

**Volume X ♦ Issue 4**

2019

# Logos & Episteme

an international journal  
of epistemology

**Romanian Academy  
Iasi Branch**



**“Gheorghe Zane” Institute  
for Economic and Social  
Research**

## **Founding Editor**

---

Teodor Dima (1939-2019)

## **Editorial Board**

---

### **Editor-in-Chief**

Eugen Huzum

### **Executive Editors**

Vasile Pleșca

Cătălina-Daniela Răducu

### **Assistant Editors**

Irina Frasin

Bogdan Ștefanachi

Ioan Alexandru Tofan

### **Web&Graphics**

Codrin Dinu Vasiliu

Virgil-Constantin Fătu

Simona-Roxana Ulman

## **Contact address:**

---

Institutul de Cercetări

Economice și Sociale „Gh.Zane”

Iași, str.T.Codrescu, nr.2, cod 700481

Tel/Fax: 004 0332 408922

Email: [logosandepisteme@yahoo.com](mailto:logosandepisteme@yahoo.com)

<http://logos-and-episteme.acadiasi.ro/>

[https://www.pdcnet.org/pdc/bvdb.nsf/journal?openform&journal=pdc\\_logos-episteme](https://www.pdcnet.org/pdc/bvdb.nsf/journal?openform&journal=pdc_logos-episteme)

## **Advisory Board**

---

Frederick R. Adams

University of Delaware, USA

Scott F. Aikin

Vanderbilt University, USA

Daniel Andler

Université Paris-Sorbonne, Paris IV, France

Alexandru Boboc

Academia Română, România

Panayot Butchvarov

University of Iowa, USA

Mircea Dumitru

Universitatea din București, România

Sanford Goldberg

Northwestern University, Evanston, USA

Alvin I. Goldman

Rutgers, The State University of New Jersey, USA

Susan Haack

University of Miami, USA

Stephen Hetherington

The University of New South Wales, Sydney, Australia

Paul Humphreys

University of Virginia, USA

Jonathan L. Kvanvig

Baylor University, USA

Thierry Martin

Université de Franche-Comté, Besançon, France

Jürgen Mittelstrab

Universität Konstanz, Germany

Christian Möckel

Humboldt-Universität zu Berlin, Germany

Maryvonne Perrot

Université de Bourgogne, Dijon, France

Olga Maria Pombo-Martins

Universidade de Lisboa, Portugal

Duncan Pritchard

University of Edinburgh, United Kingdom

Nicolas Rescher

University of Pittsburgh, USA

Rahman Shahid

Université Lille 3, France

Ernest Sosa

Rutgers University, USA

Alexandru Surdu

Academia Română, România

John F. Symons

University of Texas at El Paso, USA

# TABLE OF CONTENTS

## RESEARCH ARTICLES

Scott AIKIN, Brian RIBEIRO, Skeptical Theism and the Creep Problem.....	349
Tommaso OSTILLIO, Michal BUKAT, The Knobe Effect with Probable Outcomes and Availability Heuristic Triggers.....	363
Seungbae PARK, Surrealism Is Not an Alternative to Scientific Realism.....	379
Hans ROTT, Unstable Knowledge, Unstable Belief.....	395
Michael J. SHAFFER, The Availability Heuristic and Inference to the Best Explanation.....	409
Xintong WEI, The Permissible Norm of Truth and “Ought Implies Can”.....	433

## DISCUSSION NOTES/ DEBATE

Moti MIZRAHI, Factivity and Epistemic Certainty: A Reply to Sankey.....	443
Howard SANKEY, Why Must Justification Guarantee Truth? Reply to Mizrahi.....	445
James SIMPSON, Knowledge Doesn’t Require Epistemic Certainty: A Reply to Mizrahi.....	449
Notes on the Contributors.....	451
<i>Logos and Episteme</i> . Aims and Scope.....	455
Notes to Contributors.....	457



## RESEARCH ARTICLES



# SKEPTICAL THEISM AND THE CREEP PROBLEM

Scott AIKIN, Brian RIBEIRO

**ABSTRACT:** Skeptical theism is the view that human knowledge and understanding are severely limited, compared to that of the divine. The view is deployed as an undercutting defeater for evidential arguments from evil. However, skeptical theism has broader skeptical consequences than those for the argument from evil. The epistemic principles of this skeptical creep are identified and shown to be on the road to global skepticism.

**KEYWORDS:** problem of evil, skeptical theism, total evidence skepticism,  
global skepticism

## Introduction

Skeptical theism is deployed to undercut evidential arguments from evil. It is the view that when we consider the problem of evil, we have no good reason to believe that our conception of goods, evils, and relevance relations between them are representative of what God would consider when He permits, what seem to humans, gratuitous evils. Our view is that skeptical theism has a problem with what we call *skeptical creep* – namely, that the skeptical consequences of the view spread beyond the domain of the evidential problem of evil to theology, moral knowledge, and then at last to become a global skeptical problem. Theological and moral skeptical creep has been widely noted. Our objective is to show that a particular dialectical requirement for justification is behind the skeptical theist's challenge, and this requirement yields the creep phenomenon not only extending to theological and moral knowledge, but also to knowledge in general.

## 1. Skeptical Theism and the Problem of Evil

Skeptical theism is, in its primary instance, a dialectical view. The fact of gratuitous evils, or better put, cases of *prima facie* senseless suffering, is a problem for traditional theism. How could a God worthy of the name permit them? And so an evidential version of the argument from evil arises. It can be stated roughly as follows:

1. There are instances of evil that God could have prevented without losing some greater good or failing to prevent some greater evil.

2. If God exists, He would prevent instances of evil unless He could not do so without losing some greater good or failing to prevent some greater evil.

3. *Therefore*, God does not exist.<sup>1</sup>

The matter of import is what justification we have in believing the first premise. There appear to be many cases of suffering that confirm it, where we cannot, for all our attempts, arrive at a sufficiently satisfying reason for why God would permit them. Call the move from the breadth of what seems to be senseless suffering and our inability to think of what would justify it to Premise 1 *the inference*. Its basic form is:

*Since we humans cannot discern a justifying reason for God to allow evils, there is none.*

This is where skeptical theism plays its dialectical role. Skeptical theism is the view that we humans are significantly cognitively limited. We are so limited, especially in comparison to the divine, *the inference* is manifestly fallacious. Small children, by analogy, may hold that there is no good reason for shots or rules against eating crayons, but there clearly are. The fact that they cannot come up with them on their own is itself not a good reason to hold there are no reasons. And we, limited and fallen creatures we are, are more like children before God. His ways are not our ways, we are reminded. And so, given the way theists conceive of the gap between God's intellectual powers and ours, there are presumably many, many things He conceives and knows that we humans are in no position to know or even understand. In light of the gulf between ourselves and God, it should come as no surprise that there are events that we cannot see reason for, but for God there is perfect reason.

Notice that skeptical theism, in fact, is a reply to two coordinate problems for theism. On the one hand, it is a reply to the first-order problem of evil—that there *only seems to us* to be senseless suffering. On the other hand, skeptical theism handles the second-order problem of the long track record of failed theodicies—in particular that failed theodicy is *our* failure to understand God, not God's failure to be just. Both the fact of evils we can't see reason for and the consistent failure of theodicies seem to be evidence against theism, but the skeptical theist's move is to show that this commitment is not justified. The result,

---

<sup>1</sup> This is a modified version of the evidential argument from William Rowe and is widely glossed as the basic form of the argument. We have provided an antecedent in premise 2 to make the argument more obviously valid. See William Rowe, "The Problem of Evil and Some Varieties of Atheism," in *The Evidential Argument from Evil*, ed. Daniel Howard-Snyder (Indianapolis: Indiana University Press, 1996), 1-11.



then, is that skeptical theism's prime dialectical role is that of being a defeating consideration for a crucial premise in the argument from evil. It requires that we be skeptical about our capacities to determine what considerations would warrant God permitting evils. And as a consequence, the failure of theodicy is, too, rendered inert as evidence against theism—it is perfectly consistent with a traditional notion of God.

## 2. Skeptical Theism's Dialectical Role

Skeptical theism's dialectical role is to provide a defeating reason for our justification for believing that there are no reasons warranting God's allowing evils on the basis of there not being any we can access. The operative question is *what epistemic principle yields that defeat?*

Taking Bergmann's version of skeptical theism as exemplary, the core of skeptical theism is the three-part commitment:

- (ST) We have no good reason for thinking that the (i) possible goods, (ii) possible evils, and (iii) entailment and permission relations between goods and evils that we know of are representative of all the possible goods, evils, and permissibility relations there are.

According to skeptical theists, ST provides defeat for *the inference*. As Bergmann puts it, "we can't use our failure to think of a God-justifying reason for permitting horrendous evils... to conclude that it is unlikely that there is such a reason."<sup>2</sup>

The question, again, is how ST defeats *the inference*. At the core of ST is the relation of *representativeness*. This relation can be strict or approximate. Approximate representative samples give us information about a target class with an acceptable margin of error. So if sample A is representative of class B, then if x percent of A is F, then *approximately* x percent of B is F. Strict representativeness, however, has no margin of error. So, assuming strict representation, if x percent of A is F, then *exactly* x percent of B is F. This distinction of kinds of representativeness is important for the skeptical theist, because *the inference* requires the percentages of zero percent to be identical between the sample and target classes. So the more modest version of *the inference*

There are no known God-justifying reasons for evil

*Therefore*, there are *approximately* no God-justifying reasons for evil.

---

<sup>2</sup> Michael Bergmann, "Skeptical Theism and the Problem of Evil," in *The Oxford Handbook of Philosophical Theology*, eds. T. Flint and M. Rea (New York: Oxford University Press, 2009), 378.

would be unacceptable, because it only takes one instance to defeat the argument. Rather, what's necessary is the more strictly representative relation:

There are no known God-justifying reasons for evil

*Therefore*, there are *exactly* no God-justifying reasons for evil.

ST is a defeater for the evidential argument from evil only if *the inference* requires strict representativeness; which it seems, it must.<sup>3</sup> We can, then, state the principle that yields the defeat as follows:

- (D) If S infers *n* (exactly no B are F) from *m* (no A are F), then S has justification for *n* only if S has reason to hold sample A is strictly representative of class B.

Again, given our assumptions about the cognitive gulf between God and humans, we do not have reason to hold that the justifying reasons we know to fail are strictly representative of all the possible God-justifying reasons. And so *the inference*, it is held, is defeated.

The problem is that D seems exceedingly strong; moreover, it does not reflect ST's dialectical deployment. D is too strong, because it seems to prohibit *any* negative generalization (e.g., inferring that there are no cats in my office from a survey of where I usually see them); moreover, it fails to reflect the fact that ST is being deployed against an atheist's argument to a theist. This is because the theist will have a commitment to the great cognitive gap between humans and God. So it fails to be dialectical, in the sense that it doesn't meet its argumentative opponents where they are in the process of exchange. What's necessary is that the argument from evil be directed against well-founded notions of what God would be intellectually in comparison to us. Let's modify D to be appropriately weaker and more dialectical:

- (D') If S infers *n* (exactly no B are F) from *m* (no A are F), then S has justification for *n* only if S has reasons that would rebut well-founded challenges that S does not have reason to believe A is strictly representative of B.

Consider that the point of arguments from evil is to target *the theist's* conception of God, not the atheist's. The argument from evil is supposed to be an argument that the God of the believers doesn't exist. So if the theist has a notion of God that is itself well-founded (which we will assume here for Anselmian reasons) and which defeats the belief that the atheist's sample of God-justifying reasons is

---

<sup>3</sup> See Chris Tucker's discussion of representativeness in skeptical theological arguments for a similar analysis: "Why Skeptical Theism isn't Skeptical Enough," in *Skeptical Theism: New Essays*, eds. T. Dougherty and J. McBrayer (New York: Oxford University Press, 2014), 45-61.

representative, then the atheist's argument is not justification-affording for the relevant conclusion. That is, the atheological argument fails to be sufficiently *dialectical*, because the theist holds that God is considerably better off cognitively with regard to the relevant reasons up for consideration. This is exactly what Bergmann means to express with ST. The skeptical theist, then, defeats the inference by using D' *in conjunction with an appeal to the great cognitive gap between human minds and the divine mind*.

So D', with the dialecticality rider of the great cognitive gap, allows the skeptical theist to introduce their "well-founded challenge" of a God who is vastly cognitively superior to us: God's reasons, on the theist's conception, far exceed ours, and so *the inference* is defeated. More specifically, the induction that comprises *the inference* (which supports premise 1 of the evidential argument from evil) is undercut by an appeal to the gulf between the capacities and contents of God's mind and those of human minds. Notice that this gulf makes it so that there is little hope for justifying strict representativeness. If we grant that D' in conjunction with the appeal to the great cognitive gap defeats *the inference* that would be because *the inference* is one instance of a broader kind of theologically-inductive inference—one that is also defeated by D' in conjunction with the appeal to the great cognitive gap—which is that humans have an adequately representative sample of God's reasons for doing anything. Why would God, say, make our bodies so smelly or our elbows so ridiculous looking? Surely anyone who tries to answer that question, even with a plausible reason accessible to our minds, makes an error of presumption similar to that of *the inference*. We may have a reason available, but we do not have access to all of God's reasons, and so we have no reason to believe that our available reasons (if we have any) are strictly representative of God's reasons.

So the lesson of D', as we see it, is that it defeats *the inference* only because the dialectical requirement embedded in D' allows the skeptical theist to appeal to a substantive conception of the divine (and, thereby, to the great cognitive gap). *The inference*, then, is just one instance of a broader human presumption when reasoning about God's reasons and plan. Let us call the broader, more general category into which *the inference* fits a *theological induction*. The *negative* version of the theological induction takes the basic form:

*Since we humans cannot discern a justifying reason for God to do X or allow X, there are no such reasons.*

For the skeptical theist, the presumptiveness behind making such an induction is that our access to the reasons must be strictly representative, and we have no reason to suppose they are. Correspondingly, prohibitions on the thought

that we have strictly representative samples of God's reasons extends to the positive case of attributing our reasons to God as justifying for Him, too. Call this the *positive* version of the theological induction, and it takes the following form:

*Since humans can discern a justifying reason for God to do X or allow X, God's reason is that reason.*

If we think it possible for God to have a broader set of reasons than we have access to, perhaps even extending to reasons we cannot fathom, then both the positive and negative inductions will be unfounded. *The inference* behind the evidential problem of evil is simply a special case of (negative) theological induction, and under the skeptical theist's view, no theological induction (positive or negative) is justified or acceptable. In this way, the insight behind skeptical theism is the same as the insight behind the objection to petitionary prayer—we do not know better than or even as well as God as to what should or should not be the case.

### **3. Skeptical Creep: Undercutting Moral and Theological Knowledge**

A regular concern about skeptical theism is that it yields skeptical consequences wider than simply those on the question of whether we know the reasons why God would allow evils. Two domains of particular importance are regularly identified: moral and theological knowledge. In short, as the reasoning goes for the moral case, if God has inscrutable reason to allow what seem to us to be horrendous evils, then He may have reason to allow massive error about moral norms. The theological worry is that if God has good reason to allow toddlers to die in a rush of tsunami seawater, then he could very well have reason to permit priestly lies about the nature of salvation, the origin of evil, or His role in creation. The simple fact that we cannot think of reasons why He would do so is not reason to say that He does not have such reasons. And, in fact, us thinking of reasons for His veracity are themselves also undercut by the problem of theological induction, too. Again, the fact that we can think of reasons for God to do something does not mean that those are God's reasons or that God does not have access to defeating reasons for them. And that fact that we can think of reasons for God to do something does not mean that He has a reason to do that. Again, that is the lesson of both positive and negative theological inductions. Rational support for revealed and natural theological traditions, when put under rational scrutiny of this kind, evaporates.

The creep problem begins with the observation that skeptical theism provides defeaters for an important range of our moral knowledge as it relates to God's decisions. Once this range is defeated, the defeating conditions migrate to

other considerations beyond only God's decisions. Take any two cases of mundane moral evil, perhaps consistent child abuse that results in death. One is in the past, the other is currently transpiring. Nothing, to our knowledge, distinguishes the two, and we know for sure, assuming theism is true, the former must be justified for God to have allowed it. But what about the latter, the one happening now? Assuming time isn't a morally relevant feature, the latter, too, is justified. Or at least, we have no reason to hold it isn't. If this is the case, our ordinary moral judgment is not a reliable source of what is and what is not justified.<sup>4</sup> Skeptical theism, then, yields moral skepticism.

Skeptical theism provides dialectical defeaters for certain inferences from what we take to be the best of our (admittedly limited) knowledge. In the theological case, the inference is that we can think of no good (or undefeated) reason for God to deceive us (or allow us to be deceived) about his nature, so there is no reason.<sup>5</sup> In the moral case, the inference is that we can think of no good reason in the relevant cases of evil for God to allow evil, so we've inferred there is none. But, recall, the ST theses have run that the goods, evils, and relevance relations between them that we know provide us no justification for thinking they

---

<sup>4</sup> This argument parallels Almeida and Oppy's dilemma for the skeptical theist, since in the everyday cases of judging whether to interfere, we either *should* trust our judgment of what should be done all things considered (and so our knowledge should be representative) or if it is not representative, we *should not trust* our judgment. See Michael Almeida and Graham Oppy, "Skeptical Theism and Evidential Arguments from Evil," *Australasian Journal of Philosophy* 81 (2003): 506. The former option is not skeptical, and the latter is plenty skeptical, but morally objectionable in a way that the skeptical theist should find worth rejecting. Others who have run versions of the moral skepticism argument are William Hasker, in *Providence, Evil, and the Openness of God* (New York: Routledge, 2004), Jeffrey Jordan, in "Does Skeptical Theism lead to Moral Skepticism?," *Philosophy and Phenomenological Research* 72 (2006): 403-17, Stephen Maitzen, in "Skeptical Theism and God's Commands," *Sophia* 46 (2007): 237-43, and Aikin and Ribeiro, in "Skeptical Theism, Divine Commands, and Moral Skepticism," *International Journal for the Study of Skepticism* 3 (2013): 77-96.

<sup>5</sup> For versions of the theological skepticism argument, see Wes Morriston's "Skeptical Demonism," in *Skeptical Theism: New Essays*, eds. Dougherty and McBrayer, 221-234; Erik Wielenberg's "Divine Deception," in *Skeptical Theism: New Essays*, eds. Dougherty and McBrayer, 236-248; and John Park, "The Moral Epistemological Argument for Atheism," *European Journal for Philosophy of Religion* 7 (2015): 121-142. Further, Gale saw very early on in these discussions that "defensive skepticism" in theodicy destroys all the objects of faith and love in unclarity. See Richard Gale, "Some Difficulties in Theistic Arguments from Evil," in *The Evidential Problem of Evil*, ed. Daniel Howard-Snyder (Indianapolis: Indiana University Press, 2016), 206-218. Ireneusz Ziemiński argues that the consequences are ultimately blasphemous for theists: "The Problem of God's Existence: In Defence of Scepticism," *European Journal for Philosophy of Religion* 7 (2015): 143-163.

are strictly representative of all of them. In turn, *the same defeating reason posed for evidential atheists can be posed for theologians and moralists*. Simply, they all commit their own versions of the fallacious theological induction. And so the skepticism in skeptical theism creeps beyond its domain into theology and moral judgment.

#### 4. From Skeptical Theism to Global Skepticism

The skeptical theist's basic strategy of applying D' to yield defeat has been this: frame premise 1 of the argument from evil (viz., "There are instances of *prima facie* gratuitous evil that God could have prevented without losing some greater good or failing to prevent some greater evil") as being based on a purportedly representative sample of supporting reasons, which we have labelled *the inference* (viz., *Since we humans cannot discern a justifying reason for God to allow evils, there is none*). Now, according to the atheologian, the reasons surveyed in the sample provide appropriate justification for the claim in premise 1. But, according to the skeptical theist, what's required for appropriate justification, given the dialectical context, is that the atheologian must have reason to hold that—or at least have rebutting reasons against well-founded challenges to the claim that—the reasons available in the sample are appropriately *representative*. And, for the skeptical theist, these sampled reasons must be *strictly* representative: the atheological claims to *discern* zero reasons for God to have allowed evils, but premise 1 expresses the idea that *there are no reasons* for God to have allowed evils. So, as we put it earlier, *the inference* requires the percentages of zero percent to be identical between the sample and the target classes. But the skeptical theist then appeals to the great cognitive gap between human minds and the divine mind. Might not God have reasons we have no access to? Consequently, the requirement is that the atheologian must have some reason, from his or her limited evidence, to think that the sample evidence is strictly representative of the *total evidence*. J.L. Schellenberg has identified the inclination to make this demand as *total evidence skepticism*:

[T]otal evidence skepticism is the claim that, for any proposition expressing a belief . . . of ours, we have reason to be in doubt, or skeptical, about whether the total evidence supports that proposition.<sup>6</sup>

So, were our available evidence to support premise 1, for all we know, the total evidence (which God has) may not.

---

<sup>6</sup> "Skeptical Theism and Skeptical Atheism," in *Skeptical Theism: New Essays*, eds. Dougherty and McBrayer, 199.

Notice that Schellenberg's notion of total evidence skepticism tracks the skeptical theist's appeal to *the reasons God would have* quite exactly: skeptical theists hold that we cannot know what reasons God (an *omniscient* being) might have for permitting evils. In other words, we cannot know whether the reasons *we* have relating to the possible permission of evils are a strictly representative set of the reasons an *omniscient* being would have: namely, all-the-reasons-there-are, i.e., the "total evidence" regarding permission of evils. So, the skeptical theist's strategy is to use D' in conjunction with an appeal to the great cognitive gap to challenge *the inference*, thereby undercutting the atheologian's justification for premise 1 of the argument from evil. As we noted above, the skeptical theist holds that D', in conjunction with an appeal to the great cognitive gap, defeats *any* theological induction regarding God's reasons, positive or negative.

But D', in conjunction with an appeal to the great cognitive gap, provides a path to global skeptical creep. First, consider that if all of our induction-based beliefs had to pass the total evidence requirement in order to be justified, then arguably very few of those beliefs would pass and, hence, very few of our ordinary induction-based beliefs would be justified. For how could we establish that the evidence we do possess for any such belief is *strictly representative* of the total evidence? If the skeptical theist's appeal to the great cognitive gap is indeed a "well-founded challenge" (as required by D'), it would seem to defeat all beliefs that *derive from* or *rely upon* any inductive reasoning, not just theological inductions. This class of beliefs seems potentially very large.

Of course, D' only requires one to rebut *well-founded challenges* (concerning whether the evidence one has is strictly representative of the total evidence) to the induction. The trouble for skeptical theists is that most of the founding analogies for the godhead are those that do not guarantee that we will always have epistemically adequate access to the total evidence of *any* relevant domain of inquiry, whether induction-based or not. This is an important point, because it puts theological induction at the core of all of our foundational and inferential knowledge, and so, makes the fallaciousness of the induction a defeating condition. Consider that God is regularly analogized to a parent, and it is a standard practice for parents to shield children from many, many hard and uncomfortable truths. And so, children will have skewed samples of what the world is like, precisely because their parents have manipulated their evidence for the sake of *not* being representative of the total evidence. Or consider another analogy, that God is like a ruler or king. Again, it is a standard truth of rulers and kings that they manage their image in ways that project them in their best lights, that they keep many background issues out of the public eye, and that there are

matters that are managed so that the populace is happily ignorant of them. For sure, these manipulations are beneficent by hypothesis, but they are manipulations nevertheless. A final analogy would be that with an artificer or creator. Many products of skilled craftspeople are deceptively simple—they are designed to interface with us in ways that make them seemingly easy to understand, but in fact they are considerably more complex than our simple exchanges reveal them to be. And so, the world, our own minds, and the perceptual relations between our minds and that world, and the *a priori* justification supported by what we take as our understanding, are products of God that are designed to appear simple to us, but could in fact be cleverly crafted illusions that cover over massive complexity. In fact, they can even be complete misrepresentations of what's actually the case, as one might think that the 'close door' button on an elevator actually makes the door close faster instead of merely seeming to. We, on this well-founded analogy between God as expert craftspeople, understand very little. Our point here, again, is to show that, given the well-founded analogies between God and parents, monarchs, and craftspeople, D'-based challenges to *any* of the beliefs of the skeptical theist—whether induction-based or not—appear to be dialectically well-founded. And so, given their inability to rebut those challenges, skeptical consequences follow, but this time, they appear global.

## 5. Trying to Dam the Creep

Presumably, if skeptical theism generates global skepticism with respect to *all beliefs*, then skeptical theism fails to play any useful dialectical role for the theist. Showing that we are not justified in accepting premise 1 of the argument from evil because we are not justified in accepting *any claim whatsoever* presumably counts as a disastrous dialectical backfire for the skeptical theist.

To avoid this result, the skeptical theist might seek to limit the application of D' to certain cases. For example, if D' is the correct epistemic principle for evaluating *the inference* and *only the inference*, then worries about global skepticism evaporate. But this is not a very promising line of response. Consider what motivated D' in the first place. The thought was that, considering the limitations in our evidence, God might have reasons we don't, or even can't, know or understand: there is a great cognitive gap between God's mind and ours. So drawing conclusions about all of the reasons from our limited sampling of reasons is presumptuous and unjustified (so says the skeptical theist). But, as we pointed out in section 4, this reasoning need not be inherently theistic. The reasons God would have are, given His omniscience, simply all the reasons there are. So to compare our limited evidence to God's evidence (as skeptical theism invites us to



do in the case of undercutting the atheological argument) is exactly the same as comparing our limited evidence to the total evidence. This means that the skeptical theist's motivation for D' can be translated into theistically-neutral language very simply: considering the limitations in our evidence, the total evidence might contain reasons we don't, or even can't, know or understand, and this reflection is an undercutting defeater for any belief, whether induction-based or not. Now, the skeptical theist's own motivation for enforcing this total evidence requirement is either persuasive or not. If it's *not* persuasive, then the skeptical theist's appeals, via D', to the great cognitive gap are not adequately motivated and can be dismissed. If, on the other hand, this motivation *is* persuasive, it leads to a global skepticism for the skeptical theist.

Suppose that in response the skeptical theist says, "Yes, considering the limitations in our evidence, the total evidence might contain reasons we don't, or even can't, know or understand. So drawing conclusions about all the reasons from our limited sampling of reasons does not give us any guarantee that our beliefs or conclusions will be correct. Still, since we can do no better when deciding what to believe, we must make do and accept such *prima facie* justifications for our beliefs." This may or may not be the right response to make to total evidence skepticism. But even if it is, it won't help the skeptical theist, since this type of response would leave premise 1 as *prima facie* justified, which is the most the atheologian ever claimed for it. So to avoid leaving premise 1 unscathed, the skeptical theist would need some respectable ground for treating premise 1 *differently* from other kinds of claims. And, as we've shown above, this doesn't seem plausible. The skeptical theist's motivation for embracing D' came from making humbling comparisons between our reasons and the reasons God might have, but we have shown that this point can be detheologized and translated into the total evidence requirement. Further, it seems that well-founded notions of God's nature are perfectly amenable to extending D' well beyond the moral reasoning in theodicy cases. Thus, attempts to dam the creep fail.

As a final strategy of creep-resistance, a skeptical theist might seek to *differentiate* the beliefs they wish to maintain (distinguished from the atheologian's premise-1 claim) by advancing a common-sensist view regarding a broad class of beliefs. This is, in fact, how Michael Bergmann replies to the Schellenbergian skeptical argument. Bergmann argues that, even with the total evidence requirement, many beliefs remain immune to skeptical jeopardy: "It's true that I don't have reflective access to the total evidence bearing on whether I exist or on whether I have hands or on whether I had orange juice for breakfast

today or whether  $2+2=4$  or whether I'm in extreme pain,"<sup>7</sup> Bergmann admits. However, according to Bergmann, "in each of these cases I have knowledge or reasonable belief from which I can infer certain facts about the total evidence bearing on these propositions. For example, I can reasonably believe the total evidence supports the claim that  $2+2=4$ . I reasonably believe this even though I don't have reflective access to the total evidence bearing on that claim."<sup>8</sup> The success of this approach as an anti-skeptical strategy depends on what Bergmann calls the "epistemic force" of the claim in question (e.g.,  $2+2=4$  or that one had orange juice for breakfast). Bergman holds that from that epistemic force of the claim, one is able to make inferences about the status of the total evidence:

The point is just that from the reasonable belief that  $p$ , one can infer that the total evidence does not include a successful proof that  $p$  is false (since if  $p$  is true, the total evidence supports  $p$ , in which case it does not include a *successful* proof that  $p$  is false).<sup>9</sup>

So, on Bergmann's view, the requirement of total evidence does not provide a successful undercutting defeater for the kinds of beliefs targeted by a global skeptical creep, because those targeted beliefs enjoy sufficient, intuitively-available "epistemic force" to repel any such skeptical assault. As Bergmann sees it, the defeat a requirement like  $D'$  has for *the inference* is that  $D'$  is a requirement for *inductions*, but the epistemic force of the cases Bergmann has in mind are not instances of induction, but rather cases of non-inferential justification or intuition. Yet, as we've already argued,  $D'$  *does no work at all for the skeptical theist without the appeal to the great cognitive gap*. And it is that appeal to the great cognitive gap that is the bull in the china shop for the skeptical theist. As we put it before, once one accepts the existence of the great cognitive gap, one no longer has any guarantee that one will always have epistemically adequate access to the total evidence of *any* relevant domain of inquiry, whether induction-based or not. And we have well-founded theological reasons (from the parent, monarch, and craftsperson analogies earlier) to hold that there are defeaters for a wide range of these non-inferentially justified beliefs. To hold that the reasons we have implies that there are no reasons that run counter seems as manifestly impertinent as *the inference*.

---

<sup>7</sup> Michael Bergmann, "Theism and Total Evidence Skepticism" in *Skeptical Theism: New Essays*, eds. Dougherty and McBrayer, 209-220.

<sup>8</sup> Bergmann, "Theism and Total Evidence Skepticism," 215.

<sup>9</sup> Bergmann, "Theism and Total Evidence Skepticism," 217.

By our lights, Bergmann's common-sensist line seems out of step with precisely what is *skeptical* about skeptical theism. The kind of *epistemic humility* which seems to drive skeptical theism in its retort to atheological presumption and hubris does not seem to fit well with Bergmann's casual confidence in the "epistemic force" of his beliefs.<sup>10</sup> In other words, Bergmann's epistemic claims appear bold given the scope of challenges consistent with skeptical theism's appeal to the great cognitive gap. Again, recalling our analogies from the previous section, if God is like a parent or a monarch or an artificer, then there may be *many* things we think are simple, things which we will think we have no problem understanding, but which are, in fact, complex and significantly different from what we believe them to be, indeed perhaps even such as to be beyond our understanding. Appearances may be managed, evidence curated, functions engineered.<sup>11</sup> For the sake of argument, we can even concede that any ignorance or false beliefs humans are subjected to could all be for the good, but that point does not undercut the skeptical worry that *a beneficent god might allow such ignorance or false beliefs as products of intuition or common sense*. Thus, those simple Moorean cases Bergmann reviews, by our lights, are all in the same boat as those *prima facie* justified commitments driving the atheological argument from evil. Let us grant that they have initial epistemic plausibility, but in light of the well-founded commitments to what God's role would be, were He to exist, those beliefs are not *ultima facie* justified for the skeptical theist, because they do not, given the cognitive gulf between us and God, provide skeptical theists with any grounds for supposing they enjoy epistemically adequate access in the relevant domains.

Notice, further, that it seems open to the atheologian to take Bergmann's line of argument as a cue and apply it to the premises for the evidential argument from evil. One might say, e.g., that there is *significant epistemic force* for the thought that *there's no excuse for allowing some particular evils*, or that *some evils*

---

<sup>10</sup> See, for example, Todd Long's case for "an epistemic position of humility before God" in "Minimal Skeptical Theism," in *Skeptical Theism; New Essays*, eds. Dougherty and McBrayer, 71.

<sup>11</sup> We also hasten to add that there is a good deal of literature on whether the gods lie to and deceive humans full-stop. It seems that there is Biblical reason to think so, as it seems that God intentionally sends delusions (2 Thessalonians 2:11); and God sends prophets that He has deceived (Ezekiel 14:9). Further, it seems that gods, qua gods, are perfectly capable of and willing to deceive humans. Homer's gods, the Norse gods, and so on, in fact, provide unique reasons for skepticism in light of their inclinations and abilities. See Michael Forster's account of the Homeric reasons for skepticism in "Homeric Contributions to Skepticism," in *Skepticism: Historical and Contemporary Inquiries*, eds. G. Anthony Bruno and A. C. Rutherford (New York: Routledge, 2018), 7-23.

*are clearly gratuitous.* The problem of evil literature is replete with stories that seem to us to fit the bill, possessing the same kind of initial epistemic plausibility as Bergmann's cases. So what is to prevent the atheologian from then running the Bergmann-style argument that, since there's reason to hold premise 1 is true, we can legitimately infer that there's reason to hold that there are no defeaters in the total evidence? Surely it is reasonably intuitive to say that some things that have happened are so bad, *there's no excuse* for allowing them, and that thesis is true not as a matter of induction, but as a matter of assessing the kind of bad that has transpired. That is, there's a difference between saying that there is no reason that could justify some evil because one has surveyed a set of reasons and they fail and saying there is no excuse for some evil because the evil is so intuitively egregious—to try to justify it would fail to honor the wrong done. That's the epistemic force of the atheologian's view that there aren't God-justifying reasons for those evils. Of course, we think the skeptical theist will respond that the atheologian's Bergmann-style epistemic force argument fails because of the well-founded notion of what God is supposed to be, viz. a being so inconceivably cognitively superior to us that we are not justified in relying on what seems initially epistemically plausible to *us* as a guide to what's ultimately true. But, again, given that same well-founded commitment and the resulting position of epistemic humility, we have argued that the cases Bergmann highlights are subject to the same response. All of the instances require a background of theological induction, which *ex hypothesi*, is unfounded. Creep ensues.

## 6. Conclusion

The epistemic principle to which skeptical theists implicitly appeal, when deployed in conjunction with their appeal to the great cognitive gap between humans and God, proves to be problematically demanding and thereby generates global skepticism. We think that skeptical theists will likely find the broader skeptical consequences of their view unpalatable. For their part, they would surely wish to keep a good deal of their theological and moral views in place, and they most certainly would blanch at global epistemic collapse. As such, the creep problem for skeptical theism is a form of 'proves too much' objection to a dialectical opponent. Of course, such arguments depend on our interlocutors actually holding that the broadening skepticism *is* too much. But if our arguments convert the skeptical theist into a broader kind of skeptic, we (who are both sympathetic with the skeptics) might say this is a fortuitous conclusion.

# THE KNOBE EFFECT WITH PROBABLE OUTCOMES AND AVAILABILITY HEURISTIC TRIGGERS

Tommaso OSTILLIO, Michal BUKAT

**ABSTRACT:** This paper contributes to the existing philosophical literature on the Knobe Effect (KE) in two main ways: first, this paper disconfirms the KE by showing that the latter does not hold in contexts with probable outcomes; second, this paper shows that KE is strongly sensitive to the availability heuristic bias. In particular, this paper presents two main findings from three empirical tests carried out between 2016 and 2018: the first finding concerns the fact that if the issuer of a decision with consequences on third parties is unlikely to be perceived as unfriendly, then KE is reduced or absent; the second finding regards instead the fact that if an action has two possible outcomes (one likely to obtain with strong intensity and one likely to obtain with less intensity), then KE does not obtain for decisions whose side-effects have limited consequences on third parties.

**KEYWORDS:** experimental philosophy, Knobe effect, cognitive bias, negative externality

The concept of intentionality has played and keeps playing a dominant role in contemporary epistemology, in contemporary philosophy of mind, in contemporary philosophy of action and in contemporary meta-ethics. This is because philosophers have struggled and still struggle with finding a definition of intentionality, which leads to long-term agreement among different schools of thought.

Historically speaking, the contemporary philosophical literature on intentionality has taken two main opposite directions: on the one hand, some philosophers find an association between intentionality and the reasons to act in a particular way;<sup>1</sup> on the other hand, some philosophers find instead an association between intentionality and the aboutness (i.e. the content) of mental states.<sup>2</sup>

---

<sup>1</sup> See Gertrud Elizabeth Margaret Anscombe, *Intention* (Cambridge: Harvard University Press, 1957); Donald Davidson, "Actions, Reasons, Causes," *The Journal of Philosophy* LX, 23 (1963): 685-700.

<sup>2</sup> See Daniel Clement Dennett, "Intentional Systems," *The Journal of Philosophy* 68, 4 (1981): 87-106; John Rogers Searle, *Intentionality: An Essay in the Philosophy of Mind* (Cambridge: Cambridge University Press, 1983), 1-36.

Besides, although the literature is exceptionally vast on both sides, no perfect argument to defend a particular definition of intentionality has been found on neither side.

At the same time, philosophers' overall troubles in defining intentionality have grown bigger since the so-called experimental philosophers have shown that there exists a discrepancy between the way philosophers understand intentionality and the way folks attribute intentionality to agents.

In this respect, Malle and Knobe investigate how folks attribute intentionality to agents empirically and find that, while philosophers usually relate intentionality to purpose or mental content, folks relate intentionality to possessing the right set of skills to carry out a given course of action.<sup>3</sup> That is, according to the folks surveyed by Malle and Knobe, an action is intentional if and only if an agent is able to carry out the course of action he or she intends to carry out.<sup>4</sup>

In the light of the findings of Malle and Knobe,<sup>5</sup> Knobe carries out another survey, which relates intentionality (understood as possessing the right skills to carry out the intended course of action) to the externality of actions.<sup>6</sup> In particular, Knobe constructs two vignettes where a fictitious character, Jake, is in need for money and gains the amount of money he needs either by participating in a rifle contest or by killing his old rich aunt.<sup>7</sup> Moreover, Knobe divides each vignette case in two sub-vignettes where two assumptions are dominant: either Jake is a skilled shooter or Jake is not a skilled shooter.<sup>8</sup>

In the first vignette, Jake participates in a rifle context where he is to shoot a bull in its eye from a big distance. If Jake succeeds at shooting the bull in its eye, he gets the money, whereas, if he does not, he gets no money. Yet Jake accomplishes his goal in both sub-vignette-cases regardless of whether he is a skilled shooter or not. QED, Knobe finds that when 37 random subjects are asked whether Jake acted intentionally or not, their general answer is that he acted intentionally in the first sub-vignette-case, but he did not do so in the second sub-vignette-case.<sup>9</sup> That is,

---

<sup>3</sup> Bertram F. Malle, and Joshua Knobe, "The folk concept of intentionality," *Journal of Experimental Social Psychology* 33 (1997): 101-121.

<sup>4</sup> Malle and Knobe, "The folk concept of intentionality."

<sup>5</sup> Malle and Knobe, "The folk concept of intentionality."

<sup>6</sup> Joshua Knobe, "Intentional Action in Folk Psychology: An Experimental Investigation," *Philosophical Psychology* 16, 2 (2003): 309-324

<sup>7</sup> Knobe, "Intentional Action in Folk Psychology."

<sup>8</sup> Knobe, "Intentional Action in Folk Psychology."

<sup>9</sup> Knobe, "Intentional Action in Folk Psychology."

Jake's accomplishment is intentional as far as he possesses the right set of skills to shoot the bull in its eyes from a great distance.

By contrast, in the second vignette, Jake gets the amount of money he needs if and only if he kills his old rich aunt, while she is at home, by shooting her through the window of the house in front of hers. As in the first vignette, Jake successfully accomplishes his goal in both sub-vignettes. Yet, when 37 random subjects are asked whether Jake acted intentionally or not, their general answer is that he acted intentionally regardless of whether Jake is a skilled shooter or not.

Thus, Knobe concludes that while it holds true that folks overall relate intentionality to the ability to accomplish a given intended goal, the gathered data show also the attribution of intentionality to agents is dependent on the externality of a given action. For folks consider Jake's murder of his old aunt as intentional in both sub-vignettes.<sup>10</sup>

On this basis, Knobe constructs two more vignettes, which put a stronger emphasis on the side-effects of an action. More specifically, the two vignettes recount the story of a firm's VP who wants to implement a business project aimed at increasing his firm's profits: in the first case, the business project is implemented successfully with a positive externality (i.e. its implementation helps the environment); in the second case, the side-effect of a success implementation is a negative externality (i.e. its implementation harms the environment).<sup>11</sup> QED, Knobe finds that when 78 random subjects are asked whether the VP caused both side-effects intentionally or not, their dominant answer is that he did so in the second case, but he did not do so in the first case.<sup>12</sup>

In the philosophical literature, the effect observed by Knobe<sup>13</sup> is usually referred to as the Knobe effect (i.e. folks' tendency to consider an action intentional if and only if it has negative side-effects) and, since the findings of Knobe<sup>14</sup> have been published, the Knobe effect (KE) has been the object of important debates in philosophy and in the social sciences. In fact, the findings of Knobe<sup>15</sup> have also gained a special place in the research programs of some researchers in business and economics because KE might explain how people perceive specific business or policy decisions (yet with some limitations).

---

<sup>10</sup> Knobe, "Intentional Action in Folk Psychology."

<sup>11</sup> Knobe, "Intentional Action and Side Effects in Ordinary Language," *Analysis* 63, 3 (2003): 190-94.

<sup>12</sup> Knobe, "Intentional Action and Side Effects."

<sup>13</sup> Knobe, "Intentional Action and Side Effects."

<sup>14</sup> Knobe, "Intentional Action and Side Effects."

<sup>15</sup> Knobe, "Intentional Action and Side Effects."

In this regard, Feltz *et al.* implement an experimental setting where a random sample of subjects undergoes a two-stage treatment: in the first stage, the surveyed subjects are asked to take actions with side-effects and then evaluate how intentional their actions are on a 5-points Likert scale; in the second stage, the surveyed subjects are asked to evaluate the intentionality of some actions carried out in some vignette case, which depict the events of the first stage, on a 5-points Likert scale.<sup>16</sup> Interestingly, Feltz *et al.* find that the surveyed subjects judge their actions in the first experimental stage as being less intentional than the actions depicted in the vignette cases of the second experimental stage.<sup>17</sup> That is, Feltz *et al.* find that a change from a first-person to a third-person perspective might affect how intentionality is evaluated and attributed to agents.<sup>18</sup>

On the other hand, Utikal and Fischbacher<sup>19</sup> object that the vignette cases of Knobe<sup>20</sup> do not properly consider the economic gains of the firm harming/helping the environment. Accordingly, Utikal and Fischbacher<sup>21</sup> translate the vignettes of Knobe<sup>22</sup> into a market-like setting with three scenarios where three players play respectively the role of the firm's VP (player 1), the role of the environment (player 2) and the role of an external judge (player 3) who can punish or reward player 1 depending on the outcomes of player 1's decisions. The experimental setting designed by Utikal and Fischbacher<sup>23</sup> is divided into two stages. The first stage X represents the default economic status of all the players and is divided in three sub-stage in the following way: in the first sub-scenario, a strong active player 1 affects a weak passive player 2; whereas, in the second sub-scenario, a weak (player 1 affects a strong passive player 2; and, in the sub-third scenario, a weak active player 1 affects a weak passive player 2. The second stage Y represents the final economic status Y of player 1 and player 2 after player 2 opted for one of the three following options: a bad outcome (harm); a good outcome (help); and a neutral outcome. Figure 1 (below) shows that, in each sub-scenario, the outcomes of player 1's decisions lead to different endowment reallocation. Eventually, after having observed what outcome obtains, player 3 can either reward player 1 (i.e.

---

<sup>16</sup> Adam Feltz, Maegan Harris, and Ashley Perez, "Perspective in intentional action attribution," *Philosophical Psychology* 25, 5 (2012): 673-687.

<sup>17</sup> Feltz *et al.*, "Perspective in intentional action attribution."

<sup>18</sup> Feltz *et al.*, "Perspective in intentional action attribution."

<sup>19</sup> Verena Utikal and Urs Fischbacher, "Attribution of externalities: an economic approach to the Knobe effect," *Economics and Philosophy* 30, 2 (2014): 215-240.

<sup>20</sup> Knobe, "Intentional Action and Side Effects."

<sup>21</sup> Utikal and Fischbacher, "Attribution of externalities."

<sup>22</sup> Knobe, "Intentional Action and Side Effects."

<sup>23</sup> Utikal and Fischbacher, "Attribution of externalities."



The Knobe Effect with Probable Outcomes and Availability Heuristic Triggers

player 3 can subtract points from player 2 and reallocate them to player 1) or punish player 1 (i.e. player 3 can subtract points from player 1 and reallocate them to player 2). The latter option for player 3 represents the activation of KE.

Setting		Default X		Endowment after change Y	
		Player 1	Player 2	Player 1	Player 2
<b>StrongActiveSmallHelp</b>	Harm	50	50	60	30
	Help	50	20	60	30
	No side effect	-	-	60	30
<b>WeakActiveSmallHelp</b>	Harm	20	80	30	60
	Help	20	50	30	60
	No side effect	-	-	30	60
<b>WeakActiveBigHelp</b>	Harm	20	80	30	60
	Help	20	20	30	60
	No side effect	-	-	30	60

Figure 1 - Verena Utikal and Urs Fischbacher, "Attribution of externalities: an economic approach to the Knobe effect," 220.

In the light of the aforementioned premises, Utikal and Fischbacher<sup>24</sup> find that KE obtains only in the first scenario, while it reverses in the second and in the third scenario. That is, in the first scenario, player 1 is overall punished, whereas, in the second and in the third scenario, player 1 is overall rewarded by player 2 regardless of the option chosen by player 1. This is because, according to Utikal and Fischbacher,<sup>25</sup> Player 1 does not look unfriendly to Player 3 in the second and in the third scenario.

Most importantly, the findings of Utikal and Fischbacher<sup>26</sup> find some confirmation in an earlier study by Wible,<sup>27</sup> where 36 random subjects are asked to evaluate the following:

*The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will increase profits, and it will also help the environment.' The chairman of the board answered, 'Great! I care about*

<sup>24</sup> Utikal and Fischbacher, "Attribution of externalities."

<sup>25</sup> Utikal and Fischbacher, "Attribution of externalities."

<sup>26</sup> Utikal and Fischbacher, "Attribution of externalities."

<sup>27</sup> Andrew Wible, "Knobe, Side Effects, and the Morally Good Business," *Journal of Business Ethics* 85 (2009): 173–178.

*helping the environment. I am happy that we can help the environment. I am happy that we can help the environment and make a profit at the same time. Let's start the new program.' They started the new program. Sure enough, the environment was helped.*<sup>28</sup>

Wible finds that 55% of the surveyed subjects says that the chairman acted intentionally. In other words, the fact that the intentions of the chairman were good and clearly stated impacts how intentionality is evaluated and attributed to agents.

Thus, considering the findings of Wible<sup>29</sup> and Utikal and Fischbacher,<sup>30</sup> there is room to argue that the availability heuristic bias<sup>31</sup> might nudge the activation of the Knobe effect in case like those described by Knobe.<sup>32</sup> In fact, the vignettes of Knobe<sup>33</sup> force the surveyed subjects to attribute intentionality to agents under uncertainty in presence of restrained data, which nudge stereotype-based judgements about the wrongdoings of greedy businessmen.

Furthermore, another objection to Knobe<sup>34</sup> might be that his vignettes represent cases where the telos of the events is given and taken for granted. That is, the intended outcomes entailed by the decision of the firm's VP are granted to obtain. Yet, when business projects are implemented, this is seldom the case because the unaccounted side-effects of a business decision might be more than executives can forecast alone.

Accordingly, in order not to fall into too speculative forms of argumentation about the vignette cases of Knobe,<sup>35</sup> this paper tests empirically whether the Knobe Effect is immune to the effects of the availability heuristic bias and whether the Knobe Effect obtains once the forecasted side-effects of an action are only probable. The next section presents the results of three survey-based experiments, which were carried out by the authors of this paper between 2016 and 2018.

## Experiment 1

The first experiment took place in December 2016 within a different research project and involved two runs of testing: in the first run (Group 1), 40 master

---

<sup>28</sup> Wible, "Knobe, Side Effects, and the Morally Good Business," 174.

<sup>29</sup> Wible, "Knobe, Side Effects, and the Morally Good Business."

<sup>30</sup> Utikal and Fischbacher, "Attribution of externalities."

<sup>31</sup> See Amos Tversky and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science, New Series* 185, 4157 (Sep. 27, 1974): 1124-1131.

<sup>32</sup> Knobe, "Intentional Action and Side Effects."

<sup>33</sup> Knobe, "Intentional Action and Side Effects."

<sup>34</sup> Knobe, "Intentional Action and Side Effects."

<sup>35</sup> Knobe, "Intentional Action and Side Effects."

The Knobe Effect with Probable Outcomes and Availability Heuristic Triggers

students of Finance at Kozminski University were asked to express their judgement on the vignette presented in Task 1 offline; in the second run (Group 2), 50 random individuals recruited on Amazon Mechanical Turk were asked to express their judgement on the vignette presented in Task 1 online.

The overall goal of the experiment was to test whether the surveyed subjects overall attribute intentionality to an action whose side-effects are only probable. More specifically, following the vignettes of Knobe,<sup>36</sup> we constructed a vignette where the outcomes of a business decision are double. That is, the latter decision leads to a bigger forecasted outcome that is likely to obtain with stronger intensity and a smaller forecasted outcome that is likely to obtain with less intensity.

On this basis, as shown below, Task 1 focused only on finding out whether KE activates only in the context of the smaller forecasted outcome that is likely to obtain with less intensity:

**Task 1:** Assume that a hedge fund decides to finance a research project for the development of a new pain killer with €200M. Assume also that the project is carried out using dogs as test animals and that the dogs might either survive or die with some probability after the experiments is performed by researchers. In any case, the development of the pain killer generates returns that amount to 30% of the hedge fund's initial investment. You're asked to evaluate the following.

CASE 1: The experiment is carried out successfully, the project generates returns that amount to 30% of the hedge fund's initial investment and the dogs used as test animals survive with probability with probability 0.75, i.e. few dogs die because of the side-effects of the experiment. Did the hedge fund cause the death of few of the dogs intentionally? Mark the option you choose.

A) YES;

B) NO.

CASE 2: The experiment is carried out successfully, the project generates returns that amount to 30% of the hedge fund's initial investment and the dogs used as test animals die with probability with probability 0.75 because of the side-effects of the experiment, i.e. few dogs survive. Did the hedge fund cause the survival of few of the dogs intentionally? Mark the option you choose.

A) YES;

B) NO.

---

<sup>36</sup> Knobe, "Intentional Action and Side Effects."

RESULTS - CASE 1	Group 1 (N=40)	Group 2 (N=50)
YES	37.5%	40%
NO	62.5%	60%
<i>Significance</i>	$\chi^2 = 2.5 (1) p = 0.114$	$\chi^2 = 2 (1) p = 0.157$
RESULTS - CASE 2	Group 1 (N=40)	Group 2 (N=50)
YES	32.5%	20%
NO	67.5%	80%
<i>Significance</i>	$\chi^2 = 4.9 (1) p = 0.027$	$\chi^2 = 9.68 (1) p = 0.002$
RESULTS - COMBINED	CASE 1 (N=90)	CASE 2 (N=90)
YES	39%	30%
NO	61%	70%
<i>Significance</i>	$\chi^2 = 4.44 (1) p = 0.035$	$\chi^2 = 14.4 (1) p = 0.000$

Table 1 - Experiment 1: results

The results in Table 1 show that both Group 1 and 2 overall do not attribute intentionality to the hedge fund in CASE 1 and CASE 2. Yet the span between YES and NO is statistically significant only in CASE 2 for both Group 1 and 2. Hence, KE is not nullified.

KE is instead nullified when the results are combined. Therefore, there is room to argue that if a decision leads to a forecasted side-effect that is likely to obtain with less intensity, then there might be no attribution of intentionality on the issuer of that decision.

## Experiment 2

After having presented the results of *Experiment 1* at some conferences and workshops, we received two main objections concerning our vignettes: first, the vignettes should have accounted also for the reverse case, i.e. for the case where the bigger side-effect obtains; second, the content of the vignettes is expressed in a very neutral language and nudges a biased evaluation under uncertainty. Both objections are addressed both by *Experiment 2* and *Experiment 3*.

More specifically, as shown below in Task 2.1, Task 2.2, Task 2.3, Task 2.4, *Experiment 2* provides a more explicit version of Task 1 including both the case where the big side-effect obtains and the case where the small side-effect obtains. Task 2.1, 2.2, 2.3, 2.4 are tested against the intuitions of 102 individuals randomly selected on Amazon Mechanical Turk.

Task 2.1: A hedge fund decides to finance a research project for the development of a new painkiller with \$500M. The researchers involved in the project use dogs and cats as test animals. In short, the researchers test the effectiveness of the

## The Knobe Effect with Probable Outcomes and Availability Heuristic Triggers

painkiller by causing some big harm to dogs and cats. Depending on how big a pain the researchers will inflict to dogs and cats, the test animals can either survive or die with some probability. Either ways, the hedge fund will turn a profit that amounts to 60% of the initial investment.

The experiment is carried out successfully. The hedge fund earns a profit of 60% on top of the initially invested capital. Yet dogs and cats survive with probability 0.75, i.e. few of them die and most of them survive. Did the hedge fund cause the survival of most of the test animals intentionally?

YES; NO.

Task 2.2: A hedge fund decides to finance a research project for the development of a new painkiller with \$500M. The researchers involved in the project use dogs and cats as test animals. In short, the researchers test the effectiveness of the painkiller by causing some big harm to dogs and cats. Depending on how big a pain the researchers will inflict to dogs and cats, the test animals can either survive or die with some probability. Either ways, the hedge fund will turn a profit that amounts to 60% of the initial investment.

The experiment is carried out successfully. The hedge fund earns a profit of 60% on top of the initially invested capital. Yet dogs and cats survive with probability 0.75, i.e. few of them die and most of them survive. Did the hedge fund cause the death of few of the test animals intentionally?

YES; NO

Task 2.3: A hedge fund decides to finance a research project for the development of a new painkiller with \$500M. The researchers involved in the project use dogs and cats as test animals. In short, the researchers test the effectiveness of the painkiller by causing some big harm to dogs and cats. Depending on how big a pain the researchers will inflict to dogs and cats, the test animals can either survive or die with some probability. Either ways, the hedge fund will turn a profit that amounts to 60% of the initial investment.

The experiment is carried out successfully. The hedge fund earns a profit of 60% on top of the initially invested capital. Yet dogs and cats die with probability 0.75, i.e. few of them survive and most of them die. Did the hedge fund cause the death of most of the test animals intentionally?

YES; NO.

Task 2.4: A hedge fund decides to finance a research project for the development of a new painkiller with \$500M. The researchers involved in the project use dogs and cats as test animals. In short, the researchers test the effectiveness of the painkiller by causing some big harm to dogs and cats. Depending on how big a pain the researchers will inflict to dogs and cats, the test animals can either survive or die with some probability. Either ways, the hedge fund will turn a profit that amounts to 60% of the initial investment.

The experiment is carried out successfully. The hedge fund earns a profit of 60% on top of the initially invested capital. Yet dogs and cats die with probability 0.75, i.e. few of them survive and most of them die. Did the hedge fund cause the survival of few of the test animals intentionally?

YES; NO.

Answers (N=102)	Task 2.1	Task 2.2	Task 2.3	Task 2.4
YES	44%	59%	63%	42%
NO	56%	41%	37%	58%
<i>Significance</i>	$\chi^2 = 1.412$ (1) $p = 0.235$	$\chi^2 = 3.176$ (1) $p = 0.075$	$\chi^2 = 2.5$ (1) $p = 0.010$	$\chi^2 = 2.5$ (1) $p = 0.113$

Table 2 - Experiment 2: results

The results in Table 2 show that KE activates only in Task 2.3 because the span between YES and NO in Task 2.3 is the only statistically significant span. Indeed, while the YES are 59% in Task 2.2, there is no statistically significant span. Accordingly, there is room to argue that, regardless of the neutrality of language, KE activates only when a decision leads to a forecasted side-effect that is likely to obtain with stronger intensity. In this sense, the findings of Knobe<sup>37</sup> are correct.

### Experiment 3

The last experiment was devised in order to account mainly for the objection of language neutrality, which is only partially addressed in Task 2.1, Task 2.2, Task 2.3 and Task 2.4.

*Experiment 3* was carried out entirely online on Amazon Mechanical Turk where 69 randomly recruited individuals were asked to express their judgements concerning the following vignette cases: Task 3-6 attempt to nudge availability heuristic biases in the surveyed subjects; Task 7-8 replicate Task 1 by adding a few emotionally triggering words, e.g. investment bankers and puppies; Task 9-12 replicate the vignette of Task 2.1, 2.2, 2.3, 2.4 for a further test.

**Task 3:** A crew of firefighters is called up to extinguish a blaze that has blasted in a building where 12 people live: 4 children, 5 women (3 of which are pregnant) and 3 men.

Once the crew of firefighters reaches the building, the firefighters realize that the situation is pretty bad: the 4 children and the 3 pregnant women have remained trapped in the building. After having evaluated the gravity of the situation, the firefighters conclude that the chances of rescue success are 5%.

<sup>37</sup> Knobe, "Intentional Action and Side Effects."

## The Knobe Effect with Probable Outcomes and Availability Heuristic Triggers

Moreover, the firefighters know that they will get decorated and obtain a raise for bravery regardless of the outcomes of their action. Thus, the firefighters break into the building, but, given the situation, give up shortly after. However, they get decorated and obtain a raise for bravery.

According to you, did the firefighters intentionally act as they did just to get decorated and obtain a raise for bravery?

A) YES;

B) NO.

**Task 4:** A crew of firefighters is called up to extinguish a blaze that has blasted in a building where 12 people live: 4 children, 5 women (3 of which are pregnant) and 3 men.

Once the crew of firefighters reaches the building, the firefighters realize that the situation is pretty bad: the 4 children and the 3 pregnant women have remained trapped in the building. After having evaluated the gravity of the situation, the firefighters conclude that the chances of rescue success are 5%. Moreover, the firefighters know that they will get decorated and obtain a raise for bravery regardless of the outcomes of their action.

Nevertheless, against any rational forecast, the firefighters get into the building and manage to save the 4 children and the 3 pregnant women. Hence, they get decorated and obtain a raise for bravery.

According to you, did the firefighters intentionally act as they did just to get decorated and obtain a raise for bravery?

A) YES;

B) NO.

**Task 5:** An NGO operates in Africa where it provides locals with free vaccinations. In particular, the NGO raises funds with charity campaigns and then purchases vaccines from top pharmaceutical corporations.

According to the physicians working for the NGO, the last batch of vaccines is defective and potentially able to cause death. However, the board of the NGO does not want to ruin the good name of the NGO, which has always carried out valorous medical operations.

Thus, considering that a very bad epidemic is spreading in the countries where the NGO operates, the NGO's board decides to take the risk of handing out vaccinations to people because, in the worst case scenario, the NGO can lay the blame on its suppliers. As a result, all the people who were vaccinated survive and the name of the NGO is safe.

According to you, did the board of the NGO cause the survival of all the vaccinated people intentionally?

A) YES;

B) NO.

**Task 6:** An NGO operates in Africa where it provides locals with free vaccinations. In particular, the NGO raises funds with charity campaigns and then purchases vaccines from top pharmaceutical corporations. According to the physicians working for the NGO, the last batch of vaccines is defective and potentially able to cause death. However, the board of the NGO does not want to ruin the good name of the NGO, which has always carried out valorous medical operations.

Thus, considered that a very bad epidemic is spreading in the countries where the NGO operates, the NGO's board decides to take the risk of handing out vaccinations to people because, in the worst case scenario, the NGO can lay the blame on its suppliers. As a result, all the people who were vaccinated die. Yet the name of the NGO is safe because the press believes it's fault of the NGO's suppliers.

According to you, did the NGO cause the death of all the vaccinated people intentionally?

A) YES;

B) NO.

**Task 7:** A hedge fund run by investment bankers decides to run a project aimed at the development of a new shampoo. The fund invests \$150M in a research project that is meant to generate returns up to 50% on top of the initial investment.

The fund purchases some puppies of tigers and panthers on the black market so that the researchers involved in the research project use those puppies as test animals. Importantly, the latter shall die with a 0.81 probability, i.e. only few of them survive.

The tests are carried out successfully, the project generates the expected returns and most of the puppies die as a result of the treatments.

Did the hedge fund cause the survival of few of the puppies intentionally?

A) YES;

B) NO.

**Task 8:** A hedge fund run by investment bankers decides to run a project aimed at the development of a new shampoo. The fund invests \$150M in a research project that is meant to generate returns up to 50% on top of the initial investment.

The fund purchases some puppies of tigers and panthers on the black market so that the researchers involved in the research project use those puppies as test animals. Importantly, the latter shall survive with a 0.81 probability, i.e. few of them die.



## The Knobe Effect with Probable Outcomes and Availability Heuristic Triggers

The tests are carried out successfully, the project generates the expected returns and most of the puppies survive as a result of the treatments.

Did the hedge fund cause the death of few of the puppies intentionally?

A) YES;

B) NO.

**Task 9:** A hedge fund decides to finance a research project for the development of a new painkiller with \$500M. The researchers involved in the project use dogs and cats as test animals. In short, the researchers test the effectiveness of the painkiller by causing some big harm to dogs and cats.

Depending on how big a pain the researchers will inflict to dogs and cats, the test animals can either survive or die with some probability. Either ways, the hedge fund will turn a profit that amounts to 60% of the initial investment.

The experiment is carried out successfully. The hedge fund earns a profit of 60% on top of the initially invested capital. Yet dogs and cats survive with probability 0.75, i.e. few of them die and most of them survive. Did the hedge fund cause the survival of most of the test animals intentionally?

A) YES;

B) NO.

**Task 10:** A hedge fund decides to finance a research project for the development of a new painkiller with \$500M. The researchers involved in the project use dogs and cats as test animals. In short, the researchers test the effectiveness of the painkiller by causing some big harm to dogs and cats.

Depending on how big a pain the researchers will inflict to dogs and cats, the test animals can either survive or die with some probability. Either ways, the hedge fund will turn a profit that amounts to 60% of the initial investment.

The experiment is carried out successfully. The hedge fund earns a profit of 60% on top of the initially invested capital. Yet dogs and cats survive with probability 0.75, i.e. few of them die and most of them survive.

Did the hedge fund cause the death of few of the test animals intentionally?

A) YES;

B) NO.

**Task 11:** A hedge fund decides to finance a research project for the development of a new painkiller with \$500M. The researchers involved in the project use dogs and cats as test animals. In short, the researchers test the effectiveness of the painkiller by causing some big harm to dogs and cats.

Depending on how big a pain the researchers will inflict to dogs and cats, the test animals can either survive or die with some probability. Either ways, the hedge

fund will turn a profit that amounts to 60% of the initial investment.

The experiment is carried out successfully. The hedge fund earns a profit of 60% on top of the initially invested capital. Yet dogs and cats die with probability 0.75, i.e. few of them survive and most of them die.

Did the hedge fund cause the death of most of the test animals intentionally?

A) YES;

B) NO.

**Task 12:** A hedge fund decides to finance a research project for the development of a new painkiller with \$500M. The researchers involved in the project use dogs and cats as test animals. In short, the researchers test the effectiveness of the painkiller by causing some big harm to dogs and cats.

Depending on how big a pain the researchers will inflict to dogs and cats, the test animals can either survive or die with some probability. Either ways, the hedge fund will turn a profit that amounts to 60% of the initial investment.

The experiment is carried out successfully. The hedge fund earns a profit of 60% on top of the initially invested capital. Yet dogs and cats die with probability 0.75, i.e. few of them survive and most of them die.

Did the hedge fund cause the survival of few of the test animals intentionally?

A) YES;

B) NO.

Results	YES	NO	<i>Significance</i>
Task 3 (N=69)	51%	49%	$\chi^2 = 0.14$ (1) $p = 0.904$
Task 4 (N=69)	32%	68%	$\chi^2 = 9.058$ (1) $p = 0.003$
Task 5 (N=69)	42%	58%	$\chi^2 = 1.754$ (1) $p = 0.185$
Task 6 (N=69)	61%	39%	$\chi^2 = 3.261$ (1) $p = 0.071$
Task 7 (N=69)	25%	75%	$\chi^2 = 17.754$ (1) $p = 0.000$
Task 8 (N=69)	64%	36%	$\chi^2 = 5.232$ (1) $p = 0.022$
Task 9 (N=69)	35%	65%	$\chi^2 = 6.391$ (1) $p = 0.011$
Task 10 (N=69)	57%	43%	$\chi^2 = 1.174$ (1) $p = 0.279$
Task 11 (N=69)	70%	30%	$\chi^2 = 10.565$ (1) $p = 0.001$
Task 12 (N=69)	36%	64%	$\chi^2 = 5.232$ (1) $p = 0.022$

Table 3 - Experiment 3: results

In both Task 3-4 and Task 5-6, the Knobe Effect nullifies again as the results show that, in the harm-case, there is no dominant judgement due to the lack of

The Knobe Effect with Probable Outcomes and Availability Heuristic Triggers statistical significance. Most likely, the Knobe Effect is mitigated by the presence of both the probabilistic factor and the availability heuristic triggers in the thread of the vignette cases. Indeed, while the uncertainty factor is present, the firefighters and the NGO are unlikely to be perceived as unfriendly.

On the other hand, the same as in Task 3-6 is much evident in the reverse way. Indeed, in task 7-8 the hedge fund is run by investment bankers and there are no more dogs and cats, but puppies of panthers and tigers purchased on the black market. In this case, the Knobe Effect obtains regardless of the fact that the hedge fund takes a business decision with probable outcomes.

Eventually, once Task 2.1, 2.2, 2.3, 2.4 are repeated in Task 9-12, the experimental results of *Experiment 2* are confirmed. For, QED, the Knobe Effect obtains only for the side-effects that are likely to obtain with strong intensity.

### Concluding Remarks

In the light of the results presented in the previous section, there is room to argue that the way folks perceive intentionality might be driven by some stereotypes concerning the agent who carries some action. In this sense, a firm's VP is likelier to look more unfriendly than an NGO who operates in underdeveloped countries or than a crew of firefighters. Moreover, it seems that if two outcomes (one big and one small) take place simultaneously, then ordinary folks judge the bigger outcome as more intentional than the smaller outcome. This is the case once the protagonists of the vignette take a decision with probable outcomes and different intensity. Eventually, the presence of triggering words (e.g. harm-help or similar) affects judgement. Thus, there is room to argue that the Knobe Effect is sensitive to framing and heuristic-related problems.<sup>38,39</sup>

---

<sup>38</sup> The first experiment of this study was partially supported by the National Science Centre of Poland (NCN) under the Grant [DEC-2015/17/D/HS6/02684] assigned to Lukasz Markiewicz, PhD (Kozminski University) who has kindly shared part of his funding to support this study. The second and the third experiment of this study were partially supported with the internal funds BST 2017 assigned to the Department of Modern Philosophy of the University of Warsaw. Hence, we would like to thank Lukasz Markiewicz, PhD (Kozminski University) and Prof. Marcin Poręba (University of Warsaw) for funding our work.

<sup>39</sup> We would like to thank the reviewers of this study for their time and professionalism. Last but not least, we are very much obliged towards Prof. Domenico Buccella (Kozminski University), Prof. Katarzyna Paprzycka (University of Warsaw), Konrad Werner, PhD (University of Warsaw), and Adrian Ziolkowski (University of Warsaw) for their insightful comments about the earlier versions of this study, which eventually made us understand where our research was heading. Thank you all for your help, time and support.



# SURREALISM IS NOT AN ALTERNATIVE TO SCIENTIFIC REALISM

Seungbae PARK

**ABSTRACT:** Surrealism holds that observables behave as if T were true, whereas scientific realism holds that T is true. Surrealism and scientific realism give different explanations of why T is empirically adequate. According to surrealism, T is empirically adequate because observables behave as if it were true. According to scientific realism, T is empirically adequate because it is true. I argue that the surrealist explanation merely clarifies the concept of empirical adequacy, whereas the realist explanation makes an inductive inference about T. Therefore, the surrealist explanation is a conceptual one, whereas the realist explanation is an empirical one, and the former is not an alternative to the latter.

**KEYWORDS:** empirical adequacy, observables, scientific realism, surrealism, truth

## 1. Introduction

The term ‘surrealism’ refers to a philosophical position that is meant to be a surrogate for scientific realism.<sup>1</sup> This paper defines it as the view that observables behave as if T were true, and scientific realism as the view that T, a theory, is true. Surrealism is regarded as an alternative to scientific realism not only by Jarrett Leplin<sup>2</sup> but also by other eminent philosophers, such as Alan Musgrave,<sup>3</sup> P. Kyle Stanford,<sup>4</sup> Timothy Lyons,<sup>5</sup> and Moti Mizrahi.<sup>6</sup> This paper exposes a problem with

---

<sup>1</sup> Jarrett Leplin, “Surrealism,” *Mind* 97, 384 (1987): 519–524.

<sup>2</sup> Leplin, “Surrealism.”

<sup>3</sup> Alan Musgrave, “The Ultimate Argument for Scientific Realism,” in *Relativism and Realism in Science*, ed. Robert Nola (Dordrecht: Kluwer Academic Publishers, 1988), 229–252; Alan Musgrave, “Strict Empiricism Versus Explanation in Science,” in *Varieties of Scientific Realism: Objectivity and Truth in Science*, ed. Evandro Agazzi (Switzerland: Springer International Publishing, 2017), 71–93.

<sup>4</sup> P. Kyle Stanford, “An Antirealist Explanation of the Success Science,” *Philosophy of Science* 67, 2 (2000): 266–284.

<sup>5</sup> Timothy Lyons, “Explaining the Success of a Scientific Theory,” *Philosophy of Science* 70, 5 (2003): 891–901.

<sup>6</sup> Moti Mizrahi, “Why the Ultimate Argument for Scientific Realism Ultimately Fails,” *Studies in History and Philosophy of Science Part A* 43, 1 (2012): 132–138.

surrealism, thereby presenting an important philosophical lesson – we should distinguish between two kinds of explanations: conceptual and empirical ones.

Philosophers have proposed surrealism as a way of explaining why T is successful and why T is empirically adequate. There are many differences between these two explananda. One of them is that the success of T implies that *some* observational consequences of T are true, whereas the empirical adequacy of T implies that *all* observational consequences of T are true. The history of science abounds in successful theories that were empirically inadequate.<sup>7</sup> For example, the Ptolemaic theory and the miasma theory were successful, but empirically inadequate. In addition to the truth of some observational consequences, T must meet other conditions to be successful, e.g., the auxiliary condition, the technological condition, and the financial condition. I only bring readers' attention to Park<sup>8</sup> for the explication of these other conditions.

This paper is concerned not with the surrealist explanation that T is successful because observables behave as if it were true, but with the surrealist explanation that T is empirically adequate because observables behave as if it were true. The former has already been criticized in detail.<sup>9</sup> Put briefly, scientists deserve credit for the success of T, but the surrealist explanation attributes the credit not to scientists but to the world, thereby disappointing scientists. To use an analogy, imagine that the Wright brothers worked hard to invent the airplane, but surrealists came along and said to the Wright brothers that the air plane could fly "because there was air in the sky."<sup>10</sup> Such an explanation would have failed to recognize the Wright brothers' accomplishment and would have disappointed them.

The outline of this paper is as follows. In Section 2, I appeal to the correspondence theory of truth to argue that saying that T is true is different from saying that the world is as T says it is. In Section 3, I argue that saying that T is empirically adequate is also different from saying that observables behave as if it were true. Hence, it is not a circular explanation that T is empirically adequate because observables behave as if it were true, contrary to what Musgrave<sup>11</sup> contends. In Section 4, I argue that the surrealist explanation is a trivial one for

---

<sup>7</sup> Marc Lange, "Baseball, Pessimistic Inductions and the Turnover Fallacy," *Analysis* 62, 4 (2002): 282; Lyons, "Explaining the Success of a Scientific Theory," 898.

<sup>8</sup> Seungbae Park, "Realism Versus Surrealism," *Foundations of Science* 21, 4 (2016): 604–606.

<sup>9</sup> Park, "Realism Vs. Surrealism," 610–614.

<sup>10</sup> Park, "Realism Vs. Surrealism," 612.

<sup>11</sup> Musgrave, "The Ultimate Argument for Scientific Realism;" Musgrave, "Strict Empiricism Versus Explanation."

those who are already familiar with the concept of empirical adequacy. In addition, I distinguish between conceptual and empirical explanations, classifying the surrealist explanation as conceptual and the realist explanation as empirical. In Section 5, I reply to two objections. This paper can be useful to those who are interested in whether surrealism is an alternative to realism, under what conditions an explanation is appropriate, and how conceptual explanations differ from empirical ones.

## 2. The Correspondence Theory of Truth

If you ask correspondentists, theorists who espouse the correspondence theory of truth, to explain why T is true, they will put forward the correspondentist explanation that T is true because it corresponds to the world, i.e., because the world is as T says it is. The correspondentist explanation is composed of the following two statements:

(T) T is true.

(W) The world is as T says it is.

Are (T) and (W) substantially different assertions? Or are they merely different expressions of the same assertion? In my view, they are substantially different assertions. (T) is an assertion about T, whereas (W) is an assertion about the world. (T) attributes a semantic property to T, whereas (W) attributes a certain manner of existence to the world. (T) and (W) cannot be mere verbal variants because they are different assertions about different targets.

If (T) and (W) were mere verbal variants, the correspondence theory would be a vacuous theory of truth. The correspondence theory, however, is not a vacuous theory of truth. It rather makes a substantive claim about what makes a statement true, viz., the world is what makes a statement true. Unlike other theories of truth, it claims that the world serves as the truth-maker for true statements.<sup>12</sup> Of course, if correspondentists believe (T), they can infer (W), and vice versa. After all, that is what it means to embrace the correspondence theory. It does not follow, however, that (W) is merely a verbal variant of (T). It is one thing that we can infer (W) from (T) and vice versa; it is another that they are mere verbal variants.

When correspondentists propose that T is true if and only if the world is as T says it is, they are engaged in a conceptual analysis of the concept of truth. They aim to identify the necessary and sufficient conditions for the truth of T. To this

---

<sup>12</sup> Alvin Goldman, *Knowledge in a Social World* (Oxford: Oxford University Press, 1999), 61.

end, they claim that if the world were not as T says it is, T would not be true, i.e., that correspondence to the world is a necessary condition for the truth of T. They also claim that if the world were as T says it is, T would be true, i.e., that correspondence to the world is a sufficient condition for the truth of T. A conceptual analysis is not an *a posteriori* enterprise but an *a priori* enterprise. While an *a posteriori* enterprise involves an investigation into the world, an *a priori* enterprise does not. Correspondentists are not making any inductive inferences about the world, but are laying bare the concept of truth.

Consider the proposal that the special theory of relativity is true because the world is as it says it is. Does this proposal merely repeat the same assertion? Or does it say something interesting about why the special theory of relativity is true? If you think that (W) is just a fancy way of saying (T), you would immediately think that it is vacuous to say that the special theory of relativity is true because the world is as it says it is. By contrast, if you think that (T) and (W) are substantially different assertions, you would think that the proposal says something interesting about why the special theory of relativity is true. (T) and (W) are substantially different assertions, as we have seen above. Therefore, it is not circular to say that the special theory of relativity is true because the world is as it says it is.

This conclusion will serve as a theoretical resource for me to refute Musgrave's objection to surrealism in the next section.

### 3. The Refutation of Musgrave's View

What does it mean to say that observables behave as if T were true? It means "that observable events occur as T says they do."<sup>13</sup> What about unobservable events? They may or may not occur as T says they do, i.e., it is open to question whether unobservables behave or do not behave as T says they do. T would be true if both observables and unobservables behave as it says they do. However, in order for T to be empirically adequate, it is only necessary that observables behave as T says they do. What if observables behave as T says they do, but unobservables do not behave as T says they do? T would be empirically adequate but false. Thus, surrealists believe that T is empirically adequate, but do not believe that it is true.

Now that we are clear about the content of surrealism, we are ready to appraise the surrealist explanation that T is empirically adequate because observables behave as if it were true. The surrealist explanation is comprised of the following two statements:

---

<sup>13</sup> Park, "Realism Vs. Surrealism," 606.



(E) T is empirically adequate.

(O) Observables behave as if T were true.

Musgrave contends that (E) and (O) are not substantially different assertions but mere verbal variants. For him, saying that observables behave as if T were true “is just a fancy way of saying that *T* is observationally or empirically adequate.”<sup>14</sup> He insists that “saying that the phenomena are *as if* the theory were true is just saying that the theory is empirically adequate.”<sup>15</sup> He also maintains that “to say that a theory is empirically adequate is just to say that the phenomena are *as if* it were true.”<sup>16</sup>

Musgrave’s linguistic intuition led him to the view that (E) and (O) are merely different formulations of the same assertion, and his linguistic intuition is not groundless. We can *infer* (O) from (E) and vice versa. For example, the belief that the special theory of relativity is empirically adequate entitles us to infer that observables behave as if it were true. The belief that observables behave as if it were true entitles us to infer that it is empirically adequate. After all, that is what it is to embrace (E) or (O). So it appears that (E) and (O) are mere verbal variants. Musgrave’s view about (E) and (O) holds an important implication regarding the surrealist explanation. If his view is true, the surrealist explanation is circular, i.e., (O) is (E) in disguise. Hence, the surrealist explanation amounts to explaining “the empirical adequacy of a theory in terms of its empirical adequacy.”<sup>17</sup>

In my view, however, (E) and (O) are not mere verbal variants, but substantially different assertions. (E) is an assertion about T, whereas (O) is an assertion about the world. (E) claims that T has a certain semantic property, viz., empirical adequacy. By contrast, (O) claims that observables behave in a certain manner. Thus, (E) and (O) are different claims about different targets. Consider also that (E) is merely the restriction of (T) to observational claims, while (O) is merely the restriction of (W) to observables. So if (T) and (W) are substantially different assertions, (E) and (O) are also substantially different assertions. As we have seen in Section 2, (T) and (W) are substantially different assertions. Therefore, (E) and (O) are also substantially different assertions, *pace* Musgrave.

Musgrave takes (E) and (O) to be mere verbal variants, despite the fact that (E) is a claim about T, whereas (O) is a claim about the world. So it is natural for him to suggest that the truth of T explains why observables behave as if it were

---

<sup>14</sup> Musgrave, “The Ultimate Argument for Scientific Realism,” 243.

<sup>15</sup> Musgrave, “Strict Empiricism Versus Explanation,” 78.

<sup>16</sup> Musgrave, “Strict Empiricism Versus Explanation,” 76.

<sup>17</sup> Musgrave, “Strict Empiricism Versus Explanation,” 84.

true. He says, "T's actually being true is the best explanation of why all the observable phenomena are as if it were true."<sup>18</sup> Note that he explains the behavior of the world in terms of the semantic property of T.

In my view, however, it is wrong to do so. The world behaves as it does irrespective of how we describe it. For example, heat is as it is, and it behaves as it does, regardless of whether we describe it as caloric fluid or as the mean kinetic energy of molecules. It is incoherent to say that cold and hot objects in contact with each other assume the average temperature because the kinetic theory is true. By contrast, it is coherent to say that cold and hot objects assume the average temperature because the fast-moving molecules of the hot object slow down and the slow-moving molecules of the cold object move faster. In general, an event should be explained not in terms of a semantic property but in terms of another event.

Of course, we can make an *inference* from the truth of T to the behavioral pattern of observables. It does not follow, however, that we can *explain* the behavioral pattern of observables in terms of the truth of T. It has become an accepted point in philosophy of science that inference and explanation are two different affairs. As Sylvain Bromberger<sup>19</sup> has pointed out, it is legitimate to infer the length of the flagpole from the length of the shadow, but illegitimate to explain the length of a flagpole in terms of the length of its shadow.

Let me present a thought experiment to make my foregoing objection more convincing. Imagine a possible world in which God changes the way the world behaves via changing the truth-values of T. For example, God invests the theory of gravity with truth during the day but with falsity at night, so an apple falls downwards in the daytime, but rises upwards at night. God does not directly change the way the world behaves. He rather does so by changing the truth-values of the theory of gravity. Thus, the semantic property of the theory of gravity is the immediate cause of the way this possible world behaves. In such a possible world, it would be legitimate to explain an event in terms of a semantic property. For example, it would make perfect sense to say that the apple falls down because the theory of gravity is true.

In the actual world, however, it is wrong to say that observables behave as T says they do because it is true, or to say that observables behave as if T were true because it is empirically adequate. Such explanations are all conceptually flawed. It is only legitimate to explain the semantic property of the truth or empirical

---

<sup>18</sup> Musgrave, "Strict Empiricism Versus Explanation," 83.

<sup>19</sup> Sylvain Bromberger, "Why Questions," in *Mind and Cosmos*, ed. R. G. Colodney (Pittsburgh, PA: University of Pittsburgh Press, 1966).

adequacy of T in terms of how the world behaves. So we can say that T is true because the world behaves as T says it does, or that T is empirically adequate because observables behave as if it were true. In other words, in the actual world, (W) can explain (T), but not vice versa, and (O) can explain (E), but not vice versa. This asymmetric explanatory relation of (T) and (W) further shows that they are substantially different assertions, and that the correspondentist explanation is not circular. Similarly, the asymmetric explanatory relation of (E) and (O) further shows that they are substantially different assertions, and that the surrealist explanation is not circular either.

One might attempt to defend Musgrave's view about the surrealist explanation by appealing to deflationism, an alternative to the correspondence theory. According to deflationism, 'It is true that p' means no more than p, i.e., 'It is true that p' and 'p' are equivalent statements. It follows that (T) and (W) are equivalent. Given that (E) and (O) are just the restrictions of (T) and (W) to observables, (E) and (O) are also equivalent. It follows that (E) and (O) are mere verbal variants. Thus, under deflationism, Musgrave is right after all.

It is doubtful, however, that Musgrave would endorse the preceding deflationist defense of his view that the surrealist explanation is circular. He says that "The aim of science, realists tell us, is to have true theories about the world, where 'true' is understood in the classical correspondence sense."<sup>20</sup> In short, Musgrave operates under the correspondence theory when he argues that (E) and (O) are mere verbal variants, so the surrealist explanation is circular.

So far, I have argued that (E) and (O) are substantially different assertions, so the surrealist explanation is not circular. Let me now turn to the confrontation between the surrealist explanation and the realist explanation that T is empirically adequate because it is true. The surrealist explanation invokes the behavioral pattern of observables, whereas the realist explanation invokes the truth of T, to explain why T is empirically adequate. Neither the surrealist explanation nor the realist explanation suffers from a conceptual problem.

André Kukla,<sup>21</sup> Lyons,<sup>22</sup> and Mizrahi<sup>23</sup> would take the surrealist explanation as a serious alternative to the realist explanation. Kukla states that the "observable world behaves as if our theories are true."<sup>24</sup> In a similar vein, Lyons claims that the "mechanisms postulated by the theory and its auxiliaries would, if actual, bring

---

<sup>20</sup> Musgrave, "The Ultimate Argument for Scientific Realism," 229.

<sup>21</sup> André Kukla, *Studies in Scientific Realism* (New York: Oxford University Press, 1998).

<sup>22</sup> Lyons, "Explaining the Success of a Scientific Theory."

<sup>23</sup> Mizrahi, "Why the Ultimate Argument for Scientific Realism Ultimately Fails."

<sup>24</sup> Kukla, *Studies in Scientific Realism*, 22.

about all relevant phenomena thus far observed and some yet to be observed at time  $t$ ; and these phenomena are brought about by actual mechanisms in the world.”<sup>25</sup> Mizrahi also says that the “observable world behaves as if our mature scientific theories are true.”<sup>26</sup> He would say that there is no good reason to prefer the realist explanation over the surrealist explanation because the realist explanation is empirically no better than the surrealist explanation, i.e., “both make the same testable predictions.”<sup>27</sup>

Which one is better, the realist explanation or the surrealist explanation? Musgrave prefers the realist explanation to the surrealist explanation on the grounds that the surrealist explanation is circular. He argues that “truth explains empirical adequacy better than empirical adequacy does, because the latter ‘explanation’ is completely circular.”<sup>28</sup> As we have seen, however, the surrealist explanation is not vacuous. Hence, we are still left with the question: which explanation is better? I defend my answer to this question in the next section.

#### 4. The Real Problem

In general, an explanation is appropriate when it serves the explainers’ purposes and/ or the explainees’ purposes, and is inappropriate when it serves the purposes of neither. Suppose that a jet airliner crashes, and that investigators rush to the crash site. After investigating the wreckage, they hold a news conference. They announce, to the surprise of news reporters, that the jet airliner crashed due to the gravitational force between it and the Earth. This explanation, although conceptually sound, is inappropriate because it serves neither the explainers’ nor the explainees’ purposes. It merely makes an obvious point that interests neither the explainers nor the explainees. Such an explanation might, however, be appropriate in science classrooms, in which teachers aim to convey the concept of gravity to students. It would serve both the teachers’ purpose to teach the concept of gravity and the students’ purpose to learn the new concept. This story suggests that explainers’ and explainees’ purposes determine whether an explanation is appropriate or not.

This general point applies to the surrealist explanation. The surrealist explanation is appropriate when it serves the explainers’ and/or explainees’ purposes, and is inappropriate when it serves neither. Suppose that professors wish to share the concept of empirical adequacy with students in a philosophy of science

---

<sup>25</sup> Lyons, “Explaining the Success of a Scientific Theory,” 900.

<sup>26</sup> Mizrahi, “Why the Ultimate Argument for Scientific Realism Ultimately Fails,” 133.

<sup>27</sup> Mizrahi, “Why the Ultimate Argument for Scientific Realism Ultimately Fails,” 133.

<sup>28</sup> Musgrave, “Strict Empiricism Versus Explanation,” 87.

class. Under these circumstances, the surrealist explanation would be appropriate because it would prove illuminating to students who were previously unfamiliar with the concept of empirical adequacy, enabling them to grasp both the relationship between empirical adequacy and observables, and the relationship between empirical adequacy and truth. The surrealist explanation would serve both the explainers' purpose and the explainees' purpose.

What if the explainers and explainees are already familiar with the concept of empirical adequacy? The surrealist explanation, although conceptually sound, would be inappropriate. It would merely make an obvious point that interests no one, just as the investigators' gravitational explanation above makes an obvious point that interests no one. Hence, the surrealist explanation would serve no one's purpose.

Surrealists might reply that even if explainers and explainees are already familiar with the concept of empirical adequacy, the surrealist explanation can nevertheless serve a certain purpose, viz., to undermine the no-miracles argument.<sup>29</sup> The no-miracles argument was originally constructed to explain not the empirical adequacy of T but the success of T. It, however, can be recast to explain the empirical adequacy of T. The recast version would hold that the empirical adequacy of T would be a miracle if T were false, so the truth of T best explains the empirical adequacy of T. The argument maintains "not just that truth explains empirical adequacy, but that it is the only explanation, or at least the best explanation."<sup>30</sup> Surrealists, by providing an alternative to the realist explanation, have imposed upon scientific realists the burden of proving that the realist explanation is better than the surrealist explanation. Consequently, the surrealist explanation is an appropriate one in the scientific realism debate.

It is, however, debatable whether the surrealist explanation is an alternative to the realist explanation. The surrealist explanation is a conceptual analysis laying bare the necessary and sufficient conditions for the empirical adequacy of T, just as the correspondentist explanation is a conceptual analysis laying bare the necessary and sufficient conditions for the truth of T. Surrealists do not make an inductive inference about T, any more than correspondentists make an inductive inference about T. After all, the surrealist explanation is just the restriction of the correspondentist explanation to observational claims and observables. It is for this reason that if you were already familiar with the concept of empirical adequacy, you would immediately accept the surrealist explanation. If you do not accept the

---

<sup>29</sup> Hilary Putnam, *Mathematics, Matter and Method (Philosophical Papers, vo. 1)* (Cambridge: Cambridge University Press, 1975), 73.

<sup>30</sup> Musgrave, "Strict Empiricism Versus Explanation," 84.

surrealist explanation, that shows not that you refuse to make an inductive inference about T, but that you do not understand what it means to say that T is empirically adequate. In short, the surrealist explanation is not an inductive inference but a clarification of a concept.

By contrast, when realists advance the realist explanation, they are in the business of making an inductive inference about T. From the premise that T is empirically adequate, they inductively infer that it is true. They are not in the business of clarifying the concept of empirical adequacy. They do not say that the explanans is the necessary and sufficient condition for the explanandum. After all, it is obviously false that the truth of T is the necessary and sufficient condition for the empirical adequacy of T. The realist explanation involves an inductive inference from the empirical adequacy of T to the truth of T. For this reason, antirealists reject the realist explanation, even if they are already familiar with the concept of empirical adequacy. They reject the realist explanation not because they do not understand what it means to say that T is empirically adequate, but because they are not willing to run the epistemic risk involved in the inductive inference. In short, the realist explanation involves not a clarification of a concept but an inductive inference.

The difference between the surrealist explanation and the realist explanation discussed above calls for the distinction between what I call conceptual and empirical explanations. A conceptual explanation is an attempt to illuminate a concept by providing a necessary and a sufficient condition for it. The former is called an *analysandum*, and the latter is called an *analysans*. No inductive inference is made from the *analysandum* to the *analysans*. By contrast, an empirical explanation is an attempt to illuminate an explanandum by providing an explanans. An inductive inference is made from the explanandum to the explanans. The surrealist explanation exemplifies a conceptual explanation, whereas the realist explanation exemplifies an empirical explanation.

Consider now how the surrealist explanation and the realist explanation could be refuted. We can conceive of some counterexamples, some scientific theories, that drive a wedge between the explanandum and the explanans of the realist explanation. Suppose that von Neumann and Dirac's version of quantum mechanics is empirically adequate, and that it is empirically equivalent to Bohm's version of quantum mechanics. Given that they make incompatible claims about unobservables, they are empirically adequate rivals, and they would constitute counterexamples undermining the realist inference from the empirical adequacy of T to the truth of T, i.e., from the explanandum to the explanans of the realist explanation. In contrast, we cannot even conceive of counterexamples

undercutting the surrealist inference from the analysandum to the analysans. Suppose that critics of surrealism present certain scientific theories, and then say that though they are empirically adequate, observables do not behave as if the theories were true. What they say would indicate not that the scientific theories are counterexamples to the surrealist explanation, but that they do not know what it is for T to be empirically adequate. In short, the realist explanation is subject to an empirical refutation whereas the surrealist explanation is not. This difference provides further support for my view that the surrealist explanation is a conceptual one, whereas the realist explanation is an empirical one.

What can we conclude from my view that with the surrealist explanation, surrealists are engaged in the *a priori* enterprise of clarifying the concept of empirical adequacy, whereas with the realist explanation, realists are in the *a posteriori* enterprise of making an inductive inference from the empirical adequacy of T to the truth of T? We can conclude that the surrealist explanation cannot be an alternative to the realist explanation. To use an analogy, there are many kinds of apples: Red Delicious, Granny Smith, Yellow Newton, etc. Suppose that you claim that Red Delicious is the most delicious apple. I present you with an orange, and request that you prove that Red Delicious apples taste better than the orange. You would immediately object that my request is illegitimate, saying that you were talking about apples, but not about oranges. Realists can say the same thing about surrealists' request to prove that the realist explanation is better than the surrealist explanation. When realists say that the realist explanation is the best explanation of the empirical adequacy of T, they mean that the realist explanation is the best *empirical* explanation. The surrealist explanation is not an empirical one but a conceptual one. It follows that the surrealist explanation cannot be an alternative to the realist explanation, and that it is wrong to say that the surrealist explanation undermines the realist contention that the realist explanation is the best empirical explanation of the empirical adequacy of T.

Surrealists might now go on the offensive against the realist explanation. What purpose does the realist explanation serve? My answer is that it serves the realist purpose of arriving at the realist explanans that T is true, i.e., realists claim that T is true on the grounds that the truth of T best explains the empirical adequacy of T. Since the realist explanation serves the realist purpose of supporting the truth of T, it is appropriate in the scientific realism debate.

Such a defense cannot be made for the surrealist explanation. Surrealists cannot say that the surrealist explanation serves the surrealist purpose of arriving at the analysans that observables behave as if T were true. After all, the analysans is nothing but the necessary and sufficient condition for the analysandum that T is

empirically adequate. The surrealist explanation only clarifies the concept of empirical adequacy, just as the correspondentist explanation only clarifies the concept of truth. It follows that surrealists can only say that the surrealist explanation serves the purpose of analyzing the concept of empirical adequacy, just as the correspondentist explanation serves the purpose of analyzing the concept of truth.

Interestingly, the surrealists' analysandum coincides with the realists' explanandum. Both realists and surrealists are explaining why T is empirically adequate. Clarifying the concept of empirical adequacy is not only what surrealists should do, but also what realists should do. After all, if the concept of empirical adequacy is obscure, it is pointless for realists to say that the truth of T best explains the empirical adequacy of T. It follows that surrealists are helping realists by providing the surrealist explanation of empirical adequacy, and that realists should endorse the surrealist explanation.

## 5. Objections and Replies

I argued above that (E) and (O) are not mere verbal variants but substantially different assertions. Recall that (E) and (O) are as follows:

(E) T is empirically adequate.

(O) Observables behave as if T were true.

Surrealists might insist that (E) and (O) are mere verbal variants on the grounds that they parallel (1) and (2):

(1) A term refers.

(2) The world contains something that is picked out by the term.

(1) and (2) are mere verbal variants, although (1) is a claim about a term, whereas (2) is a claim about the world. It follows that (E) and (O) are also mere verbal variants, although (E) is a claim about T, whereas (O) is a claim about the world.

My replies are two-fold. First, (1) and (2) parallel the analysandum and the analysans of the correspondentist explanation, (T) and (W):

(T) T is true.

(W) The world is as T says it is.

It follows that if (1) and (2) were mere verbal variants, (T) and (W) would also be mere verbal variants. As we have seen in Section 2, however, (T) and (W) are not mere verbal variants but substantially different assertions. Therefore, (1) and (2) are also not mere verbal variants but substantially different assertions.



Second, just as (W) is more fundamental than (T), so (2) is more fundamental than (1). It follows that just as (W) can explain (T), but not vice versa, so (2) can explain (1), but not vice versa. It sounds right to say that T is true because the world is as T says it is. In contrast, it sounds wrong to say that the world is as T says it is because T is true. Analogously, it sounds right to say that a term refers because the world contains something that is picked out by the term. By contrast, it sounds wrong to say that the world contains something that is picked out by a term because the term refers. This asymmetrical explanatory relation between (1) and (2) further indicates that (1) and (2) are substantially different assertions.

Moreover, it sounds wrong to say that an object exists because its term refers in the actual world. It sounds right to say so only in a possible world in which God makes objects come into being and pass out of being by changing the semantic properties of terms. For example, imagine that God makes 'electron' refer during the day and not refer at night. As a result, an electron exists during the day, but does not exist at night. In such a possible world, it makes perfect sense to say that an electron exists because 'electron' refers.

Let me now turn to a different objection. Berkeleyan idealists would say that T is empirically adequate not because observables behave as if it were true, but because God implants certain ideas in my mind as if it were true. Surrealists would retort that T is empirically adequate not because God implants certain ideas in my mind as if it were true, but because observables behave as if it were true. Thus, surrealists make an inductive inference about the world, i.e., they inductively infer that observables, as opposed to certain ideas in my mind, behave as if T were true. Since the surrealist explanation makes an inductive inference, it is an empirical explanation, and it is an alternative to the realist explanation.

This objection, although brilliant, can be reduced to absurdity. If the surrealist explanation were an empirical one for the reason stated above, the correspondentist explanation would also be an empirical one for a similar reason. When correspondentists say that T is true because the world is as T says it is, they are making an inductive inference about the world, i.e., they are inductively inferring that the material world, as opposed to the ideal world, is as T says it is. To go further, it would also be an empirical explanation to say that John is a bachelor because he is an unmarried adult male. When you give this explanation, you are inductively inferring that the combination of John's body and mind, as opposed to a collection of my ideas, is an unmarried adult male. These two explanations, however, are not empirical ones but conceptual ones. Therefore, the surrealist explanation is also a conceptual one.

The fact that the foregoing objection falls prey to a *reductio ad absurdum* indicates that there is an intrinsic problem with it. The intrinsic problem is that it involves a sudden change of frameworks from the materialist framework to the idealist framework. It is due to this change of frameworks that the surrealist explanation becomes an empirical one making an inductive inference about the world. If the framework did not change, the surrealist explanation would consistently be a conceptual one clarifying the concept of empirical adequacy.

Let me flesh out this abstract point. Under the materialist framework, to say that T is empirically adequate means that observables behave as if it were true. It does not mean that certain ideas occur in my mind as if T were true. After all, T would not be empirically adequate, even if certain ideas occurred in my mind as if it were true, if observables did not behave as if it were true. Once surrealists adopt the materialist framework and say that T is empirically adequate, they have no choice but to say that T is empirically adequate because observables behave as if it were true. They cannot say that T is empirically adequate because certain ideas occur in my mind as if it were true. To say so is to change the framework suddenly from the materialist framework to the idealist framework.

By contrast, under the idealist framework, to say that T is empirically adequate means that certain ideas occur in my mind as if it were true. This does not mean that observables, immaterial objects, behave as if T were true. Once surrealists adopt the idealist framework and say that T is empirically adequate, they have no choice but to say that T is empirically adequate because certain ideas occur in my mind as if it were true. They cannot say that T is empirically adequate because observables behave as if it were true. To say so is to change the framework suddenly from the idealist framework to the materialist framework.

In short, if surrealists interpret the analysandum, (E), under the materialist framework, they should provide a materialist analysans. If they interpret it under the idealist framework, they should provide an idealist analysans. Following these rules would inevitably result in conceptual explanations.

## 6. Conclusion

Just as the correspondentist explanation makes it clear that the world makes T true, so the surrealist explanation makes it clear that the world makes T empirically adequate. In other words, just as the correspondentist explanation claims that the world is the truth-maker of T, so the surrealist explanation claims that observables are the empirical-adequacy-maker of T. It follows that it is not circular to say that T is empirically adequate because observables behave as if it were true, any more than it is circular to say that T is true because the world is as T says it is. By making

these claims, both correspondentists and surrealists are engaged in conceptual analyses, attempting to lay bare the necessary and sufficient conditions for the truth and empirical adequacy of T, respectively.

The surrealist explanation is a conceptual one, whereas the realist explanation is an empirical one. The surrealist explanation merely clarifies the analysandum in terms of the analysans, whereas the realist explanation involves an inductive inference from the explanandum to the explanans. The surrealist explanation is a trivial one for those who are already familiar with the concept of empirical adequacy, whereas the realist explanation is not a trivial one for those who are already familiar with the concept of truth. In sum, the surrealist explanation is different in kind from the realist explanation, and surrealism is not an alternative to scientific realism.<sup>31</sup>

---

<sup>31</sup> Acknowledgements: I thank anonymous referees of this journal for helpful comments. This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A5A2A01039606).



# UNSTABLE KNOWLEDGE, UNSTABLE BELIEF

Hans ROTT

**ABSTRACT:** An idea going back to Plato's *Meno* is that knowledge is stable. Recently, a seemingly stronger and more exciting thesis has been advanced, namely that rational belief is stable. I sketch two stability theories of knowledge and rational belief, and present an example intended to show that knowledge need not be stable and rational belief need not be stable either. The second claim does not follow from the first, even if we take knowledge to be a special kind of rational belief. 'Stability' is an ambiguous term that has an internally conditional structure.

**KEYWORDS:** knowledge, rational belief, subjective probability, stability

## 1. The Example

Consider the following story that ramifies into two alternative versions.

Yesterday afternoon, at four o'clock, Sam looked out of his window and saw his neighbours Ann and Ben passing by (or so he thought). Sam could see the couple very clearly in the bright sunshine. It did not occur to him at all that he might mistake some other people for his neighbours. Still, he starts doubting now whether it was really Ann and Ben who he saw yesterday. Mia, a very serious and reliable person and a very close friend of Ann and Ben's, just told Sam that ...

*(Version 1)* ... it wasn't Ann and Ben who were passing by. Mia did not want to give Sam more information, but there is no doubt that what she said is true. Sam knows Ann really well, so he is reluctant to call into question that he saw her. And exactly the same is true for Ben. Still Sam concludes, with some amazement, that it must have been another man or another woman who he saw passing by his window. As a matter of fact, the woman walking past his window was indeed Ann, but the man was Ben's twin brother Bob.

*(Version 2)* ... Ann and Ben had to present their joint paper in a Graduate Workshop at the university at 4 p.m. yesterday. Since there is no question that what Mia said is true, it is doubtful whether Ann and Ben could have been in the neighbourhood at four o'clock. Sam reconsiders the situation, and even though he still thinks there is a fair chance that it was Ann and Ben who he saw, it does not appear unlikely to him that he mistook some other persons for them.

From the description of the scenarios, it is clear that Sam *fully believed* yesterday that Ann and Ben were passing by his window, and he was *fully justified* and *rational* in so believing. In addition, it seems that Sam in fact *knew* yesterday that Ann was passing by in version 1 of the story, notwithstanding his later retraction of the belief that he saw her. His view of Ann was completely unimpaired, he could recognise her clearly, and it was in fact Ann who he saw. His successful identification of Ann is not undermined by his bad luck with the (mis)identification of Ben.

We shall see that if these intuitions about Sam's propositional attitudes are right, the story goes against the idea of *stability* that some authors have suggested to be a necessary condition for knowledge or for rational belief. In the next section, I will briefly review a stability theory of knowledge fathomed by a number of recent authors. I then use Version 1 of our story to show that stability is not necessary for knowledge. In Section 3, I present a stability theory of rational belief recently proposed and developed by Hannes Leitgeb.<sup>1</sup> The second version of our story will then be employed to show that belief need not be stable either. Section 4 clarifies the relationship between the two kinds of theories by distinguishing various meanings of the predicate 'stable.' Assuming that knowledge entails rational belief, the existence of unstable knowledge seems to entail that stability cannot be a necessary condition for rational belief either. But Section 5 explains why such an inference would be fallacious. Version 2 of the story is indeed needed for my argument that rational belief need not be stable.

## 2. The Stability Theory of Knowledge

In Plato's *Meno*, stability is suggested as a feature of knowledge that makes it more valuable than merely true belief.<sup>2</sup> Contemporary epistemological writings have rarely considered stability as a part of the definition (or nature) of knowledge. In recent semantic modellings of epistemic states, by contrast, the stability condition has been the topic of considerable discussion. Stability is defined here with reference to a multitude of possible worlds, which makes it a modal concept. Referring to a then unpublished paper of Stalnaker's, Lamarre and Shoham provided an axiomatisation and a semantics reflecting the idea that "knowledge is

---

<sup>1</sup> Leitgeb's theory is a theory about *rational belief*, even if he frequently just calls it a theory about *belief* (see the titles of his works quoted below).

<sup>2</sup> See Casey Perin, "Knowledge, Stability, and Virtue in the *Meno*," *Ancient Philosophy* 32, 1 (2012): 15–34, and the references cited therein.

belief that is ‘stable with respect to the truth.’”<sup>3</sup> Stalnaker seems to have been the first author presenting the idea of the stability analysis of knowledge:

[... an agent] *a* knows that  $\phi$  if and only if *a* believes that  $\phi$  [...], and that belief is robust with respect to the truth. [...] More precisely, the proposition that *a* knows that  $\phi$  is the set  $\{w \in W: \text{for all } \psi \text{ such that } w \in \psi, B_{a,w}(\psi) \subseteq \phi\}$ .<sup>4</sup>

Here,  $B_{a,w}(\psi)$  denotes the belief state of agent *a* in world *w* conditional on  $\psi$ , or more precisely, the posterior belief state of *a* that would be induced by learning  $\psi$  in *w*. In Stalnaker’s paper, a belief state is simply the strongest proposition believed to be true, i.e., the set of worlds that the subject believes might be the actual world. It is important that the propositions  $\psi$  on which the belief state is to be conditioned in order to determine whether agent *a*’s belief that  $\phi$  is stable are propositions that are true at *w*.

This stability analysis of knowledge is a simplified variant of the defeasibility analyses of knowledge prevalent in the 1960s and 1970s: where the former refers to a loss of belief, the latter refer to a loss of justification.<sup>5</sup> It is easier to give a semantic model of the loss of belief than to give one of the loss of justification. The stability analysis was later entertained and discussed by Rott,<sup>6</sup> Stalnaker<sup>7</sup> and Baltag and Smets,<sup>8</sup> but none of these authors has actually embraced it as a successful analysis of knowledge. Baltag and Smets occasionally use the term “Stalnaker

<sup>3</sup> Philippe Lamarre and Yoav Shoham, “Knowledge, Certainty, Belief, and Conditionalization,” in *Principles of Knowledge Representation and Reasoning (KR’94)*, eds. Jon Doyle, Erik Sandewall, and Pietro Torasso (San Francisco, CA: Morgan Kaufmann, 1994), 415–424, here 418.

<sup>4</sup> Robert Stalnaker, “Knowledge, Belief and Counterfactual Reasoning in Games,” *Economics and Philosophy* 12, 2 (1996): 133–163, here 146 and 155–156, notation adapted.

<sup>5</sup> The defeasibility analysis of knowledge is linked to philosophers like Annis, Harman, Klein, Lehrer, Paxson, Sosa and Swain. It has been criticised many times, but for some epistemologists, it still remains the most plausible approach to solving the Gettier problem; see Claudio de Almeida and João R. Fett, “Defeasibility and Gettierization: A Reminder,” *Australasian Journal of Philosophy* 94, 1 (2016): 152–169.

<sup>6</sup> “A belief  $\alpha$  is a piece of knowledge of the subject *S* iff  $\alpha$  is not given up by *S* on the basis of any true information that *S* may receive” (Hans Rott, “Stability, Strength and Sensitivity: Converting Belief into Knowledge,” *Erkenntnis* 61, 2–3 (2004): 469–493, here 471).

<sup>7</sup> “[...] define knowledge as belief (or justified belief) that is stable under any potential revision by a piece of information that is in fact true” (Robert Stalnaker, “On Logics of Knowledge and Belief,” *Philosophical Studies* 128, 1 (2006): 169–199, here 187).

<sup>8</sup> What Alexandru Baltag and Sonja Smets, “A Qualitative Theory of Dynamic Interactive Belief Revision,” in *Logic and the Foundations of Game and Decision Theory (LOFT 7)*, eds. Giacomo Bonanno, Wiebe van der Hoek, and Michael Wooldridge (Amsterdam: Amsterdam University Press, 2008), 11–58, call “Stalnaker knowledge” is “belief that is *persistent under revision with any true information*” (13).

knowledge” (in scare quotes), but in general prefer calling what is defined by the stability analysis “safe belief.”<sup>9</sup> Independently of each other, Rott and Stalnaker offered counterexamples against the stability analysis.<sup>10</sup>

Like defeasibility analyses, stability analyses have a problem with misleading evidence (or “misleading defeaters”). The first version of the story above is similar to the counterexamples advanced earlier against defeasibility and stability analyses, even though it would seem odd to call the information provided by Mia misleading. After his observation in the bright sunshine, Sam knew that Ann was passing by his window. Upon receiving the true, but belief-contravening information that it wasn’t Ann and Ben who were passing by, however, Sam drops not only his false belief that he saw Ben, but also his true belief that he saw Ann. If this interpretation of the situation is correct, then knowledge need not be stable in the sense of the stability theory of knowledge.

### 3. The Stability Theory of Rational Belief

We now turn to the question whether stability is a necessary requirement for rational belief.<sup>11</sup> The claim that belief needs to be stable is surprising, because intuitively, and also according to the Platonic Socrates, stability or strength may just be features that *distinguish* knowledge from belief.<sup>12</sup>

---

<sup>9</sup> See Baltag and Smets, “Qualitative Theory,” 13 and 27–29. They think that the stability condition is *too weak* for knowledge, and suggest that knowledge requires stability even upon receipt of arbitrary, possibly false information.

<sup>10</sup> Rott, “Stability, Strength and Sensitivity,” 482–483, and Stalnaker, “On Logics of Knowledge and Belief,” 190. The stability analysis had not been criticised either by Lamarre and Shoham, “Knowledge, Certainty, Belief, and Conditionalization,” or by Stalnaker, “Knowledge, Belief and Counterfactual Reasoning in Games.” Rott’s and Stalnaker’s examples are intended to show that the stability condition is *too strong*. Rott (Stability, Strength and Sensitivity,” 476–477) points to a general reason for the failure of the stability analysis. He shows that a belief is stable (in the above sense) just in case it is more entrenched in the subject’s belief state than *every* false belief. This is a requirement that seems very hard to meet: we probably have many false beliefs, some of them highly entrenched in our cognitive states. So meeting this requirement can hardly be a necessary condition for knowledge.

<sup>11</sup> The first authors to make the connection between the stability theories of knowledge and rational belief were Eric Raidl and Niels Skovgaard-Olsen, “Bridging Ranking Theory and the Stability Theory of Belief,” *Journal of Philosophical Logic* 46, 6 (2017): 577–609.

<sup>12</sup> See Terry Penner, “Socrates on the Strength of Knowledge: Protagoras 351B–357E,” *Archiv für Geschichte der Philosophie* 79, 2 (1997): 117–149, here 121: “Knowledge is strong while belief is *weak*.” Also compare John Hawthorne, Daniel Rothschild and Levi Spectre, “Belief is Weak,” *Philosophical Studies* 173, 5 (2016): 1393–1404, who argue that our everyday notion of belief is unambiguously a weak one.



According to Louis Loeb, however, David Hume held that the most essential elements of belief are steadiness and stability.<sup>13</sup> Every belief, qua belief, is *steady* or “infix” by a belief-forming mechanism (the senses, memory, causal inference, custom or repetition), and it may as such be called *stable in a wider sense*. But not every belief is *stable in the narrower, proper sense* of the term: not every belief is steady in its influence on thought, feeling, the will and action. Steadiness makes for justification other things being equal, but only stability proper makes for justification all things considered.<sup>14</sup> According to Loeb’s “more demanding” reading of the Hume’s stability theory, rational (justified) beliefs have to be stable under *full* or *intense* reflection. Such reflection includes an assessment of the quality of one’s belief-forming processes, as well as the elimination of incoherences among the beliefs that were infixed by the belief-forming mechanisms. Let us call this a *reflective* conception of stability.

I will not contest Loeb’s account, neither as an interpretation of the historical Hume nor as a substantive analysis of belief. Instead I want to turn to a recent alternative approach championed by Hannes Leitgeb.<sup>15</sup> He assumes that the doxastic state of a subject includes both her categorical beliefs, represented by a single proposition, and her degrees of beliefs, represented by a probability function. He picks up on Loeb’s interpretation of Hume, but his motivation can be traced even further back than to Hume. Leitgeb’s initial project was to reconcile two things: (i) the so-called *Lockean thesis*, according to which rational belief *simpliciter* is tied to high probability above a certain threshold value  $r$ , and (ii) the logical closure and consistency of rational categorical beliefs.<sup>16</sup> The lesson from the lottery paradox seems to be that this is an infeasible project. But Leitgeb

---

<sup>13</sup> Louis E. Loeb, *Stability and Justification in Hume’s Treatise* (Oxford: Oxford University Press, 2002) and Louis E. Loeb, *Reflection and the Stability of Belief: Essays on Descartes, Hume, and Reid* (Oxford: Oxford University Press, 2010).

<sup>14</sup> For this and the following, see Loeb, *Stability and Justification in Hume’s Treatise*, chapter 3, and Loeb, *Reflection and the Stability of Belief*, 16–21 and Chapter 5. “A belief might fail to be steady in its influence owing to the presence of beliefs with which it conflicts, beliefs which [...] reduce its influence on the will and action. [...] I use the term ‘stable’ as a shorthand for ‘steady in its influence on thought, passions, and action’” (Loeb, *Stability and Justification in Hume’s Treatise*, 80, and Loeb, *Stability and Justification in Hume’s Treatise*, 155–156).

<sup>15</sup> Hannes Leitgeb, “The Humean Thesis on Belief,” *Proceedings of the Aristotelian Society, Supplementary Volume* 89, 1 (2015): 143–185, and Hannes Leitgeb, *The Stability of Belief: How Rational Belief Coheres with Probability* (Oxford: Oxford University Press, 2017).

<sup>16</sup> Hannes Leitgeb, “The Stability Theory of Belief,” *Philosophical Review* 123, 2 (2014): 131–171. The label “Lockean Thesis” is due to Richard Foley, “The Epistemology of Belief and the Epistemology of Degrees of Belief,” *American Philosophical Quarterly* 29, 2 (1992): 111–124.

demonstrated that such a reconciliation is non-trivially<sup>17</sup> possible, provided that the subject's personal probability function is such that there is a proposition the probability of which does not sink below 0.5, *conditional on any information compatible with the subject's beliefs*. Leitgeb thus modifies the idea that Loeb finds in Hume, and requires stability not under reflection, but under (potential or actual) revision by new information. More precisely, he considers updates of the subject's actual beliefs by new information that is compatible with these beliefs. Here is what Leitgeb calls the *Humean thesis on rational belief*:

It is rational to believe a proposition just in case it is rational to assign a *stably* high subjective probability to it (or to have a *stably* high degree of belief in it).<sup>18</sup>

*The Humean Thesis Explicated:* If *Bel* is a perfectly rational agent's class of believed propositions at a time, and if *P* is the same agent's subjective probability measure at the same time, then for all  $\phi$ :

$\phi$  is in *Bel* if and only if for all  $\psi$ , if  $\psi$  is possible both in the all-or-nothing sense (i.e.,  $\psi$  is logically compatible with *Bel*) and the probabilistic sense (i.e.,  $\psi$  has non-zero probability), then  $P(\phi \mid \psi) > r$ .<sup>19</sup>

Here  $P(\phi \mid \psi)$  is the standard conditional probability of  $\phi$  given  $\psi$ , defined as  $P(\phi \cap \psi) / P(\psi)$ , and  $r$  is a threshold parameter lying between 0.5 and 1. Conditionalising one's probability function *P* on a proposition  $\psi$  essentially means accepting  $\psi$  either actually or hypothetically. According to the Humean thesis, it is rational to believe a proposition  $\phi$  just in case its probability remains high conditional on any proposition  $\psi$  that is doxastically possible for the agent: no such proposition  $\psi$  defeats the high degree of belief in  $\phi$ .<sup>20</sup> The idea here is similar to

---

<sup>17</sup> 'Non-trivial' here means that there are beliefs with a probability below 1. This is equivalent to there being non-tautological beliefs, if the probability function is supposed to be regular. I assume that rational agents in general aim at having non-trivial belief sets.

<sup>18</sup> Leitgeb, "The Humean Thesis on Belief," 152.

<sup>19</sup> Leitgeb, "The Humean Thesis on Belief," 163, notation adapted and some more technical clauses replaced by ordinary-language formulations. On 159–162, Leitgeb reviews five alternative ways of making the generic idea of the Humean thesis precise. His option (b) which "would correspond to a kind of coherence theory of belief" (160) is closer to (Loeb's interpretation of) Hume than option (d) which Leitgeb ultimately embraces.

<sup>20</sup> Leitgeb's move of adopting the Humean rather than the Lockean thesis, i.e., of requiring *r*-stability rather than *P*-stability (which has the constant 0.5 in place of the parameter *r*), can be interpreted as reflecting the idea that the threshold value for the conditional probabilities should be the same as for the unconditional probability, i.e., it should be *r* rather than 0.5. I find this the most natural interpretation, but Leitgeb (personal communication) is ready to apply different thresholds to conditional and unconditional beliefs. For the ranges of Lockean and Humean thresholds that are suitable for a given proposition, see Hans Rott, "Stability and Skepticism in

that of the stability theory of knowledge, with the crucial difference that the latter refers to the (hypothetical or actual) acceptance of *true* propositions while the former refers to the (hypothetical or actual) acceptance of propositions *compatible with the subject's beliefs*.

The second version of our story shows, I submit, that the stability account based on the Humean thesis does not adequately capture the intuitive notion of rational belief. Sam was fully rational in believing that Ann and Ben were passing by when he looked out of his window (independently of whether it actually was Ann and Ben who he saw). The information that Ann and Ben have had an important obligation to present their joint paper at the workshop is consistent with Sam's belief that he saw the couple walking past his window, and indeed with his full body of belief. Sam knew, after all, that their scheduled presentation might have been put off. But the news about their commitment dramatically decreases the likelihood that it was Ann and Ben who he saw. So we have found a perfectly rational belief that has a rather low subjective probability when conditionalised on information compatible with Sam's full body of beliefs. This is a counterexample to the Humean thesis.

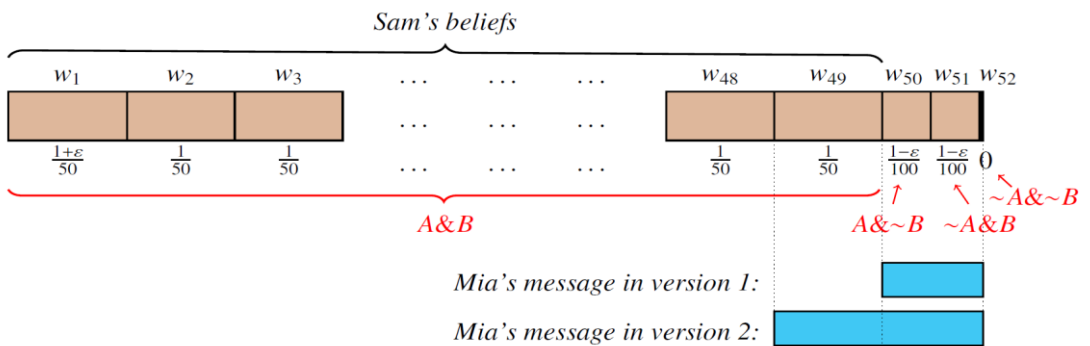


Fig. 1: Sam's doxastic state yesterday and Mia's messages

This example is meant to be compelling because it is intuitively plausible. Still, it will be reinforced and provide a better service as a counterexample if we can reproduce it in terms of the formal model used by Leitgeb. Suppose that Sam's doxastic situation yesterday looked as follows (see Fig. 1). He considered 52 worlds

the Modelling of Doxastic States: Probabilities and Plain Beliefs," *Minds and Machines* 27, 1 (2017): 167–197.

as possible that we call  $w_1, w_2, \dots, w_{52}$ . In  $w_1, \dots, w_{50}$ , Ann passes Sam's window, in  $w_{51}$  and  $w_{52}$ , she doesn't. In  $w_1, \dots, w_{49}$  and  $w_{51}$ , Ben passes Sam's window, in  $w_{50}$  and  $w_{52}$ , he doesn't. In  $w_{49}, \dots, w_{52}$ , Ann and Ben have an important obligation elsewhere (such that at least one of them ought to be present), in  $w_1, \dots, w_{48}$  they don't. Suppose further that Sam assigned the following subjective probabilities to the worlds he considered possible: for some very small positive real number  $\varepsilon$ ,  $P(w_1) = \frac{1+\varepsilon}{50}$ ,  $P(w_2) = \dots = P(w_{49}) = \frac{1}{50}$ ,  $P(w_{50}) = P(w_{51}) = \frac{1-\varepsilon}{100}$  and  $P(w_{52}) = 0$ . Assuming that Sam can think of 49 different ways his neighbours' passing by might have come about, this is a natural representation of Sam's belief state yesterday.<sup>21</sup>

The proposition  $\{w_1, \dots, w_{49}\}$  is the only non-trivial  $P$ -stable set and may thus be taken to qualify as the proposition characterising Sam's initial beliefs. Sam's conditional probability that Ann and Ben passed by, *given* the information that they had an important obligation elsewhere, is above 0.5, but only slightly so. If we take  $\varepsilon = 0.1$ , for instance, it is 0.526. As long as Leitgeb assumes that the threshold for belief is set to 0.5, he can still recommend as rational the belief that Ann and Ben passed by, since  $\{w_{49}\}$  is the only non-trivial  $P$ -stable set of Sam's subjective probabilities conditionalised on the information that Ann and Ben had an important obligation (i.e., conditionalised on the proposition  $\{w_{49}, w_{50}, w_{51}, w_{52}\}$ ). But a posterior probability of 0.526 is low, arguably too low to support belief *simpliciter*. Belief appears to require at least a moderately *high* probability, one that lies significantly above 0.5. As a consequence, Sam loses his belief that Ann and Ben were passing by yesterday.

The following *Preservation condition* may be viewed as a qualitative analogue of Leitgeb's stability condition: If a proposition is consistent with a subject's current beliefs, she should give up none of her current beliefs on accepting or on hypothetically assuming that this proposition is true. Preservation is one of the basic conditions of Alchourrón, Gärdenfors and Makinson, and it has almost universally been accepted in the belief revision literature.<sup>22</sup> But the second

---

<sup>21</sup> There are of course many alternative representations of Sam's belief state that make sense. But it is important to stress here that a single natural way of formally fleshing out the informal example is sufficient for establishing that it can serve as a serious counterexample to Leitgeb's theory. And I claim that my formal precisification is a natural one. Two potential objections do not strike me as compelling. First, there is no reason to suppose that Mia's message introduces a context change that forces a refinement of the partition of all possibilities. Second, nothing depends on there being a world with zero probability; the example could easily be modified in such a way that  $w_{52}$  has positive probability.

<sup>22</sup> Among the very few authors arguing against Preservation are Charles B. Cross, "Belief Revision, Nonmonotonic Reasoning, and the Ramsey Test," in *Knowledge Representation and Defeasible Reasoning*, eds. Henry E. Kyburg, Ronald P. Loui, and Greg N. Carlson (Boston:

version of our example may also serve as a counterexample to Preservation. Proceeding on the supposition that categorical beliefs derive from probabilities, we have just fleshed out the example in such a way that Sam appears to be fully rational in dropping his belief that it was Ann and Ben who he saw yesterday.<sup>23</sup>

#### 4. Analysis: What ‘Stability’ May Mean

The theories reviewed make use of different ideas of stability that are specialisations of a more general concept. The general stability scheme is this. A state property  $X$  is stable under the state transformation  $Y$  just in case the following holds: for all states  $S$ , if  $S$  has the property  $X$  and undergoes a transformation (of the kind)  $Y$  and no other transformation is performed on  $S$ , then the state  $S' = Y(S)$  has the property  $X$ , too.

The states  $S^a$  we want to consider in the following are mental states of a rational agent  $a$ . For any proposition  $\phi$ , let the property  $X^\phi$  of a state  $S^a$  be that in  $S^a$ ,  $a$  has a certain propositional attitude of the epistemic or doxastic kind with respect to  $\phi$ . That the state  $S^a$  has the property  $X^\phi$  means that agent  $a$  *Xes* that  $\phi$  in  $S^a$ , where ‘*Xes*’ stands for verbs such as ‘knows,’ ‘believes,’ ‘rationally believes,’ ‘expects,’ ‘surmises,’ ‘doubts,’ ‘wonders,’ ‘is certain,’ ‘is convinced,’ ‘assigns a high subjective probability,’ ‘entertains (the idea),’ etc.

The property  $X^\phi$  of  $S^a$  is called *stable under reflection* just in case the following holds: if  $a$  *Xes* that  $\phi$  in state  $S^a$  and then reflects about the system of propositions *Xed* by herself (and nothing else happens), then  $a$  still *Xes* that  $\phi$  after having finished her reflections. The property  $X^\phi$  is called *stable under updating (by eligible information)* just in case the following holds: if  $a$  *Xes* that  $\phi$  in  $S^a$  and then accepts an eligible piece of information  $\psi$  (and nothing else happens), then  $a$  still *Xes* that  $\phi$  in the updated state  $S^a * \psi$ .

---

Kluwer, 1990), 223–244, here 232–234; Włodzimierz Rabinowicz, “Stable Revision, or is Preservation Worth Preserving?” in *Logic, Action, and Information: Essays on Logic in Philosophy and Artificial Intelligence*, eds. André Fuhrmann and Hans Rott (Berlin: de Gruyter, 1996), 101–128, here 105–106; and Richard Bradley, “Restricting Preservation: A Response to Hill,” *Mind* 121, 481 (2012): 147–159, here 155–156.

<sup>23</sup> Hanti Lin and Kevin T. Kelly, “Propositional Reasoning that Tracks Probabilistic Reasoning,” *Journal of Philosophical Logic* 41, 6 (2012): 957–981, here 964, call Preservation ‘Accretion’ and give a Gettier-style example that on the face of it resembles the probabilified second version of our story. However, I find their example unconvincing since they give no argument for their claim that “the strongest proposition we accept is the disjunction of ‘Nogot’ with ‘Havit,’ namely ‘somebody.’” Their example is also criticised by Leitgeb, *The Stability of Belief*, 187.

According to Loeb,<sup>24</sup> to whom Leitgeb makes essential reference, stability under reflection is what Hume was after. Stability under updating covers the stability theories of knowledge and rational belief introduced above if we specialise ‘ $X$ ’ to ‘believes’ and to ‘assigns a probability above  $x$ ,’ respectively. For the definition of stability under updating, we still need to specify when to regard a piece of information as eligible. ‘Eligible’ is used as a generic term here that is supposed to cover different interpretations of stability. We focus on the two interpretations that shape the stability theories of knowledge and rational belief sketched above and call a proposition  $\psi$  (i) *eligible for knowledge* iff  $\psi$  is true; and (ii) *eligible for belief* iff  $\psi$  is compatible with the subject’s current beliefs (i.e., iff  $\psi$  is not belief-contravening).

## 5. No Direct Route from the Instability of Knowledge to the Instability of (Rational) Belief

Do we really need two versions of our example? It is part of almost all contemporary epistemology that knowledge is a kind of rational belief. Though this is an assumption that clearly does *not* follow from the two stability theories, let us suppose it is true for the purposes of the following considerations. On this hypothesis, the fact that knowledge need not be stable seems to entail straightaway that rational belief need not be stable either. It looks as if this can be established simply by reasoning by way of a Bocardo inference:

Some pieces of knowledge are unstable.	(major premise)
All pieces of knowledge are rational beliefs.	(minor premise)
Some rational beliefs are unstable.	(conclusion)

The Bocardo scheme has been recognised as valid ever since Aristotle’s syllogistics. But this particular inference is fallacious for two reasons. First, ‘stability’ is a syncategorematic predicate that may mean different things when applied to knowledge and when applied to belief. This is indeed the case with the stability theories of knowledge and belief: they involve different propositional attitudes and different notions of eligibility. Although both theories employ the notion of stability under updating, what makes a piece of information eligible is truth in the case of knowledge and compatibility with the subject’s beliefs in the case of belief.

The ambiguity of the stability predicate is not deeply hidden, but it is worth emphasising, and it indeed prevents version 1 of our story from being suitable as a

---

<sup>24</sup> Loeb, *Stability and Justification in Hume’s Treatise*, chapter 3.

counterexample to the stability theory of rational belief. But can't we perhaps find a more sophisticated concept of eligibility that is suitable for both knowledge and belief? I do not want to exclude this possibility. But even if the search for such a universally applicable notion of eligibility were successful, the inference above would still fail to go through. As we have seen, 'stable' is not a primitive predicate, but has an intrinsically conditional structure where the propositional attitude involved occurs both in the antecedent and the consequent of the relevant conditional. Consequently, 'unstable' has a conjunctive structure in which the propositional attitude involved occurs twice, once positively and once negatively. If we make the logical structures explicit, we realise that it is inadequate to represent the proposed inference as a Bocado like this:

$$\frac{\begin{array}{l} \exists \phi (knows(\phi) \ \& \ unstable(\phi)) \\ \forall \phi (knows(\phi) \supset \ rbelieves(\phi)) \end{array}}{\exists \phi (rbelieves(\phi) \ \& \ unstable(\phi))}$$

In its deeper structure, the inference above instantiates a scheme that is indeed logically invalid—even if we could avail ourselves of a notion of eligibility that is suitable for both knowledge and belief:

$$\frac{\begin{array}{l} \exists \phi, \psi, S (knows(\phi, S) \ \& \ eligible(\psi, S) \ \& \ \sim knows(\phi, S * \psi)) \\ \forall \phi, S (knows(\phi, S) \supset \ rbelieves(\phi, S)) \end{array}}{\exists \phi, \psi, S (rbelieves(\phi, S) \ \& \ eligible(\psi, S) \ \& \ \sim rbelieves(\phi, S * \psi))}$$

Back to our example. The first version does not show that rational belief is unstable. Sam, I claim, initially *knew* and thus *rationaly believed* that Ann was passing by. He does not believe that she was passing by any more after having received the true information that it wasn't Ann and Ben who were passing by. But the information he received from Mia was incompatible with his beliefs. So while it was eligible for knowledge, it wasn't eligible for rational belief.

The second version of the example, in contrast, does illustrate the instability of rational belief. Here the information provided by Mia is eligible for both knowledge and belief. We could actually have used this version as a counterexample to the stability theory of knowledge. However, since it is a lot more complicated than the first version (witness Fig. 1), the latter is of independent value in making a simple non-probabilistic case against the stability of knowledge.

## 6. The Stability of Knowledge and Belief Themselves

The stability theories outlined above define *knowledge* as stable true belief and *rational belief* as stably high probability. By so doing they do not immediately answer the question whether knowledge and rational belief *themselves* are stable, that is, stable under updating by propositions that are eligible in the suitable sense. In this final section, I identify some sufficient conditions for this being true, on the basis of the theories in question.

We begin with knowledge, as conceived by the qualitative stability theory. That agent *a* knows that  $\phi$  in state *S*, in symbols  $knows_a(\phi, S)$ , has been defined by

$believes_a(\phi, S)$  and for all  $\psi$ , if  $true(\psi)$ , then  $believes_a(\phi, S*\psi)$ .

We want to show that knowledge is stable under eligible updating, that is:

If  $knows_a(\phi, S)$  and  $true(\psi)$ , then  $knows_a(\phi, S*\psi)$ .

So suppose that  $knows_a(\phi, S)$  and  $true(\psi)$ . We need to show that, first, that  $believes_a(\phi, S*\psi)$  and, second, that for all  $\chi$ , if  $true(\chi)$ , then  $believes_a(\phi, S*\psi*\chi)$ . Now it follows from the definition of  $knows_a(\phi, S)$  that  $believes_a(\phi, S*\psi)$ , which gives us the first claim.

It seems that the only way to prove the second claim is to take an arbitrary true sentence  $\chi$  and show that the state  $S*\psi*\chi$  supports all beliefs supported by  $S*(\psi\&\chi)$  and that  $\psi\&\chi$  is eligible, i.e., true. Since both  $\psi$  and  $\chi$  are true, so is  $\psi\&\chi$ . That the beliefs supported by  $S*(\psi\&\chi)$  are included in the beliefs supported by  $S*\psi*\chi$  is a condition well-known in the theory of iterated belief revision. It is satisfied, among others, by irrevocable revision (also known as radical revision) and by lexicographic revision (also known as moderate revision); but it is not satisfied, for instance, by natural revision (also known as conservative revision) and restrained revision.<sup>25</sup> So knowledge in the sense defined by the stability theory of knowledge is stable if either irrevocable or lexicographic belief revision is employed, but knowledge need not be stable if any other method of iterated revision is employed.

Let us now look at rational belief, as conceived by the probabilistic stability theory. That agent *a* in state *S* rationally believes that  $\phi$ , in symbols  $rbelieves_a(\phi, S)$ , has been defined by

$hiprob_a(\phi, S)$  and for all  $\psi$ , if  $compatibles(\psi)$ , then  $hiprob_a(\phi, S*\psi)$ .

---

<sup>25</sup> For the four methods, compare Hans Rott, "Preservation and Postulation: Lessons from the New Debate on the Ramsey Test," *Mind* 126, 502 (2017): 609–626. Notice that since both  $\psi$  and  $\chi$  are true, they are compatible with each other. Notice also that we need to take belief-contravening revisions into account here, too. It is not guaranteed that  $\chi$  is consistent with  $S*\psi$ .



We want to show that rational belief is stable under eligible updating, that is:

If  $rbelieves_a(\phi, S)$  and  $compatible_S(\psi)$ , then  $rbelieves_a(\phi, S*\psi)$ .

So suppose that  $rbelieves_a(\phi, S)$  and  $compatible_S(\psi)$ . We need to show that, first,  $hiprob_a(\phi, S*\psi)$  and, second, that for all  $\chi$ , if  $compatible_{S*\psi}(\chi)$  then  $hiprob_a(\phi, S*\psi*\chi)$ . It follows from the definition of  $rbelieves_a(\phi, S)$  that  $hiprob_a(\phi, S*\psi)$ , which gives us the first claim.

The only way to prove the second claim seems to take an arbitrary sentence  $\chi$  that is compatible with  $S*\psi$  and show that the state  $S*\psi*\chi$  assigns a high probability to all propositions that are highly probable in state  $S*(\psi\&\chi)$  and that  $\psi\&\chi$  is eligible, i.e., compatible with  $S$ . Since both  $\psi$  is compatible with  $S$  and  $\chi$  is compatible with  $S*\psi$ , it is plausible to assume that  $\psi\&\chi$  is indeed compatible with  $S$ . Thus, by the definition of  $rbelieves_a(\phi, S)$ , we get  $hiprob_a(\phi, S*(\psi\&\chi))$ . If the probabilities assigned in doxastic states are changed by ordinary Bayesian conditionalization when the input or assumptions are consistent with those states, then changing a state first by compatible  $\psi$  and then by compatible  $\chi$  yields identical probabilities to changing the state only once by compatible  $\psi\&\chi$ . This gives us  $hiprob_a(\phi, S*\psi*\chi)$ , as desired. Thus on the assumptions made, rational belief is indeed stable. Other ways of changing probabilities by compatible input or assumptions may give different results.

## 7. Conclusion

I have presented a stability theory of knowledge (discussed by Stalnaker, Lamarre and Shoham, Rott, and Baltag and Smets) and a stability theory of rational belief (embraced by Leitgeb), which have not been compared in the literature before. It was shown that these theories make use of a general concept of stability which can be differentiated into two distinct species. Using two versions of a concrete example, I argued that (i) knowledge need not be stable, and that (ii) rational belief need not be stable either, in the senses intended by the two theories. The two claims are independent of each other. Even on the supposition that knowledge is a particular kind of rational belief, the existence of unstable knowledge does not entail the existence of unstable rational belief, due to the logical structure of the general stability scheme and an ambiguity in the meaning of the predicate “stable.”<sup>26</sup>

---

<sup>26</sup> Acknowledgements. I'd like to thank Tim Kraft, Hannes Leitgeb, Eric Raidl, Niels Skovgaard-Olsen and audiences in Regensburg, Stockholm, Munich, Paris, Maastricht and Dortmund for instructive comments on earlier versions of this paper. I am also grateful to the Swedish Collegium for Advanced Study in Uppsala for providing me with excellent research conditions while part of this paper was written.



# THE AVAILABILITY HEURISTIC AND INFERENCE TO THE BEST EXPLANATION

Michael J. SHAFFER

**ABSTRACT:** This paper shows how the availability heuristic can be used to justify inference to the best explanation in such a way that van Fraassen's infamous "best of a bad lot" objection can be adroitly avoided. With this end in mind, a dynamic and contextual version of the erotetic model of explanation sufficient to ground this response is presented and defended.

**KEYWORDS:** inference to the best explanation, explanation, scientific progress, heuristics, erotetic logic, contextualism

## 1. Introduction

The programs respectively associated with bounded and ecological rationality (BER) and the heuristics and biases program (HBP) have been regarded as having significant implications for many areas of philosophy and psychology. The HBP is an empirically motivated project that focuses on demonstrating why human cognitive performance with respect to tasks like probabilistic reasoning and decision-making often violates (or appears to violate) classical norms of rationality.<sup>1</sup> On a more positive note, those working in the context of this program have argued that human cognitive performance involves using variety of simple heuristics rather than conformity to the classical norms of rationality (i.e. the probability calculus, classical first-order logic, orthodox decision theory, etc.). The BER project is also an empirically minded project aimed at showing that human cognitive performance is actually rational despite the fact that such behavior often does not satisfy classical standards of rationality. BER specifically focuses on the

---

<sup>1</sup> See Daniel Kahneman, Paul Slovic and Amos Tversky, *Judgment under Uncertainty* (Cambridge: Cambridge University Press, 1982) and Gerd Gigerenzer, *The Adaptive Tool Box* (Oxford: Oxford University Press, 2000). Also, see Ken Manktelow, *Thinking and Reasoning* (New York: Psychology Press, 2012) for an excellent overview and Johnathan Howard, *Cognitive Errors and Diagnostic Mistakes* (New York: Springer, 2019) for discussion of heuristics and cognitive biases in medicine.

computational and environmental features of real cognitive performance as the key to understanding how humans are rational in terms of this alternative, heuristic-based and environmentally sensitive, account of rationality.

BER is a reaction to the pessimistic interpretation of the results of the HBP which were sometimes alleged to show that humans are badly irrational when judged against classical norms of rationality.<sup>2</sup> The defenders of the BER project effectively disputed this more pessimistic conclusion and argued that facts about human cognitive performance are better understood as evidence that the traditional norms of rationality are not the correct norms by which human cognitive performance should be judged. The opposition between these two camps is ongoing and it has led to some heated exchanges.<sup>3</sup> But, these ideas can be usefully combined to support an alternative and empirically grounded conception of rationality as adherence to heuristic rules that are normatively appropriate in certain ecological contexts and given human cognitive limitations.<sup>4</sup>

In this paper this sort of empirically based and fallibilistic approach to rationality is used to justify inference to the best explanation (IBE) and this justification is specifically based on the availability heuristic. This strategy also involves the central contention that IBE involves the more general notion of problem or question substitution.<sup>5</sup> In its relevant form, the availability heuristic is the claim that certain inferences and decisions are made on the basis of psychologically familiar factors, as opposed to all relevant factors.<sup>6</sup> Problem or question substitution is just the tactic of substituting and solving an easier version of a problem when a given problem is itself too difficult to solve. So, the availability heuristic is just a special case of problem substitution.<sup>7</sup> The contention here then is that it is rational to accept the best psychologically available

---

<sup>2</sup> See Richard Nisbett and Eugene Borgida, "Attribution and the Psychology of Prediction," *Journal of Personality and Social Psychology* 32 (1975): 932-43 and Massimo Piatelli-Palmarini, *Inevitable Illusions* (New York: John Wiley, 1994).

<sup>3</sup> See Richard Samuels, Stephen Stich, and Michael Bishop, "Ending the Rationality Wars: How to Make Disputes about Human Rationality Disappear," in *Common Sense, Reasoning and Rationality*, ed. Renee Elio (Oxford: Oxford University Press, 2002), 236-268, Daniel Kahneman and Amos Tversky, "On the Reality of Cognitive Illusions: A Reply to Gigerenzer's Critique," *Psychological Review* 103 (1996): 582-591 and Gerd Gigerenzer, "On Narrow Norms and Vague Heuristics," *Psychological Review* 103 (1996): 592-596..

<sup>4</sup> A version of this hybrid view antedates both HBP and BER and was defended in Herbert Simon, *Models of Man* (New York: Wiley, 1957).

<sup>5</sup> See Daniel Kahneman, *Thinking Fast and Slow* (New York: Farrar, Straus and Giroux, 2011).

<sup>6</sup> See Kahneman, *Thinking Fast and Slow*.

<sup>7</sup> See Kahneman, *Thinking Fast and Slow*, ch. 9 for discussion of this connection.

explanation of psychologically available data when we frame this sort of inferential practice in terms of a more naturalistic and realistic conception of rationality. In other words, it is often perfectly rational to substitute simpler explanatory problems for more complex ones. This is due to our cognitive limitations and environmental constraints. Such substitution does carry with it the possibility of cognitive bias and error, but this is no surprise when we recognize that explanatory reasoning involves uncertainty and limited cognitive resources. However, as we shall see, such reasoning also involves the possibility for the correction of such errors and the refinement of our explanatory understanding.

The model proposed here for IBE is founded on a theory that combines insights from epistemic contextualism and the erotetic theory of explanation. One important implication of this work is that it provides an answer to van Fraassen's infamous criticism of IBE.<sup>8</sup> This critical attack on IBE is based on the contention that the conclusions of such inferences should not be taken to be likely (and hence should not be accepted). This is supposed to be because such inferences are always based on a set of available hypotheses that constitutes only a small sub-set of all of the possible hypotheses that are potential explanations of a given phenomenon. So, as van Fraassen has argued, it appears to be the case that it will always be much more likely that the true explanation is among the set of unconsidered (and mostly unformulated) hypotheses. The alternative model of IBE presented in this paper neatly avoids this criticism and renders rational the acceptance of the conclusions of such inferences. In part this is because the model of IBE introduced here is both dynamic and contextual thus providing for the possibility of error correction and it is based on the insight that contextual factors fix the sets of hypotheses and evidence that are appealed to in such inferences.<sup>9</sup>

---

<sup>8</sup> See Bas van Fraassen, *Laws and Symmetry* (Oxford: Clarendon, 1989).

<sup>9</sup> The theory developed here has much in common with Hintikka's view of abduction as the search for correct explanations (i.e. as abductive search) as presented in Jaakko Hintikka, "What is Abduction? The Fundamental Problem of Contemporary Epistemology," *Transactions of the Charles S. Peirce Society* 34 (1998): 503-533. He concludes that abduction is not a form of inference at all. The view defended here is that IBE is the terminal step in abductive search and that IBE is indeed a form of inference involved in that process. But, abductive search also involves seeking evidence and constructing sets of theories that are used as inputs in IBEs. In other words, abductive search includes the construction of the sample space of theories and the marshalling of relevant evidence, which are then employed in IBE inferences. This aligns with much of Jonah Schupbach's criticism of van Fraassen's objection to IBE from "Is the Bad Lot Objection Just Misguided?" *Erkenntnis* 79 (2014): 55-64. Schupbach argues that van Fraassen's criticism of IBE is misguided in that it confuses the issue of the probity of IBE inferences with the matter of the completeness and appropriateness of the input into IBE inferences. See Kyle Stanford, *Exceeding our Grasp* (Oxford: Oxford University Press, 2010) and Finnur Dellsén,

## 2. Constructing a theory of IBE

IBE is perhaps the most basic form of reasoning that humans engage in. Perhaps more crucially, IBE plays a central role in scientific inquiry. For example, McMullin and Lipton contend that it is *the* central form of inference in science.<sup>10</sup> But, there has been much critical discussion of this sort of explanatory reasoning and considerations of the probity of explanatory reasoning as a distinct form of inference are most notably traceable back to Peirce's work on abduction.<sup>11</sup> On this basis, it should be clear that any suitable account of IBE must satisfy (at least) three important desiderata. First, the account must incorporate a plausible theory of explanation. It is straightforwardly obvious that we must know what an explanation simpliciter is if we are to hope to come to know what the best explanation of anything is. Second, the account must provide an explication of what it is for one explanation to be better than another explanation. Finally, the probative nature of this form of inference must be accounted for. This last aspect of any adequate account of IBE is especially important, as IBE arguments must provide warrant for their conclusions in such a way that we are entitled to provisionally accept such theoretical claims.<sup>12</sup> If this final desideratum is not satisfied, then it is obvious that IBE would be of no use in solving the problem of the acceptance of theoretical claims in a substantial and normative sense.

### 2.1 The Questions of Explanation

The 20<sup>th</sup> century history of the philosophy of science is replete with examples of attempts to provide adequate theories of explanation, and this fact is well-represented and summarized in Salmon's classic 1989 survey.<sup>13</sup> The most well-

---

"Reactionary Responses to the Bad Lot Objection," *Studies in History and Philosophy of Science Part A* 61 (2017): 32-40 on this issue and others related to the bad lot objection.

<sup>10</sup> See Ernan McMullin, *The Inference that Makes Science* (Marquette: Marquette University Press, 1992) and Peter Lipton, *Inference to the Best Explanation*, 2<sup>nd</sup> ed. (London: Routledge, 2004).

<sup>11</sup> See C. S. Peirce, *Collected Papers of Charles Sanders Peirce*, eds. Charles Hartshorne, Paul Weiss, and Arthur Burks, 8 vols. (Cambridge: Harvard University Press, c.1901/1931-1958).

<sup>12</sup> This is the general gist of van Fraassen's *Laws and Symmetry* criticism of IBE. See Samir Okasha, "Van Fraassen's Critique of Inference to the Best Explanation," *Studies in the History and Philosophy of Science* 31 (2000): 691-710, Stathis Psillos, "On Van Fraassen's Critique of Abductive Reasoning," *The Philosophical Quarterly* 46 (1996): 31-47, Stathis Psillos, *Scientific Realism: How Science Tracks the Truth* (London: Routledge Press, 1999), Timothy Day and Harold Kincaid, "Putting Inference to the Best Explanation in Its Place," *Synthese* 98 (1994): 271-295 and Stanford 2010 for extensive discussion of van Fraassen's argument.

<sup>13</sup> See Wesley Salmon, "Four Decades of Scientific Explanation," in Phillip Kitcher and Wesley

known theory of course is the deductive-nomological model of explanation. However, there are numerous well-known counter-examples to this account of explanation, and, for the most part, this theory has been rejected.<sup>14</sup> But, this need not worry us as there is a readily available alternative account of explanation that can be used to ground IBE. This model takes an explanation to be the answer to an explanatory question. As such, the best explanation will turn out to be the best answer to such a question. This account of explanation is promising because it ties explanation directly to understanding without begging any specific questions about what types of explanations are legitimate. In point of fact, it is compatible with the view that different kinds of explanations are perfectly legitimate in different contexts within a particular discipline, or in different disciplines, or at different times, etc. As such, it is perfectly compatible with the idea that methodological standards can vary with context. As we shall see this is a significant virtue of the account of IBE presented here. The modern work on erotetic logic that gave rise to the general idea of an erotetic model of explanation can be traced back to the work of Åqvist via the more or less independent work of Belnap and Steel, Hintikka, and Bromberger.<sup>15</sup> But, the best-known and more contemporary erotetic accounts of explanation are those presented by van Fraassen and Tuomela.<sup>16</sup> However the

---

Salmon (eds.), *Scientific Explanation* (Minneapolis: University of Minnesota Press, 1989), 3-219 and Wesley Salmon, *Scientific Explanation and the Causal Structure of the World* (Princeton: Princeton University Press, 1984).

<sup>14</sup> See Phillip Kitcher and Wesley Salmon, eds., *Scientific Explanation* (Minneapolis: University of Minnesota Press, 1989) and Bas van Fraassen, *The Scientific Image* (Oxford: Clarendon, 1980) for detailed consideration of the problems with the D-N model of explanation. This is not to say, of course, that other accounts of the nature of explanation are not also problematic. For example, as shown in Michael Shaffer, "Unification and the Myth of Purely Reductive Understanding," *Organon F* (forthcoming), the unificationist view of explanation is also afflicted with serious problems related to IBE. The unificationist view is most famously defended in Phillip Kitcher, "Explanatory Unification," *Philosophy of Science* 48 (1981): 507-531, Phillip Kitcher, *The Advancement of Science* (New York: Oxford University Press, 1993) and Michael Friedman, "Explanation and Understanding," *The Journal of Philosophy*, 71 (1974): 5-19.

<sup>15</sup> See Lennart Åqvist, *A New Approach to the Logical Theory of Interrogatives, Part 1: Analysis* (Uppsala: Filosofiska föreningen i Uppsala, 1965), Noel Belnap and Thomas Steel, *The Logic of Questions and Answers*. New Haven: Yale University Press, 1976), Jaakko Hintikka, *The Semantics of Questions and the Questions of Semantics* (Amsterdam: North Holland, 1976), Sylvain Bromberger, *On What we Know we Don't Know* (Chicago: University of Chicago Press, 1992) and Sylvain Bromberger, "Why Questions," in Robert Colodny (ed.) *Mind and Cosmos: Essays in Contemporary Science and Philosophy*, vol. 3 (Pittsburgh: University of Pittsburgh Press, 1966): 75-100.

<sup>16</sup> See van Fraassen, *The Scientific Image* and Raimo Tuomela, "Truth and Best Explanation," *Erkenntnis* 22 (1985): 271-299.

theory of IBE developed here will be more specifically based on Hintikka's account of the logic of questions and answers, though the account presented is ultimately also rather like that proposed by Tuomela.<sup>17</sup> However, before turning to the relevant details of that account it will be instructive to first outline some of the characteristic and general features of the erotetic model of explanation.

## 2.2 The Multiplicity of Explanation and Context

It has been widely acknowledged for quite some time now that a given body of data can be explained by a potentially infinite number of theories. This is just the familiar point about the underdetermination of theory by evidence. However, there is another sort of ambiguity inherent in the activity of explanation that is accentuated in the erotetic model of explanation. This is the following sort of pedagogical phenomenon. Even mild acquaintance with science and how it is generally taught should make us aware of the kind of situation in which an explanation of some phenomenon is presented, where that explanation is later revealed to be incomplete or not quite correct. For example, classical mechanics is generally taught before quantum mechanics or relativistic mechanics, and, typically the latter types of explanation of the very same phenomena are regarded as more complete and more correct. However, in general, this does not impugn the simpler explanation either as worthless or as non-explanatory. Quite the opposite is true in practice. The explanation of many phenomena in terms of classical mechanics is often retained because it is appropriate *in certain contexts*. This issue raises an aspect of explanation that has not received as much attention as it deserves from philosophers of science. This is just the context dependence of explanation.<sup>18</sup> It is however helpful for the purposes of this paper that sensitivity to context dependence has become commonplace in contemporary epistemology, and this provides us with some guidance on the matter.

The sense in which explanation appears to be context dependent is then relevantly similar to the sense in which the terms 'knowledge' and 'justification' have been said to be context dependent in relatively recent discussions in epistemology. Specifically, Keith DeRose and David Lewis have famously defended this sort of view.<sup>19</sup> The basic idea behind the concept of context dependence of

---

<sup>17</sup> Gilbert Harman, "Inference to the Best Explanation," *The Philosophical Review* 74 (1965): 88-95, and Lipton, *Inference to the Best Explanation*, 2<sup>nd</sup> ed.

<sup>18</sup> Ironically, the theory presented in Bas van Fraassen, *The Scientific Image* (Oxford: Clarendon Press, 1980) incorporates the contextual aspects of explanation most straightforwardly.

<sup>19</sup> See Keith DeRose, "Contextualism: An Explanation and Defense," in John Greco and Ernest Sosa (eds.) *The Blackwell Guide to Epistemology* (Malden: Blackwell, 1999), 187-205.



epistemological concepts like knowledge is that assumptions about the epistemic standards involved in a given situation vary from context to context and so our attributions of knowledge may also vary as a result. For example, in everyday discussion skeptical hypotheses are ignored as irrelevant while in the context of a philosophical discussion about the nature of knowledge skeptical hypotheses are taken to be relevant. As such, one may have the knowledge that there is a hand before one's face in the former context, but not in the latter context *without contradiction*. This is supposed to be the case because the standards that govern the philosophical context are much stronger than those that are in place in more ordinary, everyday, contexts. This then is the crux of the contextualist view of knowledge. Whether a particular person knows a particular proposition depends on certain contextual features of the person's epistemic situation.

What will be suggested here is that explanation has a similar sort of context dependence that has gone largely unnoticed by most philosophers of science. For example, what counts as an acceptable explanation of a phenomenon in a high school physics class is different from what counts as an acceptable explanation of that phenomenon in a graduate level physics seminar, and our theory of explanation needs to reflect this fact about scientific practice. The epistemic standards that are presupposed in the latter context are much more stringent than those at work in the former, and that makes an important difference with respect to which theory we ought to accept in a given context. The main feature of the view defended here is that context determines what kinds of explanatory standards are in place in a context, the body of explanatory hypotheses to be considered and the body of evidence to be explained. Different degrees or depths of explanatoriness are then appropriate to different contexts much like different standards of evidence apply in different contexts according to epistemic contextualism about knowledge.

In terms specific to the erotetic model of explanation, this will amount to regarding the best explanation as the best answer to some why-question or how-question given some specified explanatory context. Of course this means that we will have to say something about what contextual factors need to be taken into account in general when assessing what explanation is best in a fully specified explanatory context. However, as epistemic context appears to be highly plastic and variable, it may turn out that there is not very much of interest that we can say about general epistemic standards across contexts. So, one interesting aspect of the

---

Keith DeRose, *The Case for Contextualism: Knowledge, Skepticism and Context*, Volume I. (Oxford: Oxford University Press, 2009) and David Lewis, "Elusive Knowledge," *Australasian Journal of Philosophy* 74 (1996): 549-567.

theory of explanation presented here will concern the extent to which we can claim that there are any non-contextual methodological standards that all explanations must meet. The specific view defended here is that there are some such invariant standards, but they are rather weak. This acknowledgement of the relative plasticity of explanatory contexts then in turn helps to explain the variety of explanatory practices of practitioners in different disciplines, the variety of explanatory practices at different times in the same discipline, etc.

One might be immediately tempted to object to this general account of explanation due to the perceived relativity that it imposes on the concept of explanation, and there are at least *prima facie* reasons to be sympathetic to this initial reaction. However even though such worries appear cogent it will be argued here that they are ultimately not serious worries. For the most part, this sort of worry is the result of baggage left over from previous accounts of explanation. Going back to Hempel's classic work on explanation, 'explanation' has generally been taken to be a success term and one of the chief desiderata of an adequate explanation is that it be true. So, for example, as explanation is traditionally understood, the Ising model of magnetism in solids cannot explain anything because the Ising model of magnetism is, strictly speaking, false. Given this long-standing desideratum of theories explanation it might appear that the theory of explanation sketched above will be unacceptable as it would seemingly appear to allow both that false theoretical claims can be explanations provided the correct context is present. But this problem is really a non-issue.

This is because what does not vary is whether or not a particular theoretical claim is a *potential* explanation of a phenomenon. Whether a particular theoretical claim is, or is not, a potential answer to a given scientific question is purely a matter of erotetic logic. There may be an infinite number of such answers that can be formulated with respect to any scientific question, but this does not in any way entail relativism of any sort in and of itself. Again, on the view developed here what most importantly varies with context are the epistemic standards by which we judge the *superiority* of explanations relative to one another. This involves the acceptability of the epistemic standards in question. Should the same context arise on more than one occasion, then the same evaluative ranking in terms of 'bestness' of explanation should result provided we are considering the same set of theoretical claims with respect to the same body of evidence and background knowledge. As such, substantive worries about the relativity of explanation seem largely unfounded. Such relativity as there is in this account is simply a function of the fact that the epistemic standards for acceptance of theoretical claims can vary across epistemic situation types. But, what it really indicates is just that

explanatoriness comes in degrees and that evidential standards can vary and nothing more radical than that.<sup>20</sup>

In line with this, it is well-known that IBE is a form of nonmonotonic inference.<sup>21</sup> For nonmonotonic inferences of this sort then a given theoretical claim  $T_i$  might be the best explanation of a body of evidence  $e$  in context  $B_k$ , while  $T_j$  might be the best explanation of  $e$  &  $f$  in  $B_k$  or of  $e$  in  $B$ .<sup>22</sup> It is in this sense that inference to the best explanation is then a kind of ampliative and defeasible inference, and it seems as if we might be able to represent this property of IBE while at the same time allowing for a sense in which it is probative. So, we need then to determine how to represent such inferences and when we can regard instances of IBE as “good” in a clear sense. But first there are some important other factors concerning IBE that need to be examined. First and foremost, in these sorts of inferences we typically restrict our attention only to some factors that make up a relatively well-defined inferential context. In these restricted contexts evidence is typically limited to some sub-set of the total known evidence  $e$ , where we limit the set of theoretical claims considered to a sub set of  $T$ —the set of all competing theoretical claims with respect to some phenomenon, and/or where we fix other particular methodological features that govern inferences. If information is added to our premises or contextual factors change, then what inferences are considered to be warranted can also change. As a result, this version of IBE reflects the defeasibility of IBE and this account of IBE squares well with the fact that, in actual practice, scientists accept theories but never make such inferences from complete bodies of evidence or from exhaustive sets of theoretical claims. This is primarily because of cognitive and computational limitations.

### 2.3 IBE

Preliminaries aside, we can then introduce this account of IBE. An *explanatory scientific problem*  $S$  will be taken to be a quintuple consisting of one or more why- or how-questions  $Q_n$ , a set of all competing theoretical claims  $T$  indexed to elements of  $Q_n$  that minimally fulfill a set of logical criteria EXP for what counts as an answer to a given question  $q$ , where  $q \in Q_n$ , the total body of relevant evidence  $E$  and a context  $B$ . So, the  $i$ -th ideal explanatory scientific problem will be written as  $S = \langle Q_n, T, E, B, \text{EXP} \rangle$ . However, as most scientific problems are complex there

<sup>20</sup> See Peter Railton, “Probability, Explanation, and Information,” *Synthese* 48 (1981): 233-256.

<sup>21</sup> See Gerhard Brewka, Jurgen Dix, and Kurt Konolige, *Nonmonotonic Reasoning: An Overview* (Stanford: CSLI, 1997) and Henry Kyburg and Choh Man Teng, *Uncertain Inference* (Cambridge: Cambridge University Press, 2001).

<sup>22</sup> See Lipton, *Inference to the Best Explanation*, 2<sup>nd</sup> ed., 92.

will be several members of  $Q_n$ , but in the simplest case—what we will call a *simple problem*— $Q_n$  will be a singleton and  $q_i = Q_n$ . Where  $S$  is complex there will be an appropriate number of  $T$  indexed to the elements of  $Q_n$ , and  $B$  will be similarly indexed. The solution to a given simple explanatory scientific problem—a given  $S$  where  $Q_n$  is a singleton—is then  $T_i$ , the element of  $T$  which satisfies EXP and fares best in terms of  $E$  and the various standards encoded in  $B$ . More realistic and contextually restricted explanatory scientific problems will involve restrictions of  $T$  and of  $E$ . In a given context  $B$  a research group trying to answer a given explanatory question  $q_i$  may limit consideration to  $T_n$ —a few select members of  $T$  such that  $T_n \supset T$ —or they may limit consideration to some sub-set  $\kappa$  of the total relevant known evidence  $E_k$ . For example, one crucially important way that  $T$  is restricted by  $B$  is via the introduction of idealizing assumptions.<sup>23</sup> In such cases, when a given idealizing assumption  $I$  is imposed in a given context it effectively rules out of consideration all theoretical claims that fail to hold under  $I$ . In other words doing so restricts consideration to  $I$ -simplified theories. Other ways of limiting  $T$  are common and include restricting consideration to extant theories, or restricting consideration to highly plausible theories, or simple differential comparisons of just two competitors, etc. So, one example of a more realistic construal of the  $i$ -th simple explanatory scientific problem can be written as  $S_i = \langle q_i, T_n, \kappa, B, \text{EXP} \rangle$ . Typically this reflects the fact that real scientific research concerning a simple explanatory problem involves a finite set of theories and some sub-set of the known relevant evidence in a fixed context that determines which methodological standards will be used to evaluate the competing theories. It is here that the work on bounded and ecological rationality will ultimately play an important role in understanding the probative nature of this complex form of inference. However, let us turn our attention at this point to saying a bit more about questions and their role in scientific explanation.

Following Åqvist and Hintikka, the sorts of questions we are interested in can be analyzed in terms of epistemic imperatives to bring about certain epistemic states.<sup>24</sup> So, we can analyze questions as requests by an agent to some external source of information to bring it about that the agent knows the answer. All well-formed questions of these sorts implicitly incorporate the *presupposition* of that question. The question ‘Is  $\phi$  the case?’ presupposes that  $\phi$  is the case or that it is not the case that  $\phi$ , and the question ‘Why is  $\phi$  the case?’ presupposes that  $\phi$  is the

<sup>23</sup> See Michael Shaffer, *Counterfactuals and Scientific Realism* (New York: Palgrave-MacMillan, 2012).

<sup>24</sup> See Åqvist, *A New Approach to the Logical Theory of Interrogatives, Part I* and Hintikka, *The Semantics of Questions and the Questions of Semantics*.

case. A question admits of satisfactory answers only if the presupposition of that question is true, or at least approximately true. In general we will indicate the presupposition of a given question with an expression of the form  $PR(q)$ . Minimally acceptable answers to questions are then propositions that allow us to understand the presupposition of that question to some degree. So, a minimally acceptable answer—or a *potential answer*—to a given simple scientific problem is a theoretical claim that at least in part explains the presuppositions of a given scientific problem. Acceptable answers to specifically scientific problems are theoretical claims that allow us to understand a phenomena or the law that the question is about.

This view then naturally looks very much like an erotetic approach to Peircean abductive/explanatory inference. However, Hintikka criticized the common view that abduction is a distinct and bona fide form of inference at all.<sup>25</sup> Against this common view Hintikka suggested that abduction is really a *search strategy* in the epistemic attempt to discover truth, as opposed to a form of inference. As Hintikka ultimately saw it, abductive search is the search for true answers to why-questions and why-questions are simply requests for explanations. So, according to Hintikka, abductive search is erotetic—it is a form of explanatory inquiry—but there is no such thing as abductive *inference* per se. The view defended here is, to a significant degree, in agreement Hintikka's. As it will be understood here, abductive search is the dynamic process of searching for explanatory answers to why-questions. But, the contention made here is that IBE is the terminal and inferential stage of abductive search. So, the position defended here is that abduction is not precisely the same thing as IBE. However, against Hintikka in particular, the view defended here is that inference to the best explanation is a form of inference employed in the broader process of abductive search, even if *abductive search* itself is not a form of inference. In any case, the attempt to construe how the members of  $T$  are demarcated with respect to some problem  $S$  requires that we address explicitly what constitutes EXP, the set of logical requirements that a given theoretical claim must fulfill in order to be considered a member of  $T$  in the context of some scientific problem.

## 2.4 Potential Explanations

We can now turn our attention to satisfying one of the three desiderata for an account of IBE mentioned earlier. Specifically, we can address what it is for one claim to be explanatory with respect to another. As this conceptual issue does not

---

<sup>25</sup> See Hintikka, "What is Abduction."

incorporate any evaluative or comparative elements the minimal requirements for membership in the set of potential answers to a given scientific problem are neither especially strong nor especially interesting. In point of fact, it will be suggested here that in an ideal world where there were no computational or physical limitations on scientific practitioners, the evaluation of which explanation is best with respect to a scientific problem would be purely a matter of logic, probability and statistics in the more formal sense. However, as has been stressed in earlier sections of this paper we do not live in such a world, and so we are often forced to simplify things by limiting our concern to those relevant theoretical claims that have been formulated and which satisfy certain additional contextual constraints, and to the relevant evidence of which we are aware. In any case we can now turn to discussion of the minimal criterion that a theoretical claim must satisfy in order to be included in the set of potential answers to a given explanatory scientific problem. As we saw earlier, for a given answer to an explanatory scientific problem to be counted as an explanation it must satisfy the basic principle EXP. EXP is then understood here as follows:

(EXP) With respect to background knowledge  $B$  and where  $T_i \in B$  and  $PR(q) \in E$ , theoretical claim  $T_i$  is a member of the set of potential answers to a simple problem  $S$ , or  $T_i \in \mathbf{T}$ , if and only if (1)  $P(PR(q) \mid T_i) > P(PR(q))$  and (2) for all  $T_j$   $\neg[P(PR(q) \mid T_i \& T_j) \leq P(PR(q) \mid T_j)]$ .<sup>26</sup>

EXP is by no means especially novel and has been assumed to be a basic tenet of theories of explanation for some time. As was alluded to earlier, we should be aware here the epistemic imperative to bring it about that the agent *knows* that  $p$  used in the erotetic analysis of explanation will have to be weakened somewhat. In the context of why-questions and recognizing that explanation comes in degrees, it seems that we really need only know that a theoretical claim raises the probability of the phenomena or law in question and that there is no other theoretical claim that wholly accounts for this increase in probability in order for a theoretical claim to be counted as a potential explanation of some data or of some lower level theoretical claim

Notice however that EXP does not narrow the range of explanations very much at all. As we noted and stressed earlier, it is well known that a non-finite number of theoretical claims can be arbitrarily constructed that satisfy EXP with respect to any problem  $S$  simply by taking a theoretical claim  $T_i$  and disjoining it with arbitrary strings of expressions. This just tells us that the purely logical aspects

---

<sup>26</sup> The second conjunct on the right hand side of the bi-conditional in EXP is included in order to rule out pseudo-explanations. See Alan Goldman, *Empirical Knowledge* (Berkeley: University of California Press, 1991).

of explanation are not very interesting and that they presuppose a sort of informational omniscience with respect to evidence and theory, and that we are forced by computational, cognitive and physical constraints to consider only those theoretical claims that we deem to be relevant from among those that have been explicitly formulated. In the unrestricted case  $T$  has the form  $\{T_1 \vee T_2 \vee T_3 \vee T_4 \vee \dots \vee T_n\}$ , while in real cases we only consider  $T_n$  of finite, and often quite small, cardinality and which hold only under idealizing assumptions. These more realistic cases of confirmation of competing theoretical claims are then often themselves cases of epistemic/methodological idealization where we are simplifying the confirmational context by reducing the number of theories that are being considered as serious candidates for confirmation by some given body of evidence that is itself restricted. As should then be obvious, the real substance of the account of theory acceptance developed here is to be found in  $B$ , the contextual factors that determine the epistemic standards in terms of which a given scientific problem is considered. In particular we must pay careful attention to those standards in addition to EXP that impact the ranking of explanations in given context. So, context determines which theoretical claims are taken to be relevant, what idealizing assumptions are allowed with respect to a given scientific problem and what factors will be used to rank explanations in addition to EXP. Context thereby determines  $T_n$ ,  $e_n$ ,  $I$  and the evidential and explanatory standards that characterize that explanatory scientific problem.

## 2.5 The Contextual Aspects of Explanation

Now we can focus our attention squarely on what might be the most interesting aspect of this account of IBE, its contextual aspects. More specifically, we can consider how epistemic context relates to epistemological standards operative in explanation. Finally, we can move on to consider in detail how we evaluate which explanation is best in a given context, and with this established we can formulate a general rule of theory acceptance based on those evaluative standards.

So, what is an epistemic context? Answering this question is of central importance in explicating the sort of account of IBE offered here, and we can get some help from looking at epistemic contextualism. There are at least two forms of contextualism and we can follow DeRose's terminology in order to locate the sort of contextualism appropriate to the sorts of explanatory endeavors in the physical sciences that we have been considering. Most crucially, DeRose distinguishes between *subject* contextualism and *attributor* contextualism.<sup>27</sup> On the one hand,

---

<sup>27</sup> See DeRose, "Contextualism: An Explanation and Defense."

subject contextualists hold that features of the (physical) context of the subject of knowledge vary (e.g. location), and so whether the subject knows something or not depends on those contextual factors. Certainly environmental facts about computation and cognition can impact whether we know something or not. Also, facts about the environment in which we are located can impact whether we know certain things. When, for example, a subject inhabits an environment littered with fake barns or robot cats, we might say that he does not know that he sees a barn or a cat when she is the subject of particular sensory stimulations. When a type identical subject with type identical sensory experiences inhabits an environment that is relatively free from these sorts of deceptions, we might say that he does know that he sees a barn or a cat. On the other hand, attributor contextualism holds that contextual features of the conversational context of the attributor of knowledge *to some other subject* vary, and so whether we are warranted in saying of someone that they know varies with these contextual factors. What will vary in this sort of contextualism are the epistemic standards by which we judge of someone that they are warranted in making a knowledge attribution.<sup>28</sup>

By and large, however, this distinction is superficial and it is not really necessary to opt exclusively for one or the other. This is simply because both kinds of contextual features are epistemically important. They are both essentially elements of what has typically been referred to as background knowledge. The former kinds of contextual factors are empirical facts about our cognitive limitations, computational capacities, physical environments, etc., and the latter kinds of contextual factors are pragmatic factors about how we are going to apply the term ‘explanation’ in light of our physical and epistemic situation. Furthermore, in a sense we are all both attributors and subjects of epistemic attributions, and being aware of one’s environmental context as well as being aware of one’s conversational context may make one’s own attributions of knowledge, or of justification, to others—or even to one’s self—different. In any case, the kind of contextualism that characterizes explanatory situations involves both aspects of attributor contextualism and aspects of subject contextualism. The view developed here will be framed in terms of attributor contextualism as that view will allow us to subsume the kinds of factors that are of interest in subject contextualism. So, what we are interested in determining is when, in context *B*, an attributor *a* is justified in claiming of some subject *b* that *b* has explained *e* or *T*<sub>1</sub> to some other agent *c*. In terms of the erotetic model of explanation outlined above, we are then ultimately interested in examining when in context *B* an attributor *a* is

---

<sup>28</sup> See DeRose, *The Case for Contextualism* and David Lewis, “Scorekeeping in a Language Game,” *Journal of Philosophical Logic* 8 (1979): 339-359.



justified in claiming of some subject  $b$  that  $b$  has provided an acceptable answer to a why-question about  $e$  or  $T_i$  to some other agent  $c$ . In other words, we want to know when  $b$  has met the imperative implicit in a scientific explanatory request, at least to some degree.

## 2.6 Best Explanation and Problem Substitution in the Sciences

So, now we can turn our attention to the issue of when are we justified in claiming of someone that they have provided the *best* answer to someone's request for explanatory information in a given specific context? This is essentially the question of when in context  $B$  of an attributor  $a$ ,  $b$  has explained  $e$  or  $T_i$  to  $c$ . Given this understanding of the erotetic model of explanation and our understanding of the contextual aspect of scientific explanation, we can claim that in context  $B$   $a$  is justified in claiming of  $b$  that  $b$  has explained  $e$  (or has explained  $T_i$ ) to  $c$  if and only if  $c$  has made a request 'Why  $e$ ?' or 'Why  $T_i$ ?' to  $b$  and  $b$  has conveyed to  $c$  that ' $T_j$ ' where  $T_j \in \mathbf{T}$  and  $T_j$  satisfies EXP. More importantly, we can now see that IBE can be presented in a similar manner. In context  $B$ , an attributor  $a$  is justified in claiming of some subject  $b$  that  $b$  has best explained  $e$  (or  $T_i$ ) to  $c$  if and only if  $c$  has made the request 'Why  $e$ ?' or 'Why  $T_i$ ?' to  $b$  and  $b$  has conveyed to  $c$  that ' $T_j$ ' where  $T_j \in \mathbf{T}$ ,  $T_j$  satisfies EXP, and  $T_j$  satisfies BEST. With respect to an ideal explanatory scientific problem involving  $\mathbf{T}$  and a given body of evidence  $e$ , BEST is then characterized as follows:

(BEST) If  $T_j$  satisfies EXP, then  $T_j$  is *the best (purely logical) explanation* of  $e$  in  $B$  if and only if  $\neg(\exists T_i)[(T_i \in \mathbf{T}) \& (P(e \mid T_i \& B) > P(e \mid T_j \& B))]$ .<sup>29</sup>

What defenders of IBE assert uniformly is that if this sort of principle is satisfied, then we are *defeasibly* warranted believing that  $T_j$ . In terms of the contextualist view of explanation presented here, what we are really allowed to say of a theory that satisfies BEST is that we are warranted in believing that  $T_j$  *in*

---

<sup>29</sup> This is to be understood as a partial empirical analysis of the logical aspects of explanation in the sense articulated in Carl Hempel, *Fundamental of Concept Formation in Empirical Science* (Chicago: University of Chicago Press, 1952). Also, in Lipton's 2004 terminology, best or "loveliest" explanation is not being completely identified here with likeliest explanation. The conjecture about what explanation is best offered here is that it is the theory that is most highly ranked from among competitors based on the total set of criteria present in a given context. This is meant to stave off criticisms of (virtual) triviality that apply to stand-alone account of IBE based solely on criteria like BEST. See Christopher Hitchcock, "The Lovely and the Probable," *Philosophy and Phenomenological Research* 74 (2007): 433-440 for this criticism. See Peter Achenstein, *Evidence and Method* (Oxford: Oxford University Press, 2013) for some additional criticisms of IBE.

context *B*. For our purposes here, notice that if we adopt BEST as a core component of a rule of theory acceptance, it allows us to assess the confirmational status of theories that are more or less realistic and it can easily be applied to cases where we are dealing with restricted sets of theories or restricted bodies of evidence.

With respect to a more realistic explanatory scientific problem involving the restriction of theories considered to  $T_n$  and to a given body of evidence  $e$ , BEST can be modified to reflect this as follows:

(BEST') If  $T_j$  satisfies EXP, then  $T_j$  is *the best (purely logical) explanation* of  $e$  in  $B$  if and only if  $\neg(\exists T_i)[(T_i \in T_n) \& (P(e \mid T_i \& B) > P(e \mid T_j \& B))]$ .

This then means that we can still maintain a coherent and normative sense of inference to the best explanation with respect to both ideal and realistic contexts. In what follows we will primarily deal with BEST, and we will simply acknowledge at this point that BEST' can be substituted for BEST when dealing with more realistic cases of theory confirmation. Finally, one might then define the differential degree of confirmation of theoretical claim based on a measure of explanatory power as follows.<sup>30</sup> With respect to an ideal explanatory scientific problem involving  $T$ , a given body of evidence  $e$ , and where  $T_j$  satisfies BEST and  $T_i$  is the second most likely theory relative to  $e$ ,

(CN)  $CN(T_i) = \text{diff}[P(e \mid T_j \& B), P(e \mid T_i \& B)]$ .<sup>31</sup>

So, on this particular view the differential degree of confirmation of a given best explanation is the degree to which it is more likely than the next most likely explanation of the same evidence.<sup>32</sup> Of course this can be similarly defined for more realistic scientific problems by replacing BEST with BEST'. Real scientific problems then can be formally understood as follows:  $S = \langle q, T_n, e, B, \text{EXP}, \text{BEST}' \rangle$ . As we shall see, however, there is typically much more to rules of theory

<sup>30</sup> This is but one possibility and is in no way a necessary component of the theory defended here.

<sup>31</sup> See Johnah Schupbach, "Comparing Probabilistic Measures of Explanatory Power," *Philosophy of Science* 78 (2011): 813-829 and Jonah Schupbach and Jan Sprenger, "The Logic of Explanatory Power," *Philosophy of Science* 78 (2011): 105-127 for discussion of other measures of explanatory power.

<sup>32</sup> There may also be other measures of the degree of confirmation or evidential support, but this one seems reasonable and (importantly) it is suitably differential. See Edward Erwin and Harvey Siegel "Is Confirmation Differential?" *British Journal for the Philosophy of Science* 40 (1989): 105-119 for discussion of the differentiability of inference to the best explanation. One related alternative that looks similarly promising has been articulated by Kyburg and Teng (*Uncertain Inference*, 103). It is derived from the work in John Kemeny and Paul Oppenheim, "Degree of Factual Support," *Philosophy of Science* 19 (1952): 307-324. This differential measure can be stated as follows:  $CN^*(T_i \mid e) = P(e \mid T_i) - P(e \mid \neg T_i) / P(e \mid T_i) + P(e \mid \neg T_i)$ .

acceptance at work in given contexts than EXP and BEST and this is part of the background knowledge present in such cases. But, more importantly, why should we regard this sort of inferential scheme as probative? If we cannot justify the probative nature of this account, then we are not entitled to hold that such inferences have normative force. So, why is inference to the best explanation a probative form of inference?

## 2.7 The Probative Nature of IBE

Many philosophers have raised objections with respect to IBE for a variety of reasons, but they have typically done so without explicitly acknowledging that IBE is nonmonotonic, that it is dynamic, and that such inferences often depend on simplifying assumptions with respect to the evidence entertained and the theories considered in those inferences. With respect to this latter feature, it is crucial to understand that typical cases of IBE are normative and depend (at least) on three simplifying assumptions. The first assumption is that scientists consider only a finite set of relevant theoretical claims when assessing what is the best explanation of some phenomenon or lower level theoretical claim.<sup>33</sup> Second scientists consider only a subset of the total known evidence relevant to a scientific explanatory problem. Thirdly, scientists typically deal with theoretical claims that hold only under one or more idealizing assumption. As we shall see, all of these assumptions are fixed by contextual factors.

That said, the standard and supposedly damning criticism of IBE in the literature is, of course, due to van Fraassen. The primary worry that he infamously raised about inference to the best explanation concerns the idea that we have no good reason to accept the best explanation of some phenomenon from among a finite set of actually formulated theoretical claims unless we have reason to believe that the true explanation is a member of the set we are considering. Of course, van Fraassen claims that we only ever deal with very small sets of such theoretical claims when those sets are compared to the set of logically possible, but unformulated, theoretical claims. So, van Fraassen concludes that IBE is not probative because it is more likely that we are accepting the best of a bad lot, and if we are just accepting the best of a bad lot then IBE does not track the truth. In other words, as he sees it, it is irrational to accept the conclusion of any actual IBE as likely to be true. Van Fraassen entertains three potential types of responses to this line of argument and he refers to these three general strategies as follows: the privilege strategy, the force majeure strategy and the retrenchment strategy.

---

<sup>33</sup>See especially van Fraassen, *Laws and Symmetry*.

The privilege response essentially involves the idea that we have some special ability to track the truth and so are entitled to believe that the true theory is among those we consider in inferring the best explanation from sets of known theories. As van Fraassen puts it, the privilege strategy depends on the dubious assumption that "...we are predisposed to hit on the right range of hypotheses."<sup>34</sup> The privilege response takes both naturalistic and rationalistic forms, but neither is at all compelling. There is simply no good reason to believe that the set of known hypotheses we deal with must contain the truth. The force majeure response involves the basic idea that we simply have no alternative and so must infer the best explanation from among the relevant set of known alternatives. But, van Fraassen rejects this response because forced choices are not necessarily rational choices. So, from the fact that we must infer the best explanation from among known explanations it does not follow that the best alternative is true. The retrenchment response involves rejecting inference to the best explanation and replacing it with an alternative account of theory acceptance. So, ultimately, he claims we are not entitled to believe in the truth of our best explanations and that we should engage in radical retrenchment in epistemology. In doing so, he rejects the appeal mysterious powers, and he is right to do so. However, his argument against the probativity of IBE is flawed and his negative assessment of the probativity of IBE is over-stated. The contention made here is that this is the case because his argument against IBE is based on an uncharitable understanding of the actual practice of inferring best explanations as it is done in actual practice.<sup>35</sup> The defense against van Fraassen's argument mounted here is then best understood as a sophisticated version of the *force majeure* response, and we shall see that it is one that enjoys considerable support from the HBP as well as the BER program.

The sense in which IBE is probative needs to account for the idea that IBE is nonmonotonic and that in inference to the best explanation we deal with incomplete information (i.e. evidence) and incomplete sets of explanatory theories.<sup>36</sup> In accord with these ideas, the appropriate notion of "goodness" for IBE is nonmonotonic and is a form of ideal case reasoning. What we are entitled to

---

<sup>34</sup> van Fraassen, *Laws and Symmetry*, 143.

<sup>35</sup> Specifically, it involves all the elements of abductive search as understood in Hintikka, "What is Abduction."

<sup>36</sup> So, in his "Is the Bad Lot Objection Just Misguided?" Schupbach is correct to note that van Fraassen simply misses the point when he criticizes IBE as a probative form of inference in criticizing the quality of the inputs to which IBEs are applied. When coupled with Hintikka's understanding of the dynamic nature of abductive search from his "What is Abduction?" all of van Fraassen's worries go away. IBEs are simply inferences made in dynamic contexts where we are constantly updating the sets of hypotheses and bodies of evidence to which IBEs are applied.

assert when we use IBE is that in worlds that are more epistemically perfect than but still similar to the actual world, it is the case that (at least) one of the theoretical claims in  $T$  is more likely to be true than the others. The sense in which these worlds are ideal or perfect is that in such worlds we know of all the alternative theories, we know all the relevant evidence and we are able to assess those theories in terms of BEST (and whatever other norms are in place in a given context). Since that ideal case claim is true with respect to ideal worlds, we *should* employ IBE in actual practice and so it is an appropriate norm with respect to real world science. This is a sort of Kantian approach to normativity and it is based on the following sort of argument.<sup>37</sup> A fully rational scientist would select the best explanation from among all possible alternatives on the basis of all evidence. If a fully rational scientist would select the best explanation from among all possible alternatives on the basis of all evidence, then an imperfectly rational scientist ought to select the best explanation from among all possible alternatives on the basis of all evidence. Therefore, an imperfectly rational scientist ought to select the best explanation from among all possible alternatives on the basis of all evidence. Actual scientists are, of course, imperfectly rational. Therefore, actual scientists ought to select the best explanation from among all possible alternatives on the basis of all evidence. But, we can only be reasonably expected to obey norms to the degree that we can actually do so. So, we can further reason as follows. If actual scientists ought to select the best explanation from among all possible alternatives on the basis of all evidence but they are not capable of doing this at time  $t$ , then actual scientists ought only to do their best to select the best explanation from among all possible alternatives on the basis of all evidence at time  $t$ . Therefore, actual scientists ought only to do their best to select the best explanation from among all possible alternatives on the basis of all evidence at time  $t$ . So, the best actual scientists can hope to achieve in any given context at a given time is to select the best explanation of a phenomenon from among known hypothesis on the basis of known evidence. That is typically the best that we can do in our imperfect circumstances. We are limited beings in environments that constrain our abilities to reason and so we must often substitute more easily solvable problems for those that are beyond our abilities in a given context.

So, the purely probabilistic rule BEST (in conjunction with any additional norms in our background knowledge) tells us how to evaluate theories on the basis

---

<sup>37</sup> The argument presented here depends heavily on the interpretation of Kant from Robert Holmes, *Basic Moral Theory*, 4<sup>th</sup> ed. (New York: Cengage, 2006). See Michael Shaffer, "Bealer on the Autonomy of Philosophical and Scientific Knowledge," *Metaphilosophy* 38 (2007): 44-54 for discussion of ideal case counterfactuals.

of evidence in such situations, and in such cases we are warranted in accepting the theoretical claim that maximizes likelihood *even if we do not actually meet the preconditions of the ideal case claim*. We can be governed by the ideal norm and yet also be warranted in following its real world correlate because we cannot do any better. The normatively correct acceptance of theories in real world contexts then amounts to our being warranted in accepting the best of a *known* lot of hypothesis on the basis of *known* evidence in a given context. In other words, it is rational for us to employ the availability heuristic. In such cases we are entitled to accept the theory that maximizes likelihood from among known theories on the basis of known evidence, at least pending the introduction of more evidence, or the introduction of new theoretical claims, or other changes in context. In essence, we must settle and accept that if the restricted set of theoretical claims *were* the set of all possible theoretical claims and the evidence of which we are aware *were* all of the evidence, then we would be entitled to accept that theoretical claim which maximizes likelihood on that evidence as true in that context. What else could we do in such a situation? In fact, to claim that IBE of this sort is irrational would commit us to wholesale skepticism about explanation and about science and it would be totally at odds with actual practice. The history of scientific practice just is the history of explaining to the degree that we currently are able and so problem substitution is the bread and butter of explanatory science. We seek to solve simpler explanatory problems first and then attempt to deal with their more complex incarnations.

However, it is clear that in typical scientific contexts there are more norms at work than just BEST. Since we do science in the actual world and not in normatively perfect worlds, we also have to do our best to close the gap between the actual world and the normatively ideal world. Properly conducted science typically requires us to attempt to gather more evidence, to generate new and better evidence using new methods, and so on. It also typically requires us to formulate and consider new competing hypotheses. As such, science is typically conducted under the assumption of the following two additional norms, the norm of evidential generation and the norm of theoretical innovation:

(EVG) We should gather and generate evidence using the best means available.

(THI) We should formulate and consider hypotheses.<sup>38</sup>

---

<sup>38</sup> These norms are part of the more broad process of abductive search as understood in Hintikka, "What is Abduction" and IBE can them be understood as the terminal and inferential stage of such abductive inquiry.

These are then norms of bias correction that allow us to alleviate worries about the kinds of biases that can arise from the kind of problem substitution that the availability heuristic involves. EVG and THI then allow us to offer an answer to van Fraassen's worries about IBE based on the nonmonotonic and dynamic practice of inferring explanations on the basis evidence. Dynamic and contextual IBE is a defeasible but probative form of inference that says that we should always accept the best available explanation of the available evidence in a given context, *but that is by no means the end of the story at all*. We should also strive to satisfy EVG and THI so that we come closer to satisfying the ideal case norm by correcting biases over time. So, while it is true that in some context at some time we may be accepting the best of a bad lot this need not be true in the long run. From the fact that actual conditions are not normatively perfect, it does not follow that it IBE is irrational and it does not follow that it does not track the truth in the long run. In effect, what we can see is that real scientific problems are dynamic in nature. So, real dynamic scientific problems are sequences of problems with the following form:  $S = \langle q, T_n, \kappa, B, EXP, BEST', EVG, THI \rangle$ . They are instances of the application of problem substitution involving the availability heuristic to ideal problems of the form:  $S = \langle Q_n, T, E, B, EXP, BEST \rangle$ . Given EVG and THI such sequences of  $S$ s will involve sets  $T_n$  and  $\kappa$  that are being expanded sequentially as we become aware of new evidence and new theories in our search for the truth. Typical, environmentally situated, members of such sequences will be simplified version of a complete and far more complex problem. But, solving the simpler problems very often yields insight into the answers to those complete problems. The simpler explanation provide partial understanding of the very same phenomena that the more complex explanations more fully explain. There are however some other aspects of this theory of explanation that are in need of a bit more detailed discussion, especially as they pertain to the robust evaluation of what theory is the best explanation in a given context.

## 2.8 The Variety of Explanatory Practices

As stressed at the beginning of this paper what is then important to recognize is that given this very general account of explanation, we can account for the variety of explanatory practices in the various sciences and their respective sub-fields in terms of the different additional methodological norms that are elements of the contexts that characterize those disciplines. So, the standards required for the confirmation of the existence of a particle in high-energy physics may be very high, this need not be true for the confirmation of a claim that a patient has a particular psychological disorder in clinical psychology. Moreover, some scientific

contexts may require that acceptable explanations are causal/mechanical, while others may require only statistical models. Some contexts may allow black box explanations, while others may not. Similarly, in some scientific contexts that characterize problems in physics or chemistry general laws may be required to explain, whereas in others such as biology or archaeology only singular causal explanations may be required to explain. Finally, we may find that more general methodological norms like simplicity, predictive novelty, conservativeness and so on characterize scientific practice in different contexts. What is of great importance is that we recognize that this aspect of the contextual theory of IBE is an asset as opposed to a problem. This is because, while the theory developed here ties explanation to understanding in a minimal and partial way via EXP and BEST and thereby unifies explanatory practice in a normative way at a very generic level, it is compatible with the observed variety of explanatory practices in the sciences and the variety of additional methodological norms that characterize individual contexts. This means then that BEST is not a full account of IBE. It is merely a core part of the theory of what counts as the best explanation in a given context and this rule can be supplemented with all sorts of additional criteria that might be elements of our background knowledge. How these additional features count in ranking hypothesis beyond the ranking imposed on the set of potential answers to a given scientific problem will itself be a function of the background knowledge present in the context of that explanatory problem. This then further suggests that there are different epistemically virtuous senses of understanding as well that correspond to the satisfaction of different sets of scientific and methodological desiderata and also that there are different degrees of explanatory understanding. So, as suggested earlier, this view is particularly well suited to the naturalistic studies of the sciences and the study of the diversity of methodological practices that we find therein. With respect to the theory developed here, what this amounts to is just the idea that we cannot really assess the confirmational status of theoretical claims absent some serious understanding of the methodological features of actual scientific contexts. Nevertheless, once we have established the details of a given context the confirmational status of a given theory can be assessed in terms of EXP, BEST' and whatever additional norms happen to characterize that context.

### **3. Rational Heuristics, Ecological Rationality and Explanatory Contextualism**

What is then worth emphasizing here is that, from the perspective of the voluminous literature on the psychology of human reasoning, the quasi-formal and philosophical view of explanation developed in this paper enjoys considerable



empirical support. This is secured via its natural relationship to the expansive body of work on fast and frugal reasoning heuristics for problem solving and some of its close relatives, including the BER. In particular the work of Gerd Gigerenzer and Daniel Kahneman, Paul Slovic and Amos Tversky are of special importance here.<sup>39</sup> As noted throughout this discussion, one core idea behind the concepts of the HBP and of BER is that real agents do not have unlimited computational capacities, time, complete information, etc. and that the heuristic rules of inference and decision-making that real agents use are normatively appropriate only relative to specific environments for which they have been evolutionarily developed. The idea then is that we need to explore the manner in which real inferences and decisions are made by actual cognizers in order to see how it is that such reasoning is done quickly and frugally based on our actual abilities. The second core idea relevant here is the concept of ecological rationality. The idea here is that real reasoning is not the result of a generic, domain-independent, capacity to deliberate and reason in accordance with some universal rules of rationality cashed out in terms of informational omniscience. As a result, the heuristics for reasoning and decision-making advocated by this approach are the results of and work only in the specific environments in which they are generated, presumably by evolutionary adaptation.

What is then important for the purposes of this paper is that the formal model of explanation developed here is readily compatible with this more general and realistic model of reasoning and decision-making. This is primarily because of two reasons. First, inferring best explanations from known sets of hypotheses and data can be understood to be a normative heuristic guided process that reflects our finite epistemic abilities. It crucially involves problem substitution and the availability heuristic. The availability heuristic is an epistemic norm that we ought to follow, but, more importantly, it is one which we can follow. It is normative in the short run in the sense that the best available explanation of the available evidence is the most likely explanation *from that set*. It is normative in the long run in the sense that we ought to continue to gather new and better evidence and to formulate new and better theories in order to combat the kinds of biases that the availability heuristic can introduce in its short run applications. So, the dynamic aspects of the account allow for the idea that such inferences are normative but revisable in light of newly acquired evidence and newly formulated theories. The process of explanatory reasoning is dynamically rational in the nonmonotonic sense. Second, the central role that contextuality plays in the account of IBE

---

<sup>39</sup> See Kahneman, Slovic and Tversky, *Judgment under Uncertainty* and Gigerenzer, *The Adaptive Tool Box*.

developed here is simply a way of formally representing the ecological aspects of real-world reasoning. We infer best explanations in real contexts governed by a variety of constraints that are the result of our epistemic finitude, our real environments and our background knowledge. So, explanatory contextualism is usefully be understood to be a formal analog of the ecological facts that constrain actual human reasoning that motivate problem substitution. Facts about our abilities and the environments we inhabit constrain us in the process of abductive search in general and specifically in the ultimate stage of such inquiry, IBE. It is virtually platitudinous to assert that we can only reason in terms of what is psychologically available to us given our computational abilities. But, we can ultimately be successful in explaining and understanding the world when we realize that IBE is also dynamic. Having the best explanation of some phenomenon in one simplified context is by no means the end of abductive inquiry. The employment of the availability heuristic opens the door to bias and incompleteness, but such biases and lacuna are correctable because reasoning is dynamic and problem contexts change over time. This allows us to search for deeper and more complex explanations as context changes and we are able to contend with greater complexity or become aware of new theories and evidence.

#### **4. Conclusion: Dynamic Contextual IBE and Abductive Search for the Truth**

So, by taking the HBP and BER conception of rationality seriously—specifically by appeal to the availability heuristic and the more general notion of problem substitution—we can see that IBE, the terminal inferential stage of abductive search, is rationally grounded. Moreover, this approach to IBE allows for a more sophisticated understanding of IBE as a dynamic and contextual sort of reasoning that functions in the context of the search for explanations. So understood IBE can be defended against van Fraassen’s “best of a bad lot” objection to IBE and, contrary to van Fraassen’s claims, it is rational to accept the conclusions of IBEs even if we are not in possession of the total set of logically possible explanatory theories of some body of evidence. But, IBE is not a static kind of inference and it yields provisionally true conclusions that hold relative to the context in which they are made, but context can change and so the specific standards used to judge bestness of explanations, the set of theories considered and the body of evidence explained can change. All of this reflects actual explanatory practice in the sciences much more accurately than does the static view of IBE.

# THE PERMISSIBLE NORM OF TRUTH AND “OUGHT IMPLIES CAN”

Xintong WEI

**ABSTRACT:** Many philosophers hold that a norm of truth governs the propositional attitude of belief. According to one popular construal of normativity, normativity is prescriptive in nature. The prescriptive norm can be formulated either in terms of obligation or permission: one ought to or may believe that  $p$  just in case  $p$  is true. It has been argued that the obligation norm is jointly incompatible with the maxim *ought implies can* and the assumption that there exists some truth that we cannot believe. The problem of the incompatible triad has motivated some to adopt the permissible norm of truth. I argue that the permissible norm faces an analogous problem of the incompatible triad.

**KEYWORDS:** epistemic norms, ought implies can, nature of belief, the truth norm of belief

## 1. Introduction

Most philosophers hold that there is a standard of correctness for belief: a belief that  $p$  is correct if and only if  $p$  is true. Belief is subject to the norm of truth. Philosophers disagree, however, about whether the norm of truth is genuinely normative and whether belief is *essentially* subject to the norm of truth. According to one popular construal, normativity is prescriptive in nature, i.e., a prescriptive norm is essentially capable of guiding and it issues requirements, permissions or prohibitions. Genuine norms tell one what one ought (not) to do under given circumstances.<sup>1</sup>

Assuming the prescriptive construal of normativity, there are two intuitive ways to formulate the norm of truth governing the attitude of believing:

( $\vec{T}_O$ ) For any  $S$ ,  $p$ :  $S$  *ought to* believe that  $p$  if and only if  $p$  is true.

( $\vec{T}_P$ ) For any  $S$ ,  $p$ :  $S$  *may* believe that  $p$  if and only if  $p$  is true.

---

<sup>1</sup> The prescriptive construal of the truth norm is widely endorsed, for discussion of alternative construal in evaluative and teleological terms, see, Conor McHugh and Daniel Whiting, "The Normativity of Belief," *Analysis* 74, 4 (2014).

One problem with  $\vec{\mathcal{T}}_O$  is that we *cannot* believe every truth that is out there in the world, as such,  $\vec{\mathcal{T}}_O$  clashes with the principle *ought implies can* (OIC). In other words,  $\vec{\mathcal{T}}_O$ , OIC and the claim that there are cases where if p is true S cannot believe that p are jointly incompatible. Call this the problem of the incompatible triad.

The problem of the incompatible triad has motivated *normativists* to either revise  $\vec{\mathcal{T}}_O$  or adopt  $\vec{\mathcal{T}}_P$ .<sup>2</sup> In this paper, I will focus on the second strategy as developed by Daniel Whiting. I will first present the problem of the incompatible triad and show how it motivates Whiting's permissible norm  $\vec{\mathcal{T}}_P$ . I then show that  $\vec{\mathcal{T}}_P$  faces an analogous version of the incompatible triad. I will conclude by briefly considering the implication of the result in the debate concerning the truth norm of belief.

## 2. The Problem of the Incompatible Triad

According to Whiting,<sup>3</sup> the prescriptive formulation of the truth norm  $\vec{\mathcal{T}}_O$  seems too demanding, given that we are ordinary epistemic agents with finite cognitive powers. Since there are infinitely many truths in the world, and S cannot, surely, believe every single one of them,  $\vec{\mathcal{T}}_O$ , therefore, faces the following incompatible triad:

( $\vec{\mathcal{T}}_O$ ) For any S, p: S ought to believe that p if and only if p is true.

(OIC) For any S,  $\phi$ : Necessarily, if S ought to  $\phi$  then S can  $\phi$ .

(Limited Capacity (LC)) There are cases where if p is true, S cannot believe that p.<sup>4</sup>

---

<sup>2</sup> For instance, Paul Boghossian proposes a weaker version of  $\vec{\mathcal{T}}_O$  by dropping the biconditional—for any S, p: S ought to believe that p *only if* p is true, in his "The normativity of content," *Philosophical Issues* 13, 1 (2003): 37. Ralph Wedgwood suggests that  $\vec{\mathcal{T}}_O$  should be restricted to propositions that one considers in his "Doxastic Correctness," *Aristotelian Society Supplementary Volume* 87, 1 (2013); "The Right Thing to Believe," in *The Aim of Belief*, ed. Timothy Chan (Oxford University Press, 2013).

<sup>3</sup> Daniel Whiting, "Should I Believe the Truth?" *Dialectica* 64, 2 (2010):213-224.

<sup>4</sup> According to doxastic involuntarism, belief-formation is not under our voluntary control. But given OIC,  $\vec{\mathcal{T}}_O$  implies that we have voluntary control over our belief-formation. Therefore, OIC, doxastic involuntarism, and  $\vec{\mathcal{T}}_O$  also seem jointly incompatible. For the classic discussion on doxastic involuntarism, see William P. Alston, "The deontological conception of epistemic justification," *Philosophical Perspectives* 2 (1988). I assume that some form of doxastic voluntarism is correct.

Motivated by the problem of the incompatible triad, Whiting contends that we should reject  $\vec{T}_O$  and instead adopt  $\vec{T}_P$ . After all, by weakening the deontic requirement from an *obligation* to a *permission*, we avoid the triad.  $\vec{T}_P$  is compatible with OIC and LC, since  $\vec{T}_P$  does not say anything about what one ought to believe. There is no relevantly parallel principle of “*may implies can*,” by which we could derive a statement about what one can believe under  $\vec{T}_P$ . Adopting  $\vec{T}_P$  therefore solves the original problem of the incompatible triad.

### 3. Does $\vec{T}_P$ Escape the Incompatible Triad?

Upon a closer inspection, however,  $\vec{T}_P$  faces an analogous problem. To see this, I will first show that the permission norm implies a falsity norm and, second, identify a corresponding claim of Limited Capacity\* (LC\*) that is incompatible with the falsity norm and OIC.

To facilitate our discussion, I shall follow the notations in standard deontic logic (SDL).<sup>5</sup> “Ought” is understood in terms of the propositional operator **OB** (It is obligatory that...). According to SDL, **OB** is a modal operator and the deontic formulas are evaluated with respect to sets of worlds, in which some are ideal. For our purpose, I adopt the standard semantics for deontic operators, which appeals to possible worlds semantics in which all worlds are ranked—some worlds are better than others. I will leave it to the reader to decide how to best construe ideality with the background theory they prefer (nothing in particular will hinge on this here).<sup>6</sup> The dual concept of “ought,” i.e. “may,” is abbreviated using the operator **PE**. As is common, the modal operator **PE** is defined in terms of **OB**:

$$\mathbf{PE} x =_{\text{def}} \neg \mathbf{OB} \neg x.$$

It is not difficult to show that  $\vec{T}_P$  entails a falsity norm.  $\vec{T}_P$  can be broken into two conditionals:

$$(\vec{T}_P) \text{ For any } S, p: p \text{ is true} \rightarrow \mathbf{PE} (S \text{ believes that } p)$$

$$(\vec{T}_P) \text{ For any } S, p: \mathbf{PE} (S \text{ believes that } p) \rightarrow p \text{ is true}$$

Using contraposition,  $\vec{T}_P$  is equivalent to:

<sup>5</sup> See, Paul McNamara, “Deontic logic,” in *Stanford Encyclopedia of Philosophy*, ed. Ed Zalta (2010), <<https://plato.stanford.edu/entries/logic-deontic/>>.

<sup>6</sup> The standard semantics is defended by pioneering deontic logicians such as, Lennart Åqvist, “Interpretations of Deontic Logic,” 73, 290 (1964); David K. Lewis, *Counterfactuals* (Blackwell, 1973). More recently, it is also defended by Ralph Wedgwood, *The Nature of Normativity*, vol. 2 (Oxford University Press, 2007), chapter 5.

For any S, p: p is false  $\rightarrow$   $\neg$ PE (S believes that p)

Hence, given that PE  $x =_{def}$   $\neg$ OB $\neg$ x,  $\vec{\mathcal{T}}_p$  is equivalent to the following falsity norm:

( $\vec{\mathcal{F}}_O$ ) For any S, p: p is false  $\rightarrow$  OB ( $\neg$  S believes that p)

In other words,  $\vec{\mathcal{T}}_p$  entails that for any S, p, if p is false then S ought not to believe that p. Whiting is aware that  $\vec{\mathcal{T}}_p$  entails the obligation norm  $\vec{\mathcal{F}}_O$ , after all, obligation and permission are dual deontic concepts. He regards this as a welcome result because it offers a response to the criticism that  $\vec{\mathcal{T}}_p$  is not normatively interesting.<sup>7</sup> According to Whiting,  $\vec{\mathcal{T}}_p$  is normatively interesting just because it is capable of guiding our belief-formation through  $\vec{\mathcal{F}}_O$ , which tells us that we ought to refrain from believing p when p is false. Moreover, on Whiting's view,  $\vec{\mathcal{F}}_O$  captures a more fundamental aim of belief, namely, to avoid falsity.

However, given OIC,  $\vec{\mathcal{F}}_O$  implies that we can refrain from believing whatever that is false. There is the analogue of the incompatible triad for  $\vec{\mathcal{T}}_p$ , since  $\vec{\mathcal{F}}_O$  and OIC are jointly incompatible with the following claim:

(LC\*) There are cases where if p is false, S cannot refrain from believing that p.

Whiting quickly dismisses the problem by rejecting (LC\*). He considers a case where someone is said to be psychologically unable to refrain from believing that there are aliens. Suppose that there are no aliens. Does example like this show that LC\* is true? Whiting thinks not. First, he complains that the relevant modality of "can" figured in OIC is weaker than psychological possibility. He suggests that ought to  $\phi$  implies that it is *humanly possible* to  $\phi$ . Second, he argues that critics of  $\vec{\mathcal{T}}_p$  have not shown there are cases where if p is false, it is humanly impossible to refrain from believing that p. Finally, he claims that even if the critic of  $\vec{\mathcal{T}}_p$  can show that there are such cases, there is a further question whether the attitude S has towards p counts as a genuine belief.

The question regarding the modality of "can" is indeed an important one. However, Whiting's suggestion that the "can  $\phi$ " figured in OIC should be understood as "humanly possible" to  $\phi$  seems ill-motivated and lacks reference to the relevant literature on OIC. According to the standard interpretation of "can  $\phi$ ," one can  $\phi$  just in case one (1) has the ability to  $\phi$  and (2) has the opportunity to exercise that ability to  $\phi$ .<sup>8</sup> On one influential view, one has an opportunity to  $\phi$  if

<sup>7</sup> See, for instance, Kathrin Glüer and Åsa Wikforss, "Against Belief Normativity," in *The Aim of Belief*, ed. Timothy Chan (Oxford University Press, 2013). See also Krister Bykvist and Anandi Hattiangadi, "Does Thought Imply Ought?," *Analysis* 67, 296 (2007).

<sup>8</sup> Such formulation is widely adopted in the debate concerning OIC. See, for instance, David

there is a non-zero objective chance to  $\phi$  assigned by the relevant psychological laws, where psychological laws are laws that are broadly based on folk-psychology and deal with agent's actions and attitudes.<sup>9</sup>

On this common interpretation of “can  $\phi$ ” as having the ability and opportunity to  $\phi$ , given the psychological laws governing agent's actions and attitudes, we have at least some reasonably good grasp of what “can  $\phi$ ” amounts to, broadly based on folk-psychology. By contrast, Whiting does not explain his notion of “humanly possible” to  $\phi$ . On the face of it, whether it is “humanly possible” to  $\phi$  would depend on the kind of creature we are, empirically speaking. If that's right, a natural way to flesh out what is “humanly possible” to  $\phi$  is just the standard interpretation of *can  $\phi$* . It is *humanly possible* for S to  $\phi$  just in case S has the ability and opportunity to  $\phi$ , given the psychological laws governing agent's actions and attitudes.

That being said, I agree with Whiting that critics of  $\vec{T}_P$  are yet to show that LC\* is true. The case Whiting offers on behalf of his critics—that of a person who cannot refrain from believing that there are aliens does not lend much support to LC\* because it is hardly convincing that, in so far as how the case is described, that the person genuinely cannot refrain from believing that there are aliens. I now turn to the task of offering three more persuasive cases in support of LC\*.

First, some beliefs might be deeply integrated in our psychological make-up that we cannot refrain from having them. Consider forms of clinical delusions, e.g. patients with Capgras delusion cannot refrain from believing that a close relative has been replaced by an impostor, often due to cognitive failure including abnormal perceptual experiences (as a result of a malfunctioning face recognition system) and possibly also with a deficit in their belief evaluation system.<sup>10</sup> Now, of course, few of us suffer from clinical delusions, yet I think some of our core beliefs may be psychologically impossible to shake off in a rather similar way as a result of how we are hard-wired to perceive the world. In fact, many philosophical theories, if correct, would render some of our core beliefs false. For instance, if error theories about mathematics and ethics are correct, none of our mathematical and

---

Copp, “‘Ought’ Implies ‘Can’ and the Derivation of the Principle of Alternate Possibilities,” *Analysis* 68, 297 (2008): 67 fn2; P. A. Graham, “‘Ought’ and Ability,” *Philosophical Review* 120, 3 (2011); Moti Mizrahi, “Does ‘Ought’ Imply ‘Can’ from an Epistemic Point of View?,” *Philosophia* 40, 4 (2012); Moti Mizrahi, “‘Ought’ Does Not Imply ‘Can’,” *Philosophical Frontiers* 4, 1 (2009); Peter B. M. Vranas, “I Ought, Therefore I Can,” *Philosophical Studies* 136, 2 (2007); Ralph Wedgwood, “Rational ‘Ought’ Implies ‘Can’,” *Philosophical Issues* 23, 1 (2013).

<sup>9</sup> See Wedgwood, “Rational ‘Ought’ Implies ‘Can’,” 87.

<sup>10</sup> For a recent overview of neuropsychological accounts of delusions, see Lisa Bortolotti, *Delusions and Other Irrational Beliefs* (Oxford University Press, 2009).

ethical beliefs are literally true.<sup>11</sup> If the B-theory of time is correct, then the passage of time is an illusion and the present is not ontologically privileged.<sup>12</sup> And yet, arguably, we cannot refrain from having beliefs about temporal experiences, that  $2+2=4$ , or that murder is wrong.

Second, it is not always within our power to avoid falsity if all evidence available to our epistemic community supports the false belief in question. For example, we might say that the best evidence available to the ancient Greek supports the claim that Phosphorus and Hesperus are two different celestial bodies. Given the restricted epistemic circumstances back then, arguably one cannot revise the false belief that Phosphorus and Hesperus are two celestial bodies. Similarly, some of our current scientific beliefs may turn out false, yet we may not be able to revise them if they are supported by what our best evidence suggests. Of course, given the development of science and technology, more evidence will become available and we will be able to spot more falsehoods and revise our beliefs accordingly. Indeed, this is the story of our scientific progress. However, for any given period of time, our epistemic position is always limited and we cannot revise our false beliefs if they are supported by the best evidence available at the time.

Third, some propositions are deeply integrated in our epistemic life, such as the so-called cornerstone propositions. We cannot refrain from accepting them despite the possibility that they are false.<sup>13</sup> If I were a brain in a vat, then those cornerstone propositions would be false. Yet, can I genuinely refrain from believing those cornerstone propositions? Perhaps in an epistemology seminar I can momentarily refrain from believing cornerstone propositions while entertaining the sceptical scenarios. However, it is hard to imagine that we can carry on refraining from believing cornerstone propositions if we were to live a normal epistemic life, since if I did not believe that I am not a brain in vat, I would not be able to have the ordinary empirical beliefs which are crucial for me to navigate through the world. Of course, the point here is not to claim that scepticism is true. Rather, the point is to emphasize that there are some propositions at the core of our belief system that we cannot refrain from believing, given the kind of creatures we are. As such, if scepticism were true, we would not

---

<sup>11</sup> See, notably, Hartry Field, *Realism, Mathematics & Modality* (Blackwell, 1989); John L. Mackie, *Ethics: Inventing Right and Wrong* (Penguin Books, 1977).

<sup>12</sup> For an influential account of B-theory of time, see, for instance, Theodore Sider, *Four Dimensionalism: An Ontology of Persistence and Time*, vol. 3 (Oxford University Press, 2001).

<sup>13</sup> The concept of cornerstone proposition is first coined in Crispin Wright, "Warrant for Nothing (and Foundations for Free)?," 78, 1 (2004), which is inspired by Wittgenstein's idea of hinge proposition in his *On Certainty*, eds. G.E.M. Anscombe and G.H. von Wright (Harper Torchbooks, 1969). My use of cornerstone proposition simplifies the details of Wright's account.



be able to refrain from believing false cornerstone propositions and hence  $LC^*$  is true.

In short, the underlying thought is this: given the psychological and cognitive constraints, and the fact that the world is not always cooperative, we cannot avoid all falsity. “Seek all truths” and “avoid all falsity” are really two sides of the same coin. If we think the former clashes with OIC, there is *prima facie* reason to think the same applies to the latter, given that we are finite epistemic agents. Hence  $\vec{T}_p$  faces an analogue of the incompatible triad, and so in that respect does not fare any better than the obligation norm which it aims to replace.

Now you might point out that Whiting could still maintain that even if we have shown that  $LC^*$  is true, there is a further question as to whether the attitude in question is in fact a belief. Whiting might insist that if the above three kinds of cases are cases where a subject cannot but have a belief-like attitude towards the propositions in question, then that attitude is not that of belief. Suppose that I cannot refrain from believing, say, that  $2+2=4$ , even in the presence of overwhelming evidence that mathematical fictionalism is true, then, it may be argued that my attitude towards the proposition  $2+2=4$  is not that of belief.

I do not see how Whiting can maintain this point without presupposing a normative account of belief—the very claim that is at issue in the debate. On a normative account of belief, belief is essentially governed by the truth norm, as such, an attitude that is insensitive to evidence and fails to be revised according to the truth norm cannot count as belief. However, to assume this normative account of belief is to beg the question against the critics of the truth norm, who are likely to deny that belief is essentially governed by the truth norm. Without presupposing a normative account of belief, it is hard to see why my attitude towards that  $2+2=4$  fails to be a belief, as long as the attitude plays the kind of functional role belief plays in one’s mental economy.

## 5. Conclusion

If the case for  $LC^*$  is successful, then the problem of the incompatible triad poses a challenge not only for  $\vec{T}_o$ , but for  $\vec{T}_p$  and  $\vec{F}_o$  as well. In so far as one endorses the principle OIC, one cannot avoid the triad by weakening the deontic requirement from an obligation to a permission. Neither can one avoid the triad by adopting an obligatory norm of avoiding falsity.

It is also worth pointing out that Wedgwood’s version of  $\vec{T}_o$  would not escape the triad either. On Wedgwood’s account, one ought to believe a true proposition if one considers that proposition, which is compatible with OIC and  $LC$  since the revised truth norm does not require one to believe the infinite many

truths out there that one never ever entertains. Now, if I am right about LC\*, then Wedgwood's version of  $\vec{J}_O$  does not escape the problem of the triad because it is incompatible with OIC and LC\*. Why? Presumably, in virtue of having an occurrent belief that p, one does consider the proposition involved in that belief. So, if the proposition is in fact false, then on Wedgwood's account, one ought to revise that belief, which may be something one cannot do given that for some p, one cannot but believe that p.

The normativists' hands are therefore tight. There remain two options. The normativist may appeal to a different construal of normativity that is not necessarily prescriptive in nature. For instance, many have developed an evaluative account of the truth norm.<sup>14</sup> The idea, roughly, is that it is good or ideal to have true beliefs, even if one cannot always believe the truth. The evaluative construal of  $\vec{J}_O$  can avoid the original incompatible triad since it does not issue any requirement. Alternatively, the normativist could simply reject OIC. Numerous authors have recently challenged OIC in light of empirical evidence and counterexamples, independently of the problem that concerns us here.<sup>15</sup> However, neither option is available to Whiting. If he wants to maintain the original motivation for adopting  $\vec{J}_P$ , as based on its role in resolving the original incompatible triad, he is *ipso facto* committed to both the prescriptive construal of normativity and the truth of OIC.<sup>16</sup>

---

<sup>14</sup> For evaluative construal of the truth norm see, for instance, William P. Alston, "Concepts of Epistemic Justification," *The Monist* 68, 2 (1985); Matthew Chrisman, "Ought to Believe," *Journal of Philosophy* 105, 7 (2008); Davide Fassio, "Belief, Correctness and Normativity," *Logique Et Analyse* 54, 216 (2011); Conor McHugh, "The Truth Norm of Belief," *Pacific Philosophical Quarterly* 93, 1 (2012); Conor McHugh, "Fitting Belief," *Proceedings of the Aristotelian Society* 114, 2 (2014).

<sup>15</sup> For recent arguments against OIC, see Graham, "'Ought' and Ability;" Mizrahi, "'Ought' Does Not Imply 'Can';" "Does 'Ought' Imply 'Can' from an Epistemic Point of View?;" Paul Henne et al., "An Empirical Refutation of 'Ought' Implies 'Can'," *Analysis* 76, 3 (2016).

<sup>16</sup> Acknowledgment: I am particularly grateful to Philip Ebert and Krister Bykvist for their extremely thoughtful and invaluable comments on earlier drafts of this paper. I would also like to thank the audience of the 6<sup>th</sup> Stockholm Graduate Conference where the paper was presented for their helpful questions and suggestions. This work was supported by the John Templeton Foundation under Grant 58450.

## **DISCUSSION NOTES/DEBATE**



# FACTIVITY AND EPISTEMIC CERTAINTY: A REPLY TO SANKEY

Moti MIZRAHI

ABSTRACT: This is a reply to Howard Sankey's comment ("Factivity or Grounds? Comment on Mizrahi") on my paper, "You Can't Handle the Truth: Knowledge = Epistemic Certainty," in which I present an argument from the factivity of knowledge for the conclusion that knowledge is epistemic certainty. While Sankey is right that factivity does not entail epistemic certainty, the factivity of knowledge does entail that knowledge is epistemic certainty.

KEYWORDS: epistemic certainty, factivity, fallibilism, knowledge

I am grateful to Howard Sankey for commenting on my paper, "You Can't Handle the Truth: Knowledge = Epistemic Certainty," in which I present an argument from the factivity of knowledge for the conclusion that knowledge is epistemic certainty.<sup>1</sup> The argument runs as follows:

- 1) If  $S$  knows that  $p$  on the grounds that  $e$ , then  $p$  cannot be false given  $e$ .
- 2) If  $p$  cannot be false given  $e$ , then  $e$  makes  $p$  epistemically certain.
- 3) Therefore, if  $S$  knows that  $p$  on the grounds that  $e$ , then  $e$  makes  $p$  epistemically certain.<sup>2</sup>

Sankey argues that it is the notion of grounds that is doing the work in this argument, not the notion of factivity.<sup>3</sup> As Sankey puts it:

the argument that Mizrahi presents does not in fact proceed from the factivity of knowledge to knowledge being epistemic certainty. Rather, the argument proceeds from an assumption about the relation between grounds and knowledge to the conclusion about epistemic certainty.<sup>4</sup>

Sankey argues that this argument proceeds from an assumption about grounds, not factivity, because, to say that knowledge is factive is to say that

---

<sup>1</sup> Moti Mizrahi, "You Can't Handle the Truth: Knowledge = Epistemic Certainty," *Logos & Episteme* 10, 2 (2019): 225-227.

<sup>2</sup> Mizrahi, "You Can't Handle the Truth," 225.

<sup>3</sup> Howard Sankey, "Factivity or Grounds? Comment on Mizrahi," *Logos & Episteme* 10, 3 (2019): 333-334.

<sup>4</sup> Sankey, "Factivity or Grounds?" 333.

“knowledge requires truth,”<sup>5</sup> whereas “the claim that knowledge is factive says nothing about a relation between grounds and knowledge.”<sup>6</sup>

Now, Sankey is right that, strictly speaking, *factivity* “says nothing about a relation between grounds and knowledge.”<sup>7</sup> But the claim that *knowledge* is factive does say something about a relation between grounds and knowledge. For, just as “knowledge requires truth,”<sup>8</sup> knowledge also requires justification. Just as it “is not possible to know a proposition if that proposition is false,”<sup>9</sup> it is also not possible to know a proposition if that proposition is unjustified. Accordingly, if *S* has no grounds for believing that *p*, then *S* cannot be said to know that *p*. On the other hand, if *S* knows that *p*, then *p* must be not only true but also justified. Therefore, the claim that knowledge is factive does say something about the relation between knowledge and grounds insofar as knowledge requires justification. And justification (i.e., reasons or evidence) is that which makes a proposition epistemically certain because, if *S* knows that *p*, then *p* cannot be false.<sup>10</sup>

Nevertheless, I suspect that the argument sketched above can be made without the explicit mention of justification or evidence for *p*, given that knowledge requires justification in much the same way that knowledge requires truth. That is:

- 1) If *S* knows that *p*, then *p* cannot be false.
- 2) If *p* cannot be false, then *p* is epistemically certain.
- 3) Therefore, if *S* knows that *p*, then *p* is epistemically certain.

I think that this argument works just as well as the previous one in showing that knowledge is epistemic certainty. Again, what guarantees the truth of *p*, i.e., what makes it such that *p* cannot be false, is *S*'s justification for *p*; justification that *S* must have if *S* can be said to know that *p*. Since *p* cannot be false because knowledge is factive, it follows that *S*'s justification for *p* must be such that it makes *p* epistemically certain. That is why knowledge = epistemic certainty.

---

<sup>5</sup> Sankey, “Factivity or Grounds?” 333.

<sup>6</sup> Sankey, “Factivity or Grounds?” 334.

<sup>7</sup> Sankey, “Factivity or Grounds?” 334.

<sup>8</sup> Sankey, “Factivity or Grounds?” 333.

<sup>9</sup> Sankey, “Factivity or Grounds?” 333.

<sup>10</sup> Of course, the claim that knowledge requires justification is an assumption just as much as the claim that knowledge requires truth is. Both, however, are assumptions that are widely accepted among epistemologists. Even those that reject traditional analyses of knowledge and embrace a so-called “knowledge first” epistemology, agree that knowledge entails truth, belief, and justification. See, for example, Timothy Williamson, *Knowledge and Its Limits* (Oxford: Oxford University Press, 2000).

# WHY MUST JUSTIFICATION GUARANTEE TRUTH? REPLY TO MIZRAHI

Howard SANKEY

**ABSTRACT:** This reply provides further grounds to doubt Mizrahi's argument for an infallibilist theory of knowledge. It is pointed out that the fact that knowledge requires both truth and justification does not entail that the level of justification required for knowledge be sufficient to guarantee truth. In addition, an argument presented by Mizrahi appears to equivocate with respect to the interpretation of the phrase "*p* cannot be false".

**KEYWORDS:** Moti Mizrahi, factivity, epistemic certainty, fallibilism, knowledge

## I.

In "You Can't Handle the Truth: Knowledge = Epistemic Certainty," Moti Mizrahi claims that the factivity of knowledge entails that knowledge is epistemic certainty.<sup>1</sup> In "Factivity or Grounds? Comment on Mizrahi," I pointed out that Mizrahi's argument that knowledge is epistemic certainty requires more than the simple assumption that knowledge is factive.<sup>2</sup> In addition, Mizrahi must also adopt an assumption about the relationship between grounds (or evidence) and knowledge.

In "Factivity and Epistemic Certainty: A Reply to Sankey," Mizrahi agrees with me on the above point.<sup>3</sup> He agrees that "strictly speaking" the assumption of factivity tells us nothing about the relationship between grounds and knowledge. However, he thinks that a version of his original claim can still be maintained. He asserts that "the claim that *knowledge* is factive does say something about a

---

<sup>1</sup> Moti Mizrahi, "You Can't Handle the Truth: Knowledge = Epistemic Certainty", *Logos & Episteme* X, 2 (2019): 225-227.

<sup>2</sup> Howard Sankey, "Factivity or Grounds? Comment on Mizrahi," *Logos & Episteme* X, 3 (2019): 333-334.

<sup>3</sup> Moti Mizrahi, "Factivity and Epistemic Certainty: A Reply to Sankey," *Logos & Episteme* X, 4 (2019): 443-444.

relation between grounds and knowledge.”<sup>4</sup> The reason is that in the same way that knowledge requires truth, it “also requires justification.”<sup>5</sup>

Mizrahi writes in more detail as follows:

... if *S* has no grounds for believing that *p*, then *S* cannot be said to know that *p*. On the other hand, if *S* knows that *p*, then *p* must be not only true but also justified. Therefore, the claim that knowledge is factive does say something about the relation between knowledge and grounds insofar as knowledge requires justification. And justification (i.e. reasons or evidence) is that which makes a proposition epistemically certain.<sup>6</sup>

In other words, it is because knowledge requires *both* truth *and* justification that the level of justification required for knowledge must be sufficiently high to guarantee truth. It is not just that knowledge is factive, but that it is factive and it requires justification.

## II.

Mizrahi assumes that knowledge requires truth. That is what is meant in this context by saying that knowledge is factive. He also assumes that knowledge requires justification. Hence, knowledge requires both truth and justification. Mizrahi takes the fact that knowledge requires both truth and justification to entail that justification must guarantee truth. For this reason, he assumes that the level of justification required for knowledge is certainty. For it is only if justification is epistemic certainty that justification may guarantee truth.

I regard the assumption that justification must guarantee truth as problematic. Like Mizrahi, I assume that knowledge requires both truth and justification. Truth and justification are necessary conditions for knowledge. But they are distinct conditions for knowledge: one condition may be met without the other being met. The assumption that knowledge requires truth and justification does not entail that the level of justification of a belief be sufficient to guarantee truth of the belief.

Mizrahi assumes that in order for a justified true belief to constitute knowledge the justification of the belief must guarantee the truth of the belief. In other words, justification must guarantee truth. It is entirely possible that an argument might be given for this assumption. But, so far as I can see, no such argument has been supplied by Mizrahi. The simple point that knowledge requires

---

<sup>4</sup> Mizrahi, “Factivity and Epistemic Certainty,” 443.

<sup>5</sup> Mizrahi, “Factivity and Epistemic Certainty,” 443.

<sup>6</sup> Mizrahi, “Factivity and Epistemic Certainty,” 444.



both truth and justification does not by itself entail that justification be a guarantor of truth.

### III.

Toward the end of his reply, Mizrahi offers the following argument for his view:

- (1) If  $S$  knows that  $p$ , then  $p$  cannot be false
- (2) If  $p$  cannot be false, then  $p$  is epistemically certain.
- (3) Therefore, if  $S$  knows that  $p$ , then  $p$  is epistemically certain.<sup>7</sup>

This argument may at first blush appear to be valid. On closer inspection, it appears to equivocate with respect to the phrase “ $p$  cannot be false.” In its first occurrence in premise (1), the phrase “ $p$  cannot be false” is taken to state a necessary condition for knowledge. But in its second occurrence in premise (2), the very same phrase is taken to either mean or entail that  $p$  must be certain. But the fact that, if  $p$  is false,  $S$  does not know that  $p$ , does not entail that  $p$  must be certain. Truth is a necessary condition for knowledge. To say that truth is a necessary condition for knowledge is not to say that knowledge requires certainty. It is just to say that if the proposition believed by the subject is false, then justified belief in that proposition does not constitute knowledge. It fails to be knowledge because the proposition in question is false.

### IV.

I do not wish to suggest that no argument may be given for the infallibilist view that the level of justification required for knowledge is epistemic certainty. What I do wish to suggest is that, in his original note and subsequent reply, Mizrahi has not provided such an argument. I have no doubt that one might have an intuition to the effect that justification must guarantee truth. But, without an argument, those of us who do not share that intuition are left without grounds to adopt the infallibilist point of view.

---

<sup>7</sup> Mizrahi, “Factivity and Epistemic Certainty,” 444.



# KNOWLEDGE DOESN'T REQUIRE EPISTEMIC CERTAINTY: A REPLY TO MIZRAHI

James SIMPSON

**ABSTRACT:** In a recent discussion note in this journal, Moti Mizrahi offers us the following argument for the conclusion that knowledge requires epistemic certainty:

- 1) If S knows that p on the grounds that e, then p cannot be false given e.
- 2) If p cannot be false given e, then e makes p epistemically certain.
- 3) Therefore, if S knows that p on the grounds that e, then e makes p epistemically certain.

I'll argue that (2) of Mizrahi's argument is false, and so, Mizrahi's argument is unsound.

**KEYWORDS:** knowledge, epistemic certainty, possible circumstance

In a recent discussion note in this journal, Moti Mizrahi<sup>1</sup> provides the following argument for the conclusion that knowledge requires epistemic certainty:

- 1) If S knows that p on the grounds that e, then p cannot be false given e.
- 2) If p cannot be false given e, then e makes p epistemically certain.<sup>2</sup>
- 3) Therefore, if S knows that p on the grounds that e, then e makes p epistemically certain.

Let's call this Mizrahi's Argument.

I'll argue that (2) of Mizrahi's Argument is false, and so Mizrahi's Argument is unsound. To see this, consider the following scenario:

**Math.** Suppose my mathematician dad, an honest and reliable fellow, tells me that  $2+2=4$ . On this basis, I come to believe that  $2+2=4$ .

---

<sup>1</sup> Moti Mizrahi, "You can't handle the truth: knowledge = epistemic certainty," *Logos & Episteme* X, 2 (2019): 225-227.

<sup>2</sup> Following Mizrahi (*ibid.*, 225) and Peter Klein, *Certainty: A Refutation of Scepticism* (University of Minnesota Press, 1981), 185, I'll take "e makes p epistemically certain" to mean e guarantees the truth of p.

James Simpson

Now, observe, that  $2+2=4$  is necessarily true, and so it cannot be false in any logically possible circumstance. If  $2+2=4$  cannot be false in any possible circumstance, then it cannot be false, given that my honest and reliable mathematician dad tells me that  $2+2=4$  and that I come to believe that  $2+2=4$  on this basis. Yet, intuitively, my honest and reliable mathematician dad telling me that it's true that  $2+2=4$  doesn't *guarantee* that it is true that  $2+2=4$ . After all, honest and reliable experts tell people things all the time that aren't true. In Math, of course, my dad tells me something that's necessarily true, and so he couldn't have told me something that isn't true, if he tells me that  $2+2=4$ . But, quite plausibly, what guarantees the truth of  $2+2=4$  isn't my dad telling me, in Math, that it's true that  $2+2=4$ . It's that, in fact,  $2+2=4$ .

If this interpretation of Math is correct, as is very plausible, then (2) of Mizrahi's Argument must be false. Since there's some metaphysically possible circumstance where it's true that  $p$  cannot be false given  $e$ , but it's false that  $e$  makes  $p$  epistemically certain—i.e., it's false that  $e$  guarantees the truth of  $p$ . Thus, Mizrahi's Argument is unsound.

## NOTES ON THE CONTRIBUTORS

**Scott Aikin** is Assistant Professor of Philosophy and Director of Undergraduate Studies at Vanderbilt University. His research is focused primarily in epistemology, ancient philosophy, and argumentation theory. He is the single author of two books: *Epistemology and the Regress Problem* (Routledge, 2011), and *Evidentialism and the Will to Believe* (Bloomsbury, 2014), and he has recently co-authored with Robert Talisse *Pragmatism, Pluralism, and the Nature of Philosophy* (Routledge, 2018) and *Political Argument in a Polarized Age* (Polity, 2020). Contact: scott.f.aikin@vanderbilt.edu.

**Michał Bukat** is a doctoral student of finance at Kozminski University in Warsaw, Poland. He holds a MSc double degree diploma in International Business from the Prague University of Economics and Jean Moulin University in Lyon, and a BA degree from the University of Warsaw. His main research interests are within the fields of behavioral economics, psychology, philosophy and all that relates to labor markets and investment decisions. Contact: mimakb@wp.pl; 1173-phdf@kozminski.edu.pl.

**Moti Mizrahi** is Associate Professor of Philosophy in the School of Arts and Communication at the Florida Institute of Technology. He is also an Associate and Book Reviews Editor of *Philosophia* and the editor of *The Kuhnian Image of Science: Time for a Decisive Transformation?* (Rowman & Littlefield, 2018). He has teaching and research interests in argumentation, epistemology, ethics, logic, philosophy of religion, and philosophy of science. He has published extensively on the philosophy of science, the scientific realism/anti-realism debate, the epistemology of philosophy, and argumentation. His work has appeared in journals such as *Argumentation*, *Erkenntnis*, *Philosophical Studies*, *Studies in History and Philosophy of Science*, and *Synthese*. Contact: motimizra@gmail.com.

**Tommaso Ostillio** is currently a doctoral student of Philosophy (University of Warsaw, Poland) and a doctoral student of Finance (Kozminski University in Warsaw, Poland). He holds a bachelor's degree in Philosophy and a master's of science in Strategic Management. His main research interests are: cognitive psychology, behavioural economics, experimental economics, experimental philosophy, epistemology and philosophy of economics. Contact: t.ostillio@student.uw.edu.pl; tostillio@kozminski.edu.pl.

**Seungbae Park** is a philosophy professor at Ulsan National Institute of Science and Technology in the Republic of Korea. He received his Ph.D. from the University of Arizona in 2001, having specialized in philosophy of science under the guidance of Prof. Richard Healey. He taught at the University of Arizona, the University of Maryland, and POSTECH before coming to his current institution in 2009. He served as an anonymous referee for such journals as *Journal of Philosophy*, *Philosophy of Science*, *Synthese*, *European Journal for Philosophy of Science*, *Foundations of Science*, *Philosophia*, *Philosophy of the Social Sciences*, *International Studies in the Philosophy of Science*, *Journal for General Philosophy of Science*, *Philosophical Papers*, and *Journal of Ethics*. Homepage: <https://philpeople.org/profiles/seungbae-park>. Contact: [nature@unist.ac.kr](mailto:nature@unist.ac.kr).

**Brian Ribeiro** is UC Foundation Professor of Philosophy at the University of Tennessee at Chattanooga. His research is focused on the history and significance of the Western skeptical tradition, with special attention to the philosophies of Sextus Empiricus, Michel de Montaigne, and David Hume. He also does work in philosophy of religion from a skeptical perspective. His published work has appeared in many journals, including *The Monist*, *Ancient Philosophy*, *British Journal of Aesthetics*, *Journal of Scottish Philosophy*, *Ratio*, and *International Journal for the Study of Skepticism*. Contact: [Brian-Ribeiro@utc.edu](mailto:Brian-Ribeiro@utc.edu).

**Hans Rott** is Professor of Philosophy at the University of Regensburg, Germany. He is the author of *Change, Choice and Inference* (Oxford University Press 2001) and of papers appearing in *Annals of Mathematics and Artificial Intelligence*, *Artificial Intelligence*, *Erkenntnis*, *Journal of Logic and Computation*, *Journal of Philosophical Logic*, *Journal of Philosophy*, *Journal of Symbolic Logic*, *Kant-Studien*, *Linguistics and Philosophy*, *Mind*, *Minds and Machines*, *Studia Logica*, *Synthese*, *Topoi* and *Zeitschrift für philosophische Forschung*. Much of his research focusses on the logical analysis of belief, the dynamics of belief systems and defeasible reasoning. He is also interested in doxastic voluntarism, the logic of conditionals and the analysis of disputes and disagreements. Contact: [hans.rott@ur.de](mailto:hans.rott@ur.de).

**Howard Sankey** is Associate Professor of Philosophy in the School of Historical and Philosophical Studies at the University of Melbourne (Australia). He teaches in epistemology and philosophy of science. He has published on the problem of incommensurability, epistemic relativism and scientific realism. He is the author of *Scientific Realism and the Rationality of Science* (Ashgate, 2008), *Theories of Scientific Method: An Introduction* (Acumen, 2007, with Robert Nola), *Rationality, Relativism and Incommensurability* (Ashgate, 1997) and *The*

*Incommensurability Thesis* (Avebury, 1994). More information may be found at <https://philpeople.org/profiles/howard-sankey>. Contact: [chs@unimelb.edu.au](mailto:chs@unimelb.edu.au).

**Michael J. Shaffer** is currently professor of philosophy at St. Cloud State University. He is also a fellow of the center for formal epistemology at Carnegie-Mellon University, a fellow of the Rotman Institute for Science and Values at the University of Western Ontario, a Lakatos fellow at the London School of Economics, a fellow of the University of Cologne's summer institute for epistemology and an NEH fellow at the University of Utah. His main areas of research interest are in epistemology, logic and the philosophy of science, and he has published more than fifty articles and book chapters on various topics in these areas. He is co-editor of *What Place for the A Priori?* (Open Court, 2011) and is the author of *Counterfactuals and Scientific Realism* (Palgrave-MacMillan, 2012), *Quasi-factive Belief and Knowledge-like States* (Lexington, forthcoming) and *The Experimental Turn and the Methods of Philosophy* (Routledge, forthcoming). Contact: [shaffermphil@hotmail.com](mailto:shaffermphil@hotmail.com).

**James Simpson** is a Ph.D. candidate at the University of Florida. His interests are in epistemology, ethics, metaphysics, the philosophy of action, and the philosophy of religion. His most pressing current research project involves a dissertation-length defense of a novel analysis of knowledge. Contact: [simpson.james@ufl.edu](mailto:simpson.james@ufl.edu).

**Xintong Wei** is a PhD student at the St Andrews and Stirling Graduate Programme in Philosophy (SASP); and a member of the *Knowledge Beyond Natural Science* research project based in the Philosophy Department at the University of Stirling. She works mainly in epistemology and normativity. Contact: [xw30@st-andrews.ac.uk](mailto:xw30@st-andrews.ac.uk).





## ***LOGOS & EPISTEME*: AIMS & SCOPE**

*Logos & Episteme* is a quarterly open-access international journal of epistemology that appears at the end of March, June, September, and December. Its fundamental mission is to support philosophical research on human knowledge in all its aspects, forms, types, dimensions or practices.

For this purpose, the journal publishes articles, reviews or discussion notes focused as well on problems concerning the general theory of knowledge, as on problems specific to the philosophy, methodology and ethics of science, philosophical logic, metaphilosophy, moral epistemology, epistemology of art, epistemology of religion, social or political epistemology, epistemology of communication. Studies in the history of science and of the philosophy of knowledge, or studies in the sociology of knowledge, cognitive psychology, and cognitive science are also welcome.

The journal promotes all methods, perspectives and traditions in the philosophical analysis of knowledge, from the normative to the naturalistic and experimental, and from the Anglo-American to the Continental or Eastern.

The journal accepts for publication texts in English, French and German, which satisfy the norms of clarity and rigour in exposition and argumentation.

*Logos & Episteme* is published and financed by the "Gheorghe Zane" Institute for Economic and Social Research of The Romanian Academy, Iasi Branch. The publication is free of any fees or charges.

For further information, please see the Notes to Contributors.

Contact: [logosandepisteme@yahoo.com](mailto:logosandepisteme@yahoo.com).



# NOTES TO CONTRIBUTORS

## 1. Accepted Submissions

The journal accepts for publication articles, discussion notes and book reviews.

Please submit your manuscripts electronically at: [logosandepisteme@yahoo.com](mailto:logosandepisteme@yahoo.com). Authors will receive an e-mail confirming the submission. All subsequent correspondence with the authors will be carried via e-mail. When a paper is co-written, only one author should be identified as the corresponding author.

There are no submission fees or page charges for our journal.

## 2. Publication Ethics

The journal accepts for publication papers submitted exclusively to *Logos & Episteme* and not published, in whole or substantial part, elsewhere. The submitted papers should be the author's own work. All (and only) persons who have a reasonable claim to authorship must be named as co-authors.

The papers suspected of plagiarism, self-plagiarism, redundant publications, unwarranted ('honorary') authorship, unwarranted citations, omitting relevant citations, citing sources that were not read, participation in citation groups (and/or other forms of scholarly misconduct) or the papers containing racist and sexist (or any other kind of offensive, abusive, defamatory, obscene or fraudulent) opinions will be rejected. The authors will be informed about the reasons of the rejection. The editors of *Logos & Episteme* reserve the right to take any other legitimate sanctions against the authors proven of scholarly misconduct (such as refusing all future submissions belonging to these authors).

## 3. Paper Size

The articles should normally not exceed 12000 words in length, including footnotes and references. Articles exceeding 12000 words will be accepted only occasionally and upon a reasonable justification from their authors. The discussion notes must be no longer than 3000 words and the book reviews must not exceed 4000 words, including footnotes and references. The editors reserve the right to ask the authors to shorten their texts when necessary.

#### **4. Manuscript Format**

Manuscripts should be formatted in Rich Text Format file (\*.rtf) or Microsoft Word document (\*.docx) and must be double-spaced, including quotes and footnotes, in 12 point Times New Roman font. Where manuscripts contain special symbols, characters and diagrams, the authors are advised to also submit their paper in PDF format. Each page must be numbered and footnotes should be numbered consecutively in the main body of the text and appear at footer of page. For all references authors must use the Humanities style, as it is presented in The Chicago Manual of Style, 15th edition. Large quotations should be set off clearly, by indenting the left margin of the manuscript or by using a smaller font size. Double quotation marks should be used for direct quotations and single quotation marks should be used for quotations within quotations and for words or phrases used in a special sense.

#### **5. Official Languages**

The official languages of the journal are: English, French and German. Authors who submit papers not written in their native language are advised to have the article checked for style and grammar by a native speaker. Articles which are not linguistically acceptable may be rejected.

#### **6. Abstract**

All submitted articles must have a short abstract not exceeding 200 words in English and 3 to 6 keywords. The abstract must not contain any undefined abbreviations or unspecified references. Authors are asked to compile their manuscripts in the following order: title; abstract; keywords; main text; appendices (as appropriate); references.

#### **7. Author's CV**

A short CV including the author's affiliation and professional postal and email address must be sent in a separate file. All special acknowledgements on behalf of the authors must not appear in the submitted text and should be sent in the separate file. When the manuscript is accepted for publication in the journal, the special acknowledgement will be included in a footnote on the first page of the paper.

## **8. Review Process**

The reason for these requests is that all articles which pass the editorial review, with the exception of articles from the invited contributors, will be subject to a strict double anonymous-review process. Therefore the authors should avoid in their manuscripts any mention to their previous work or use an impersonal or neutral form when referring to it.

The submissions will be sent to at least two reviewers recognized as specialists in their topics. The editors will take the necessary measures to assure that no conflict of interest is involved in the review process.

The review process is intended to be as quick as possible and to take no more than three months. Authors not receiving any answer during the mentioned period are kindly asked to get in contact with the editors.

The authors will be notified by the editors via e-mail about the acceptance or rejection of their papers.

The editors reserve their right to ask the authors to revise their papers and the right to require reformatting of accepted manuscripts if they do not meet the norms of the journal.

## **9. Acceptance of the Papers**

The editorial committee has the final decision on the acceptance of the papers. Papers accepted will be published, as far as possible, in the order in which they are received and they will appear in the journal in the alphabetical order of their authors.

## **10. Responsibilities**

Authors bear full responsibility for the contents of their own contributions. The opinions expressed in the texts published do not necessarily express the views of the editors. It is the responsibility of the author to obtain written permission for quotations from unpublished material, or for all quotations that exceed the limits provided in the copyright regulations.

## **11. Checking Proofs**

Authors should retain a copy of their paper against which to check proofs. The final proofs will be sent to the corresponding author in PDF format. The author

must send an answer within 3 days. Only minor corrections are accepted and should be sent in a separate file as an e-mail attachment.

## 12. Reviews

Authors who wish to have their books reviewed in the journal should send them at the following address:

Institutul de Cercetări Economice și Sociale „Gh. Zane”

Academia Română, Filiala Iași

Str. Teodor Codrescu, Nr. 2

700481

Iași, România.

The authors of the books are asked to give a valid e-mail address where they will be notified concerning the publishing of a review of their book in our journal. The editors do not guarantee that all the books sent will be reviewed in the journal. The books sent for reviews will not be returned.

## 13. Property & Royalties

Articles accepted for publication will become the property of *Logos & Episteme* and may not be reprinted or translated without the previous notification to the editors. No manuscripts will be returned to their authors. The journal does not pay royalties.

## 14. Permissions

Authors have the right to use their papers in whole and in part for non-commercial purposes. They do not need to ask permission to re-publish their papers but they are kindly asked to inform the Editorial Board of their intention and to provide acknowledgement of the original publication in *Logos & Episteme*, including the title of the article, the journal name, volume, issue number, page number and year of publication. All articles are free for anybody to read and download. They can also be distributed, copied and transmitted on the web, but only for non-commercial purposes, and provided that the journal copyright is acknowledged.

### **15. Electronic Archives**

The journal is archived on the Romanian Academy, Iasi Branch web page. The electronic archives of *Logos & Episteme* are also freely available on Philosophy Documentation Center web page.