophical Essays on Pornography and Objectification*. Oxford: Oxford University Press, 173–197.

Li, Z. 2021. "Immorality and Transgressive Art: An Argument for Immoralism in the Philosophy of Art." *The Philosophical Quarterly* 71: 481–501.

Liao, S. and Protasi, S. (2013). "The Fictional Character of Pornography." In H. Maes (ed.). *Pornographic Art and the Aesthetics of Pornography*. Basingstoke: Palgrave, 100–118.

Litsou, K., Byron, P., Mackee, A. and Ingham, R. 2020. "Learning from Pornography: Results of a Mixed Methods Systematic Review." *Sex Education* 21: 236–252.

Longino, H. 1980. "Pornography, Oppression, and Freedom: A Closer Look." In L. Lederer (ed.). *Take Back the Night: Women on Pornography*. New York: William Morrow, 40–54.

Malamuth, N. M, Hald, G. M. and Koss, M. 2012. "Pornography, Individual Differences in Risk and Men's Acceptance of Violence Against Women in a Representative Sample." *Sex Roles* 66: 427–439.

MacKinnon, C. 1987. *Feminism Unmodified*. Cambridge MA: Harvard University Press.

MacKinnon, C. 1991. "Pornography as Defamation and Discrimination." *Boston University Law Review* 71: 793–815.

McGlynn, A. 2021. "Blurred Lines: How Fictional is Pornography?" *Philosophy Compass* 16: e12721. https://doi.org/10.1111/phc3.12721.

McGowan, M. K. 2005. "On Pornography: MacKinnon, Speech Acts, and 'False' Construction." *Hypatia* 20: 22–49.

McGowan, M. K. 2009. "On Silencing and Sexual Refusal." *Journal of Political Philosophy* 17: 487–494.

McGowan, M. K. 2012. "Sincerity Silencing." *Hypatia* 29: 458–473.

Mikkola, M. 2013. "Pornography, Art, and Porno-Art." In H. Maes (ed.). *Pornographic Art and the Aesthetics of Pornography*. Basingstoke: Palgrave, 27–42.

Mumford, S. 2013. "A Pornographic Way of Seeing". In H. Maes (ed.). *Pornographic Art and the Aesthetics of Pornography*. Basingstoke: Palgrave, 58–72.

Nussbaum, M. 1999. *Sex and Social Justice*. Oxford: Oxford University Press.

Predelli, S. 1998. "I Am Not Here Now." *Analysis* 58: 107–115.

Saul, J. 2006. "Pornography, Speech Acts and Context." *Proceedings of the Aristotelian Society* 106: 229–248.

Shrage, L. 2005. "Exposing the Fallacies of Anti-porn Feminism." *Feminist Theory* 6: 45–65.

Stokes, D. R. 2006. "The Evaluative Character of Imaginative Resistance." *British Journal of Aesthetics* 46: 347–405.

Thomasson, A. L. 2003. "Realism and Human Kinds." *Philosophy and Phenomenological Research* 67: 580–609.

Vadas, M. 2005. "The Manufacture-for-use of Pornography and Women's Inequality." *Journal of Political Philosophy* 13: 174–193.

Voltolini, A. 2021a. "Beliefs, Make-beliefs, and Making Believe that Beliefs Are Not Make-beliefs." *Synthese* 199: 5061–5078.

Voltolini, A. 2021b. "What We Can Learn from Literary Authors." *Acta Analytica* 36: 479–499.

Walton, K. L. 1978. "Fearing Fictions." *The Journal of Philosophy* 75: 5–25.

Walton, K. L. 1990. *Mimesis as Make-believe*. Cambridge MA: Harvard University Press.

Walton, K. L. 1997. "Spelunking, Simulation, and Slime." In M. Hjort and S. Laver (eds.). *Emotion and the Arts*. Oxford: Oxford University Press, 37–49.

Zamir, T. 2013. "Pornography and Acting." In H. Maes (ed.). *Pornographic Art and the Aesthetics of Pornography*. Basingstoke: Palgrave, 75–99.

Health and Disease Concepts Cannot Be Grounded in Social Justice Alone

WALTER VEIT
University of Reading, Reading, UK

Kukla (2014) has argued that we should abandon naturalistic and social constructivist considerations in attempts to define health due to their alleged failure to account for their normativity and instead define them purely in terms of 'social justice.' Here, I shall argue that such a purely normativist project is self-defeating, and hence, that health and disease cannot be grounded in social justice alone.

Keywords: Health; disease; conceptual engineering; conceptual analysis; naturalism.

1. *Introduction*

Within the last decade, the philosophy of medicine has largely moved on from hardened fronts between so-called naturalists, social constructivists, and normativists about how to define the concepts of health, disease, pathology, and the like. More and more authors are defending the possibility of hybrid accounts that keep what is best about other approaches (see Simon 2007; Kingma 2014; Powell and Scarffe 2019; Broadbent 2020; Conley and Glackin 2021), and yet, there hardly appears to be any progress in developing a consensus on how these notions should be defined.

In response, Quill Kukla (2014) [writing as Rebecca Kukla] has thus argued that we should abandon naturalistic and social constructivist considerations in attempts to define health due to their alleged failure to account for their *normativity* and instead define them purely in terms of 'social justice.' This makes her account one of the first explicit attempts for the conceptual revision and design of the concepts of health and disease for the sake of morality. Health and disease, Kukla argues, are intuitive and normative concepts and hence do not naturally fit with the explications by a "social constructionist understanding of

health, wherein health and disease are whatever we take them to be, and a scientistic understanding of health, wherein health and disease are biological concepts" (2014: 525). Instead, they should help us in the normative projects of deciding how health institutions should be designed and who deserves medical treatment. But as long as "we think that health has to be either a natural, biological category or a mere social construction" Kukla maintains that we cannot use the concept for normative purposes (2014: 525).

While I agree with the set-up of their argument—the sentiment that the naturalist-constructivist framing is a false dilemma, and the fact that the folk concept of health and disease has an explicitly normative dimension, I strongly disagree with their conclusion that the concepts of health and disease are to be designed as conditions that *should* or *ought* to be medicalized. Indeed, I shall here argue that any attempt to ground the concepts of health and disease in social justice alone must be self-defeating since it would eliminate their distinctiveness from other conditions of moral concern in addition to making the institution of medicine inevitably blur recognizably with all other institutions seeking to promote social justice.

Article Outline

This article is structured as follows: In Section 2, I expand on Kukla's sketch of the debate, explicating the three competing projects in the philosophical discussion on health and disease. In Section 3, I draw on a recent distinction between two kinds of conceptual engineering, utilizing them to show that the goals of these groups are ultimately irreconcilable. In Section 4, I use Kukla's proposal to explicate the idea of a purely normativist approach to health and disease. In Section 5, I argue that Kukla's social justice account of health—and for that matter any purely normativist account—must ultimately prove self-defeating. Finally, I conclude the discussion and respond to possible objections to my arguments in Section 6.

2. *Three Competing Projects*

Instead of framing the debate in the usual terms of conceptual analysis, regarding whether naturalism or social constructivism is correct, Kukla (2014) distinguishes between the different goals of both approaches. This is praiseworthy. Whereas theorists such as Boorse (1977) have attempted to capture health and disease in biomedical terms, appealing to the idea of *dysfunction* and *normal functioning* of a biological organism as it is used in medical practice, social constructivists such as Glackin (2010) have highlighted the importance of 'medicalization' within the social and institutional practices of medicine. During the medicalization of a condition, "clusters of symptoms are identified as unified diseases and brought under medical surveillance and manage-

ment" (Kukla 2014: 515). Both of these approaches have been met with much opposition, occupying much of the literature in a back-and-forth volley of counterexamples. Kukla (2014) thinks that both approaches are inherently misguided and thus ultimately fail even when some people attempt to provide hybrid accounts. This is because, Kukla argues, health is an "intuitive notion and not a technical term" (2014: 515) and should ultimately be used to inform policy and ethical decision-making regarding the treatment of those suffering from a disease.

So far, so good. But what does it mean to assert, as Kukla does, that this sense of health is 'intuitive?' Is it merely the idea that humans talked about health and disease *prior* to the arrival of modern medicine? Perhaps even prior to any form of medication? Since many species have been found to engage in grooming and self-medication behavior, such as the consumption of plants with the propensity to reduce or prevent harmful effects of pathogens and parasites (see Clayton and Wolfe 1993; Martin and Ewan 2008; de Roode et al. 2013; Neco et al. 2019), and this has been found to be especially prevalent in primates (see Huffman et al. 1997; Huffman 1997; Huffman and Hirata 2004), it is probable that our species, *Homo sapiens*, has always engaged in at least a minimal form of proto-medical practice. Perhaps Kukla intends to say that we don't need to know the biological basis or the causal underpinnings of injury and disease to recognize them as detriments to health.

Maybe Kukla's opening paragraph highlighting the 'intuitiveness' of health is thus intended to capture our corresponding folk concept. That is, in the words of Canguilhem, a different way of life:

> In the final analysis, would it not be appropriate to say that the pathological can be distinguished as such, that is, as an alteration of the normal state, only at the level of organic totality, and when it concerns man, at the level of conscious individual totality, where disease becomes a kind of evil? To be sick means that a man really lives another life, even in the biological sense of the word. (Canguilhem 1991: 87–88)

For Canguilhem the lived experience of disease came prior and he argued it should be central in our understanding of it. It is unclear, however, how this recognition necessarily lends itself to the social-justice based account of health and disease that Kukla has in mind. Indeed, it is unclear why the folk concept of health and disease must lend itself at all to Kukla's alternative project to locate the concepts of health and disease within what they call "social justice projects" (2014: 516) which roughly corresponds to what I have dubbed 'real normativism.' Capturing all of the intuitions associated with the folk concept within a single definition has proven to be exceedingly difficult, if not impossible. It is as if the concept has to do *too much* for a single definition to achieve all of these ends. There is plenty of reason to think that different approaches can focus on different components of the folk concept without thereby claiming that it must be *the* right approach. Pluralism may well be the right approach here.

Indeed, Kukla explicitly recognizes that despite the 'intuitiveness' of health and disease it has been far from straightforward to arrive at an agreed-upon definition. Nevertheless, to motivate their alternative approach, Kukla distinguishes two markedly different projects, one they call 'scientistic projects' and the other, as just noted, 'social justice projects.' Here, it is best to cite them in full:

> 1. *Scientistic* projects: The primary goal of such projects is to understand health and disease as respectable concepts from the point of view of the natural sciences. This is possible only if we can characterize what counts as a disease or a state of health independent of our specific, contingent social categories and practices. Such accounts avoid appeals to social or personal values, as these play no role in the categories and explanatory strategies of the natural sciences. Instead, they appeal to notions such as statistical normalcy, adaptive fitness, and biological function.

> 2. *Social justice* projects. In this context, an understanding of health and disease is a part of a specific type of *normative* project—namely, that of determining the role that health *should* play in a larger theory of social justice. Political philosophers, policy makers, and others ask questions such as: To what extent and in what sense is there a universal right to health, or health care? What counts as a fair social distribution of health resources? When does a health inequity count as a justice issue in need of moral redress? How shall we balance health needs with other social needs in a just state? To answer such questions, we need an understanding of what health is. But not any old understanding will do: This has to be the kind of understanding that will guide and clarify health policy and normative questions about the role of health care in a just society. (Kukla 2014: 515–516)

These naturally need not be the only projects, but it is perhaps possible to idealize and cluster many different projects under these two separate and broad headings. Because Kukla introduces their project by comparing naturalist and social constructivist approaches, however, readers might be misled into thinking that social justice projects map onto the latter. There is something slightly disingenuous about this false dichotomy, since we are thus invited to conclude that we have to either embrace the much-criticized naturalist accounts of health and disease such as that of Christopher Boorse (1977, 1997, 2014) or realize that the concept should ultimately be grounded in concerns of justice. For purposes of clarity, it is thus useful to sketch a third kind of project in this debate that we may analogously call a *social science* project:

> 3. *Social science* projects: The primary goal of such projects is to understand health and disease as concepts used by particular linguistic communities at a particular time and place in history (including the present). Here, contingent social categories and practices that have been deemed irrelevant in the naturalist project, play the central role. In these projects, homosexuality and drapetomania may be accurately called diseases at a particular time and place, even though they are no longer today. The concepts of health and disease are thus here unlike in the other two projects – relative to the norms of a society, depending on the social processes and mechanism of *medicalization*.

This three-fold distinction between three different projects will help us to map the terrain of goals in the philosophical debate on health and disease. Indeed, it strengthens Kukla's (2014) insight that one of the reasons why "the various attempts to define health and disease have been so unsatisfactory is that those using the notion are driven by deeply diverse theoretical and practical goals" (2014: 515). Unlike Kukla, however, I maintain that the different goals for which the concept has been put to use are *the* very reason for the lack of progress in the debate. This will become apparent once we turn away from traditional conceptual analysis and instead focus on conceptual engineering. In the next section, I draw on a recent distinction between two different kinds of conceptual engineering in order to cash out the futility of trying to achieve a satisfying definition of health and disease.

3. *Two Kinds of Conceptual Engineering*

Historically, *conceptual analysis* (i.e. the descriptive analysis of a concept) has been assumed to play the central, if not only, role in settling the philosophical debate on health and disease (Schwartz 2007; Lemoine 2013; Schwartz 2014). The goal was to arrive at a list of necessary and sufficient conditions that would allow us to tidy up the world into conditions that are diseases and those that are not. This approach might be expressed quite ambitiously as the search for the true meaning or more moderately as the search for "criteria of application" that people use when employing the concept (Neander 1991: 171). Neither of these goals, however, is particularly well suited for the application of conceptual analysis. Due to considerations of space, rather than arguing for it independently, I merely wish here to announce my alignment with those who have already argued that conceptual analysis within this debate is flawed and should be replaced with conceptual engineering (Schwartz 2007; Lemoine 2013; Schwartz 2014; Matthewson and Griffiths 2017; Griffiths and Matthewson 2018; Veit 2021a, 2021b).

Throughout the last decade, methodological debates about the tools and methods of philosophy itself have resurfaced.[1] The origins of this debate can be located in Sally Haslanger (2005), who argued that we should ameliorate our concepts, rather than just analyze them. Concepts ought to be 'engineered.' *Conceptual engineering* is focused on the purposes and goals a specific concept is intended to fulfill. This is just what we need in order to make progress in the philosophical debate on health and disease. But while conceptual engineering has been a core tool among philosophers since the very origins of the field (Burgess et al. 2020), philosophers have only recently begun to seriously engage in meta-philosophical discussions about the nature of this activity.

---

[1] Cf. Cappelen et al. (2016) and Sytsma and Buckwalter (2016).

In another essay coauthored with Heather Browning, I have made a distinction between two kinds of conceptual engineering that bear similarities to the two projects Kukla (2014) has sketched, although I do see them as quite a bit broader and have applied them to the various positions in the 'normativism vs. naturalism' debate (Veit and Browning 2020). The first of these, I have called *naturalist conceptual engineering* (NCE) (Veit and Browning 2020: 10):

> *Naturalist Conceptual Engineering* = (i) The scientific assessment of concepts, categories, and classificatory systems, (ii) determination of their relevant context and purposes to which they are and should be put to use, (iii) reflections on and proposal for how to improve them, and (iv) proposals for and active participation in the implementation of the suggested improvements.

NCE may appear quite familiar to anyone who is acquainted with Carnap's (1950) concept of 'explication,' yet, this understanding would narrow it down too much. The way Carnap understood explication, was to replace vague and inexact prescientific concepts with a new concept (within the context of science) that is to be made as precise as possible (Carnap 1950: 3). However, precision is not only desiderata for evaluating scientific concepts. Here, we simply need to pay attention of the goals of the scientists, which may change substantially depending on their discipline. Carnapian explication is thus merely one form of NCE, and we can simply allow for pluralism regarding the desiderata we can use to evaluate alternative concepts.

The important lesson here, is that this way of 'designing' concept contrasts strongly with the second kind, I have dubbed *moral conceptual engineering* (MCE):

> *Moral Conceptual Engineering* = (i) The moral, political, and social assessment of concepts, categories, and classificatory systems, (ii) determination of their relevant context and purposes to which they are and should be put to use, (iii) reflections on and proposal for how to improve them, and (iv) proposals for and active participation in the implementation of the suggested improvements. (Veit and Browning 2020: 9)

In the case of health and disease, these two kinds of conceptual engineering match well with two of the three projects outlined in Section 2. Indeed, they perhaps allow us to understand why Kukla (2014) didn't include those projects I called *social science* projects. Whereas what Kukla called *scientistic* projects and *social justice* projects design concepts for a particular purpose—i.e. they are ameliorative—the social science project is merely descriptive. There, we are merely interested with how a specific community uses or has used the term. This fits better with traditional conceptual analysis, or perhaps with some of the tools advocated by experimental philosophers.

Kukla's goal is ultimately MCE, i.e. the amelioration of the concepts of health and disease to serve the purposes of what they call social justice by furthering collective wellbeing. Other purely normativist accounts may target a different moral value, but they would nevertheless

still constitute MCE for being aimed at a moral end. As I shall argue, however, their own arguments may put a premature end to the very idea of this project. Indeed, Kukla recognizes "that there is no prima facie reason to think that our best attempts to specify a scientifically rigorous definition of health and our best attempts to specify a politically and normatively useful notion of health will correspond with one another" (2014: 516). Kukla expresses skepticism that health and disease can be expressed within unified concepts that would prove satisfactory with regards to the different goals to which the concepts are put to use. Once we have moved away from the traditional method of conceptual analysis we should become skeptical that they can be thought of as natural kinds or that there is anything like a single essence only waiting to be discovered by an ingenious philosopher. Kukla's opposition to this idea may stem from their endorsement of MCE.

In passing, they note that disease could possibly be understood as biological pathology from a scientific point of view. But this is not the project Kukla is engaged in, since they seem to endorse a variant of Canguilhem's view of medicine, with an appeal to the folk concept of health conditions as something that ought to be treated. Notice that this fits somewhat uneasily with their goal of conceptual engineering. After all, it is precisely the goal of refining the folk concepts of health and disease that drives attempts at a conceptual analysis of these notions. For instance, Kukla (2014) refers to the common idea that medicine as an *institution* is "designed, first and foremost, to promote, restore, and protect health" and that the "protection of health and distribution of health services is, almost all societies would agree, an important component of justice" (2014: 515). The patient, and their suffering, comes first. Unlike Canguilhem, however, Kukla's view is oriented not on the patient-doctor relationship but rather the collective relationship between humans and medicine as an institution, hence the emphasis on *social justice*. This emphasis, Kukla argues, may ultimately lead to a different perspective on health, such as "poor nutrition among low-income children" even when biological science treats it only as a state that is causally linked to actual diseases (Kukla 2014: 516).[2] However, this emphasis on only one aspect of how talk of "health" and "disease" is used, makes it unclear why the account would constitute why social constructivist or naturalist accounts must fail. If they have different goals, then it won't constitute a failure of such accounts to miss out on the context of these concepts within political and ethical decisions about which conditions should be treated. It is only through recourse on what Kukla perceives as the most important feature of folk discourse that such statements that other accounts are mistaken can be justified, but this would just make it one of many proposals of a

[2] The attested inadequacy of the naturalist position may be premature. Multiple authors (Griffiths and Matthewson 2018; Matthewson and Griffiths 2017; Veit 2021a) have argued that a naturalist account of health and disease may very well be able to account for categorising such states as pathological.

conceptual analysis of how we should *really* understand folk discourse. The problem for a conceptual analysis of elements of our folk psychology, is as Murphy (2012) notes, not only that they often lack coherence to begin—with ordinary language terms often used in very heterogenous ways that would make a conceptual analysis impossible—but also that conceptual analysis only tells us how ordinary people think about terms, not what their actual referent really consists in (2012: 22). The latter requires a scientific empirical analysis, which may undermine the original goals to make folk terms more precise entirely. If Kukla were to limit themselves to the purpose of conceptually engineering a health concept for the purposes of social justice, this worry could be avoided, but then we would also have to cease all talk of replacing naturalist and social constructivist accounts that simply have different purposes. However, as I shall shortly argue, even such a less ambitious version of Kukla's account will remain self-defeating.

I whole-heartedly agree with the suggestion that "in considering the best definition of health, we need to keep clearly in view the theoretical and practical purposes to which we want to put the concept, while keeping an open mind as to how unified a definition is possible" (Kukla 2014: 516). While NCE and MCE do not have to come apart, this will only be the case if the goals of each project are not in conflict. In the case of health and disease we should be skeptical that the widely different goals of the different parties can be satisfied with a single concept (see also Veit 2021b). Let us therefore examine Kukla's proposal for an account of health and disease that serves the purposes of social justice—an account that, as I shall argue, demonstrates that the very notion of a purely normativist account of health and disease must ultimately fail.

4. *Engineering 'Health' for Justice*

In Kukla's paper, we are presented with Boorse's (1977, 1997, 2014) biostatistical theory (BST) account as the paradigm example for what Kukla locates within the 'scientistic project.' The BST takes, as the name would lead one to expect, statistical normal function as the core of health. Normal functioning for Boorse concerns the body (both as parts and as a whole) of an individual within a particular population (class) in which "a statistically typical contribution by it to their individual survival and reproduction" (Boorse 1977: 555). Health, for Boorse, is merely the absence of disease, which in turn "reduces one or more functional abilities below typical efficiency" (Boorse 1977). While evolutionary concepts (survival and reproduction) play a role in the BST account, one should resist Kukla's appeal to classify Boorse's account as an evolutionary one—indeed, Boorse explicitly argues against the selected-effects view of functions and has argued that evolutionary biology has little to add to our understanding of health and disease (see Boorse 1976). Boorse's argument boils down to the alleged irrelevance

of the evolutionary history of traits to disease definitions within medicine. This, of course, can be granted—but is also shows why Boorse's account isn't really a naturalist one. While Boorse is interested in the medical usage of term 'disease,' rather than its folk usage, the methodology of trying to provide an analysis of the discourse of these terms (within a context) remains the same. A truly naturalist account must revise these notions in the light of scientific analysis, not mere social facts about language use in the medical profession. Boorse's account may thus be better classified as a social constructivist account that focuses on the concepts of health and disease as they are employed by the medical profession.[3] As I have argued elsewhere, a naturalistic understanding of health and pathology requires an understanding of the species' evolutionary history since this allows us to understand their design and health correspondingly as the optimal response to trade-offs from pathological complexity (Veit 2022, 2023; Veit and Browning 2021b).

By disassociating social constructivism from normativism, we can see that some of the problems of Boorse's account may stem from its uncomfortable hybrid role as both a naturalist account of health and disease and a social constructivist account of actual medical practice. These may obviously come apart. The way scientists conceptualize a concept and the target phenomena they are trying to capture can obviously be mismatched. And if the science is a value-laden one such as medicine, there are reasonable expectations that moral values may have slipped into the concept of disease. Since these various goals can take different shapes in their own right, there is little hope for thinking that there must be something like a uniquely correct concept of health and disease that would address all of these concerns.

An important, but often neglected point that Boorse (1997) once made, is that "there can be diseases that are neither disvaluable nor worthy of therapy" and conversely, "physicians can be justified in non-therapeutic activities. So the concepts of health and disease are far from settling all clinical or social questions" (1997: 99) even if this is often assumed and taken to be a substantive criticism of Boorse's account. Boorse thus emphatically denies that his project has anything to do with what I dubbed MCE. It is therefore, as Kukla (2014) recognizes, "explicitly devoid of normative force or practical upshot" (2014: 517). Any naturalist account that arises from NCE makes it impossible to simply assume, as Kukla notes, "that there are any ethical or practical implications that follow in any direct way from determining that something is a disease, or that someone (or some group of people) is (or is especially likely to be) in ill health" (2014: 517). This does not mean that a naturalist account cannot lead to normative facts, but rather that it cannot be *a priori* assumed that it will. And it is precisely this

---

[3] See also Griffiths and Matthewson (2018); Matthewson and Griffiths (2017); Veit (2021a).

reason why many have been dissatisfied with Boorse's analysis, yet it elegantly shows how MCE and NCE can pull in entirely different directions. But to require that health *must* somehow be conceptually linked to justice is, as we shall see, a poor argument even within a social justice project.

Firstly, we can deny that a concept such as disease conceptually entails some sort of moral right for treatment, while nevertheless recognizing that both for evolutionary and empirical reasons—pathological states are strongly linked to reductions in wellbeing, autonomy, and other 'intrinsically' important features of human (or for that matter, animal) life (Veit and Browning 2021a). It is hard to see why there must be a conceptual link between health and justice in something like an entailment relationship, as opposed to an empirical link via the bridging concepts of, say, wellbeing. Doctors, after all, frequently engage in procedures to improve the wellbeing of patients, regardless of whether their intervention is properly classified as the treatment of a disease and sometimes do so even at the cost of a patient's health such as the use of strong opioids. It is unclear why, *even if* health is intuitively a moral good, our best account of health and disease must turn this into a conceptual truism. Kukla's repeated emphasis of the folk concept of health is an odd move to say the least in a paper that attempts to use moral conceptual engineering, which allows for the possibility of a drastic change from the usual folk understanding of a term.

Yet, the goal to have an account of health and disease that satisfies both MCE and NCE is what motivated many in the debate to declare Boorse's account (and any other purely naturalist accounts) as inadequate. They maintain instead that we need something like Wakefield's (2001) hybrid account for the purposes of policy-making, in order to account for both sets of goals. Like Kukla (2014), I believe that such hybrid accounts will ultimately fail to provide consensus. The projects are undermined by the very idea that we can have a single concept that satisfies the demands of both MCE and NCE. While something like a equilibrium point is a theoretical possibility, it has rarely been attempted to make the trade-offs and conflicts between these two goals explicit. I have my doubts that we will ever create a consensus on the topic of how much weight should be given to moral and naturalist considerations.

This is not to say that hybrid accounts cannot be provided—indeed, I suspect that many of the accounts usually seen as naturalist or normativist turn out to be hybrid accounts once we make a more fine-grained distinction between naturalism, normativism, and social constructivism.[4] And these commitments can come in different gradations and varieties. Engelhardt (1986), for instance, is straightforwardly both a social constructivist and normativist. Nordenfelt (1993, 1995), however, while coming close to being a 'real' normativist in his

---

[4] Recall Boorse.

defense of a holistic account of health and disease, appears to (at least implicitly) allow some role for social constructivism due to his emphasis on the role of conceptual analysis, rather than conceptual engineering. But the mere fact that many, if not most, philosophers have in actuality defended hybrid accounts does not, of course, undermine the existence and usefulness of drawing the distinctions I made in Sections 2 and 3. These are distinctive projects and it is in principle possible to conceive of a purely descriptive account of how these terms are used within a linguistic community (although this may be the task for a social scientist or linguist rather than a philosopher), and the possibility of a purely naturalist conception of these terms to describe a natural phenomenon in, say, evolutionary dynamics between predators, prey, and pathogens. The problem with hybrid—unlike with pure—accounts is that there is no *one* standard on which to measure these accounts, since there is no *a priori* weighting that can be attached to the different goals for which the concepts is put to use.

However, Kukla moves from discussing the alleged failures of hybrid accounts to arguing that no truly normativist account—whether based on naturalist or social constructivist foundations—can be given. This is because such accounts fail to capture the social significance of whether conditions are medicalized, though as I argued previously, this may simply not be the goal of such accounts. If someone is interested in pure constructivism, they are simply engaged in social science and will define that a "condition or state counts as a disease if and only if it is medicalized, where medicalization is a social and institutional process, and health is the absence of disease" (Kukla 2014: 517). But the mere fact that conditions such as homosexuality or drapetomania were once seen as diseases provides us with no guidance of whether they *should* be seen as diseases, i.e. whether they should be cured.

As I argued in Section 3, the social constructivist is engaged in a descriptive project. Even when they are trying to provide a hybrid between a normativist (in the sense of justice) and a descriptive (social science) project—as for instance Engelhardt and Nordenfelt—the disparate goals between the two endeavours may pull even more strongly in opposite directions than they did between naturalism and normativism, that could at least plausibly be bridged through the concepts of wellbeing and autonomy. Glackin (2019), who comes close to something like a pure social constructivism, neverthelesss rejects Kukla's argument and sees it as "no objection to SOCIAL CONSTRUCTIVISM, or to any other normativist account of disease, that it does not provide us with an expedited route to socially just treatment of patients" since "no version of the concept is going to do that" (2019: 273). He argues, that if "we want social justice [...] we must do the hard, patient work of argument and advocacy for it; just agreeing on the descriptive facts will not be enough" (2019: 273). When Glackin speaks here of normativism, he has social contructivism in mind—an excellent showcase for why

the label normativism is confusing, since it denotes both the project of identifying what *is* being disvalued as a disease and the question of what *should* be cured or treated. While I agree with Glackin's opposition to the pure social justice project, he gives little argument for the claim that no version of the concept could possibly succeed at promoting social justice. Again, we are only presented with assertion and it is the goal of this paper to remedy this omission.

An elegant philosophical move made by Kukla was to turn social constructivism on its head by replacing the *what is being* medicalized component of social constructivism with a *what should be* medicalized ingredient.

> *The Institutional Definition of Health*: A condition or state counts as a *health condition* if and only if, given our resources and situation, it *would be best for our collective wellbeing* if it were medicalized—that is, if health professionals and institutions played a substantial role in understanding, identifying, managing and/or mitigating it. In turn, *health* is a relative absence of health conditions (and concomitantly a relative lack of dependence upon the institutions of medicine). (Kukla 2014: 526)

This account has obvious appeal, as it denies the naturalist treatment of homosexuality as a disease and the social constructivist treatment of drapetomania or masturbation as diseases in the past. It seems to be able to treat these judgements as mistakes, without appealing to ad-hoc additions of value criteria. As Kukla puts it, the "connection to justice is *built in* [...] from the start" (2014: 529). But there are number of decisive arguments against this approach, that ultimately undermine the very goal of the moral normativist to offer an alternative account of health and disease.

5. *Why Pure Normativism is Self-Defeating*

While Kukla's arguments are a welcome contribution from the anti-naturalist side and expose many of the underlying conceptual problems in the debate, Kukla's proposal is ultimately more flawed than the accounts they have criticized. Rather than give up on the idea that the notions of health and disease must intrinsically be valued or (dis)valued, Kukla (2014) seeks to detach the concepts of health and disease from their intended targets within both biology and ordinary discourse, instead labelling them as whatever would contribute to social justice if it were medicalized. But the problem with treating statistically abnormal sexual preferences such as homosexuality or 'gender identity disorder' as mental disorders is not a mis-characterization of biological reality per se, but the empirical fact that medicalization has the unintended side-effect of treating these conditions as 'bad'—as something that should be cured, something that it would be better not to have.

Kukla (2014), instead of abolishing this problematic part of the folk concept of health and disease—one that has been criticized by utilitarians and disability rights advocates alike as something that should not

intrinsically matter—embraces it and discards any underlying biological or social phenomena. One should immediately be worried as to why it is the normativity, rather than say the naturalness of the folk concept of health that should be our focus. Idealizing away all factors aside from the moral role of these concepts is, of course, only a move worth making if the underlying goal of the pure normativist to define health exclusively in terms of justice could thus be better promoted. But is this actually an instance of MCE? Kukla argues that their account (or at least a purely normativist account of some kind) must be right since there is an asserted intrinsic association in people's minds between the abnormal and the 'bad.' But the mere fact that an empirical study or conceptual analysis of the common usage of these terms would reveal a normative component is irrelevant for the conceptual engineer interested in revising the concept for a specific goal. We could equally take the naturalist route that revises the concept in a way such that there is no longer a conceptual connection between what is called a disease and what should be treated. Despite appealing to the goals of conceptual engineering, Kukla falls prey to the old ideals of conceptual analysis.

The resulting problem is precisely what Kukla has criticized hybrid accounts for: they fail to carve nature such that disease constitutes a special moral domain. Indeed, this is precisely what numerous bioethicists in the enhancement literature have argued for: there is no important moral distinction between the treatment of a disease and an enhancement beyond what is typically considered healthy (Savulescu et al. 2011; Veit 2018b, 2018a). Both methods enhance human bodies to promote the wellbeing of the patient; whether the underlying condition is understood as a disease or not is irrelevant. Note that this is MCE, without proposing a new definition of health. They simply maintain that we should use different criteria, such as autonomy and wellbeing, when making medical decisions. Our *collective wellbeing* could be promoted in all kinds of ways by medicalizing certain states: think of hair loss in old age and many other conditions that are perhaps unfortunately left untreated because they are a natural result of the aging process. Since Kukla (2014) gives up the dysfunction criteria of disease, many conditions that aren't currently treated by medical practitioners, on the sole ground that they don't constitute actual diseases, would have to be reevaluated. This would naturally lead to a radical revision of current medical practice. But here I want to step in: why then keep the concepts of health and disease at all? What is gained by keeping these terms? Why do we need this intermediary concept between facts about the body and concerns of justice, if medical professionals are now simply in the 'business' of using the current tools of medicine for the promotion of what Kukla calls *social justice*? In fact, Kukla appears unaware that their own argument would lead to a slippery slope that is ultimately self-defeating. Let us spell this important point out in more detail.

Why is a purely normativist account bound to fail? Kukla's account provides a beautiful example for why those interested in justice cannot simply define health, disease, and pathology in terms of moral concerns. The problem lies in the connection between the institution of medicine and the concepts of health and disease. Let us for the sake of the argument assume that medical practice, medical practicioners, and the tools of medicine are simply a given. Those like Kukla, or us for that matter, who are concerned that medical institutions can misuse their authority to promote unwanted goals such as racism or homophobia. It is historically well documented, for instance, that homosexuals have been discriminated against on grounds of living a supposedly 'unnatural' life-style, something that was assumed must be pathological since it lowers one's fitness. Homophobia has been justified by hiding behind the veil of medical authority. Neither the naturalist nor the social constructivist account of health and disease *seems* to offer much to prevent such misuse. This is why Kukla wants to put the normative component of health and disease centre stage—eliminating the need for any naturalist or social constructivist basis of health. There is an intuitive appeal to the idea that we should simply look at our institution of medicine and then think about which conditions *should* be considered diseases or health-problems in order to promote social justice.

The first major problem is this: Kukla leaves social justice entirely undefined, treating it loosely as some concept of collective wellbeing. Indeed, Kukla responds to this possible criticism by treating it as a *strength* of their account: "[w]hether one is a consequentialist, a libertarian, a Rawlsian, or whatever else, one can be invested in what we have called the normative project of figuring out how a just state should manage health policy and health needs, and our definition of health can be slotted into any such project" (2014: 526). But this apparent strength of flexibility weakens the account. It amounts to little more than the unhelpful statement that we should define our terms in a way that promotes justice—whatever it is. The conceptual possibility proof is philosophically useful in terms of further exploring the conceptual possibility space, but it is pragmatically useless, since our modern societies obviously do not consist of a homogenous group in which everyone agrees about what justice should entail. If we accepted Kukla's proposal, the very concepts of health and disease would become another battleground for those with widely different moral views. Despite aiming to accommodate the apparent 'failures' of the naturalist and social constructivist to condemn the medicalization of homosexuality and drapetomania, Kukla does, in fact, do the opposite.

Consider for instance a society in which strict conservative religious views are in the majority, leading to the medicalization of attitudes like an unwillingess to bear the child of one's rapist, the desire to love someone of the opposite sex, and the opposition to the dominant religion, classified as mental disorders on the grounds of 'collective wellbeing.' It

is thus not hard to imagine that Kukla's own proposal would be used to justify the very things they aimed to condemn. Naturalists and social constructivists, on the other hand, can simply criticize old definitions of health as having been biased by the moral views of those that endorsed the medicalization of drapetomania. This I do not see as a failure, any more than the biological definition of a human being may fail to provide animals with human rights.[5] Leaving the content of social justice empty is thus a weakness of Kukla's account, not a strength.

Furthermore, I simply do not see how such a world would be preferable to our current one, in which our institutional definitions for health and disease are widely shared and pragmatically accepted among many as something that deserves treatment—not because justice is somehow built into these concepts, but simply because we *know* that biological wrongs are highly correlated with losses in autonomy, agency, and wellbeing. So unless we were to live in a world where everyone shares the same concept of justice, it would appear that the institutional definition of health Kukla proposes would surprisingly fail to promote the goals of collective wellbeing—even though this was precisely the *one* goal it was supposed to achieve. Indeed, we may simply be better off by accepting that we should respond to those conditions that lead to losses in wellbeing, regardless of whether these are diseases—something that can almost universally be agreed upon regardless of one's ultimate view on justice.

The first problem also emphasizes a larger problem that any purely normativist account of health and disease will share: a failure to ground health and disease states as distinctive from other states of moral importance and concern. To explicate this second and much more fundamental problem that underlies the motivation of this paper, let us assume for the moment that we had a universally agreed upon definition of collective wellbeing and social justice. In that case the institutional account of health would inevitably classify as a health condition all and only those things that are perceived to be something social justice should address. The institutional account is incoherent because it fails to recognize that institutions are inherently flexible and can change over time, changes that would lead to excessive broadening of the concepts beyond the point of usefulness.

Let me elaborate on my argument in more detail: If one approaches medicine from the perspective of justice, it is natural to ask what the tools of medicine should be used for. If one then rejects any connection between medicine and disease as a natural phenomenon, it becomes tempting to argue that medicine should treat and classify conditions as diseases if their medicalization by the institution of medicine would benefit collective wellbeing. What makes this suggestion incoherent is a neglect of the simple fact that the tools of medicine have been designed, first and foremost, to deal with diseases. These tools are con-

[5] Although animal rights advocates may disagree.

tinuously improved and expanded for that very purpose. We no longer use outdated practices because they have been shown to be flawed or replaced by better ones. If we now classify any possible condition that current medical tools could address to improve collective wellbeing as a health disorder, the tools of medicine will inevitably shift to become better at dealing with those conditions. In fact, Kukla's institutional account underestimates the plurality of medical practice that already exists: many of the current tools used by medical practitioners can and often are used to improve people's lives. One only needs to think of cosmetic surgeries or mood-enhancing drugs, regardless of whether these conditions are classified as health disorders.

Nevertheless, medical research has historically been tied to a biomedical, rather than a social justice, understanding of health and has constrained its scope accordingly. Kukla's proposal not only changes the definition of health conditions into something much more flexible but also alters the very institution of medicine. The tools of medicine will evolve to better address concerns of justice, and more and more states we deem concerns of justice will thus be classified as health conditions—precisely because the medical toolkit will inevitably expand to address these concerns. Unless Kukla (2014) insists that the current definition of medicalization remains fixed, medical practice would ultimately co-evolve into the practice of 'social justice promotion,' thereby losing the distinctiveness that the concept of health and disease is supposed to capture. This is a highly unattractive proposal because it would turn anything seen as an injustice into a health condition. Justice would no longer be built into the concept of health: health would simply become justice. The very criticism Kukla applied to hybrid accounts, i.e. that they fail to demarcate a unique normative role for these concepts, appears to apply even more forcefully to pure normativism.

It would lead to Rudolf Virchow's famous dictum that "politics is nothing else but medicine on a large scale"[6] except for the qualifier *on a large scale* being eliminated. Surely, such a result must be considered self-defeating, yet what could possibly stop it unless we draw on either naturalist or social constructivist resources to constrain what can legitimately be considered within the domain of medicine? Pure normativism can be "pure" in name only. It must rely on some grounding in one of the other frameworks.

As a result, an application of MCE, rather than explicating a novel concept of health and disease, would lead us to draw on either naturalism, social constructivism, or both, resulting in something like an implicitly hybrid account. Another approach would be to reject this entirely and, as some bioethicists suggest, focus on wellbeing and autonomy instead of health. I maintain that what Kukla has demonstrated with their account is not that we should build moral values into the concepts of health and disease, but rather that this part of our folk conception is

[6] See Ashton (2006).

no less problematic than the intuitions driving the naturalist or social constructivist in their accounts. The popular notion that health and disease are intrinsically moral concepts serves neither the goals of the naturalist nor, as I have demonstrated here, those of the normativist.

Before concluding this article, let us consider whether an amendment could salvage Kukla's view. The most apparent solution is to impose constraints on either (i) the institution of medicine or (ii) a purely normativist account of health.

We may, for instance, postulate that the tools of medicine are to remain fixed. This would arguably stop the slippery slope of health conditions becoming a meaningless term. However, this approach is also untenable due to the very goals of the account. If we seek to improve medicine with social justice and our collective wellbeing in mind, we would certainly want new medical tools to be developed, e.g. better forms of hormonal replacement therapies for transgender individuals seeking to transition.

Adding naturalist considerations to a social justice criterion, e.g. a selected-effects view of dysfunction, may seem like a plausible solution and one I believe holds promise. Yet, this approach also faces a fundamental problem since Kukla's entire argument for their account rests on the idea that hybrid accounts are doomed to failure. Thus, both amendments seem to only further strengthen my argument that Kukla's account is self-defeating. If it is "impossible to build a normative, social justice project on top of a scientistic conception of health and disease" (Kukla 2014: 519), then perhaps we should abandon the assumption that these concepts must have this kind of normativity built into them from the start.

6. *Conclusion*

The idea that the concepts of health and disease can serve the goals of naturalism, social science, and justice is ambitious, to say the least. Decades of debate should make us wary of thinking that there is a single concept waiting for philosophers to discover, one that preserves all of its 'intuitively' compelling properties. Kukla's article highlights an insoluble dilemma within the concepts of health and disease. Naturalists, social constructivists, and moral normativists simply have different goals for how these concepts are to be used. An obvious solution, then, is to embrace a more pluralist view, in which there could be at least three alternative accounts of health and disease, corresponding to each of these projects.[7] Purely naturalist and social constructivist accounts have been proposed in the past, but they have gained very little traction. This raises the question of whether a pure normativism could be more compelling. Kukla's institutional definition of health is

---

[7] Similar debates of medical vs social models occur in discussion on disability (Browning and Veit 2025).

one of the first attempts to achieve this through conceptual engineering rather than conceptual analysis. However, what this article aimed to demonstrate is that a purely normativist account will ultimately prove self-defeating, despite its perhaps 'intuitive' appeal.

The first reason Kukla's account may appear intuitively compelling as a purely normativist account of health and disease is the deliberate refusal to define what 'collective well-being' or 'social justice' are—terms that sound nice but will inevitably cause much more conceptual disagreement than the old debate about the proper conceptual analysis of health and disease. Secondly, there is a neglect to admit that for the account to work, the institution of medicine itself would have to be held fixed, thus making it a purely normativist account in name only.

The first problem makes Kukla's definition an excellent model to demonstrate that the arguments presented here will undermine any purely normativist account unless there is complete moral consensus. However, this is hardly a feature worth wanting if one is interested in defending such an account. For the purposes of MCE, we may very well also want to engage in some NCE or social construction to constrain these concepts or replace them with alternative notions such as wellbeing and autonomy to ground moral decision-making. For health and disease states to matter, it would be sufficient if these states have some empirical, rather than conceptual, link with those notions. To assume that something must intrinsically matter in order to be morally relevant, or to be used in decisions regarding public policy, is nothing more than an illusion.

The second problem reveals the self-defeating nature of the very idea of a purely normativist approach. If we ask what the unique tools and institution of medicine are, we must do so through recourse to our concepts of 'health' and 'disease.' This is precisely why this question has remained at the heart of the philosophy of medicine. But unless these concepts can somehow be held fixed or constrained through naturalist or social constructivist means, medicine would simply become whatever promotes justice. Yet a definition that fails to distinguish health from justice and disease from injustice can hardly be considered a definition at all. Pure normativism must fail, but this does not mean that the concepts of health and disease cannot play an important role within moral deliberation and public policy, nor that social justice considerations should play no role in these decisions.

References

Ashton, J. 2006. "Virchow Misquoted, Part-Quoted, and the Real McCoy." *Journal of Epidemiology & Community Health* 60 (8): 671.

Boorse, C. 1976. "Wright on Functions." *The Philosophical Review* 85 (1): 70–86.

Boorse, C. 1977. "Health as a Theoretical Concept." *Philosophy of Science* 44 (4): 542–573.

Boorse, C. 1997. "A Rebuttal on Health." In J. M. Humber and R. F. Almeder (eds.). *What is Disease?* Totowa: Humana Press, 1–134.

Boorse, C. 2014. "A Second Rebuttal on Health." *Journal of Medicine and Philosophy* 39 (6): 683–724.

Broadbent, A. 2020. "Health as a Secondary Property." *British Journal for the Philosophy of Science* 70 (2): 609-627.

Browning, H. and Veit, W. 2025. "The Neurodiversity Model and Medical Model: Competitors or Alternative Perspectives?" *AJOB Neuroscience* 16 (1): 32–34. https://doi.org/10.1080/21507740.2024.2437994

Burgess, A., Cappelen, H. and Plunkett, D. 2020. *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Canguilhem, G. 1991. *The Normal and the Pathological*. Trans. C. R. Fawcett. New York: Zone Books.

Cappelen, H., Gendler, T. and Hawthorne, J. P. 2016. *The Oxford Handbook of Philosophical Methodology*. Oxford: Oxford University Press.

Carnap, R. 1950. *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Clayton, D. H. and Wolfe, N. D. 1993. "The Adaptive Significance of Self-Medication." *Trends in Ecology & Evolution* 8 (2): 60–63.

Conley, B. A. and Glackin, S. N. 2021. "How to Be a Naturalist and a Social Constructivist about Diseases." *Philosophy of Medicine* 2 (1): 1–21.

Daniels, N. 2007. *Just Health: Meeting Health Needs Fairly*. Cambridge: Cambridge University Press.

de Roode, J. C., Lefèvre, T., and Hunter, M. D. 2013. "Self-Medication in Animals." *Science* 340 (6129): 150–151.

Engelhardt, H. T. 1986. *The Foundations of Bioethics*. Oxford: Oxford University Press.

Glackin, S. N. 2010. "Tolerance and Illness: The Politics of Medical and Psychiatric Classification." *Journal of Medicine and Philosophy* 35 (4): 449–465.

Glackin, S. N. 2019. "Grounded Disease: Constructing the Social from the Biological in Medicine." *The Philosophical Quarterly* 69 (275): 258–276.

Griffiths, P. E. and Matthewson, J. 2018. "Evolution, Dysfunction, and Disease: A Reappraisal." *The British Journal for the Philosophy of Science* 69 (2): 301–327.

Haslanger, S. 2005. "What Are We Talking About? The Semantics and Politics of Social Kinds." *Hypatia* 20 (4): 10–26.

Huffman, M. A. 1997. "Current Evidence for Self-Medication in Primates: A Multidisciplinary Perspective." *American Journal of Physical Anthropology* 104 (S25): 171–200.

Huffman, M. A., Gotoh, S., Turner, L. A., Hamai, M. and Yoshida, K. 1997. "Seasonal Trends in Intestinal Nematode Infection and Medicinal Plant

Use among Chimpanzees in the Mahale Mountains, Tanzania." *Primates* 38 (2): 111–125.

Huffman, M. A., and Hirata, S. 2004. "An Experimental Study of Leaf Swallowing in Captive Chimpanzees: Insights into the Origin of a Self-Medicative Behavior and the Role of Social Learning." *Primates* 45 (2): 113–118.

Kingma, E. 2014. "Naturalism about Health and Disease: Adding Nuance for Progress." *Journal of Medicine and Philosophy* 39: 590–608.

Kukla, R. 2014. "Medicalization, 'Normal Function,' and the Definition of Health." In J. D. Arras, E. Fenton and R. Kukla (eds.). *The Routledge Companion to Bioethics*. London: Routledge, 539–554.

Lemoine, M. 2013. "Defining Disease beyond Conceptual Analysis: An Analysis of Conceptual Analysis in Philosophy of Medicine." *Theoretical Medicine and Bioethics* 34 (4): 309–325.

Martin, T. J. and Ewan, E. 2008. "Chronic Pain Alters Drug Self-Administration: Implications for Addiction and Pain Mechanisms." *Experimental and Clinical Psychopharmacology* 16 (5): 357–366.

Matthewson, J. and Griffiths, P. E. 2017. "Biological Criteria of Disease: Four Ways of Going Wrong." *Journal of Medicine and Philosophy* 42 (4): 447–466.

Murphy, D. 2012. *Psychiatry in the Scientific Image*. Cambridge, MA: MIT Press.

Neander, K. 1991. "Functions as Selected Effects: The Conceptual Analyst's Defense." *Philosophy of Science* 58 (2): 168–184.

Neco, L. C., Abelson, E. S., Brown, A., Natterson-Horowitz, B. and Blumstein, D. T. 2019. "The Evolution of Self-Medication Behaviour in Mammals." *Biological Journal of the Linnean Society*: 373–378.

Nordenfelt, L. 1993. *Quality of Life, Health and Happiness*. Aldershot: Ashgate Publishing.

Nordenfelt, L. 1995. *On the Nature of Health: An Action-Theoretic Approach*. Dordrecht: Kluwer.

Powell, R. and Scarffe, E. 2019. "Rehabilitating 'Disease': Function, Value, and Objectivity in Medicine." *Philosophy of Science* 86 (5): 1168–1178.

Powers, M. and Faden, R. R. 1999. *Social Justice: The Moral Foundations of Public Health and Health Policy*. Oxford: Oxford University Press.

Savulescu, J., Ter Meulen, R., and Kahane, G. 2011. *Enhancing Human Capacities*. Chichester: John Wiley & Sons.

Schwartz, P. H. 2007. "Decision and Discovery in Defining 'Disease'." In H. Kincaid and J. McKitrick (eds.). *Establishing Medical Reality: Essays in the Metaphysics and Epistemology of Biomedical Science*. Dordrecht: Springer, 47–64.

Schwartz, P. H. 2014. "Reframing the Disease Debate and Defending the Biostatistical Theory." *Journal of Medicine and Philosophy* 39 (6): 572–589.

Simon, J. 2007. "Beyond Naturalism and Normativism: Reconceiving the 'Disease' Debate." *Philosophical Papers* 36 (3): 343–370.

Sytsma, J. and Buckwalter, W. 2016. *A Companion to Experimental Philosophy*. Chichester: John Wiley & Sons.

Veit, W. 2018a. "Cognitive Enhancement and the Threat of Inequality." *Journal of Cognitive Enhancement* 2 (4): 404–410.

Veit, W. 2018b. "Procreative Beneficence and Genetic Enhancement." *Kriterion – Journal of Philosophy* 32 (1): 75–92.

Veit, W. 2021a. "Biological Normativity: A New Hope for Naturalism?" *Medicine, Health Care and Philosophy* 24 (2): 291–301.

Veit, W. 2021b. "Experimental Philosophy of Medicine and the Concepts of Health and Disease." *Theoretical Medicine and Bioethics*. https://doi.org/10.1007/s11017-021-09550-3

Veit, W. 2022. *Health, Agency, and the Evolution of Consciousness*. PhD diss., University of Sydney. https://hdl.handle.net/2123/29836

Veit, W. 2023. "Health, Consciousness, and the Evolution of Subjects." *Synthese*. https://doi.org/10.1007/s11229-022-03998-z

Veit, W. and Browning, H. 2020. "Two Kinds of Conceptual Engineering." *Preprint*. https://philsci-archive.pitt.edu/17452/

Veit, W. and Browning, H. 2021a. "Phenomenology Applied to Animal Health and Suffering." In S. Ferrarello (ed.). *Phenomenology of Bioethics: Technoethics and Lived-Experience*. Cham: Springer, 73–88.

Veit, W., and Browning, H. 2021b. "Developmental Programming, Evolution, and Animal Welfare: A Case for Evolutionary Veterinary Science." *Journal of Applied Animal Welfare Science*: 552–564.

Wakefield, J. C. 2001. "Evolutionary History versus Current Causal Role in the Definition of Disorder: Reply to McNally." *Behaviour Research and Therapy* 39 (3): 347–366.